We define a new distance measure for ranking data using a mixture of copula functions. Our distance measure evaluates the dissimilarity of subjects’ ranking preferences to segment them via hierarchical cluster analysis. The proposed distance measure builds upon Spearman grade correlation coefficient on a copula transformation of rank denoting the level of importance assigned by subjects on the classification of objects. These mixtures of copulae enable flexible modeling of the different types of dependence structures found in data and the consideration of various circumstances in the classification process. For example, by using mixtures of copulae with lower and upper tail dependence, we can emphasize the agreement on extreme ranks when they are considered important.

KEYWORDS
distance measure, hierarchical cluster analysis, mixture of copulae, ranking data

1 INTRODUCTION

Clustering of ranking data aims to identify groups of subjects with homogenous and common preference behavior. Ranking data occur when subjects are asked to rank a list of objects according to their personal preferences. The input to cluster analysis is a distance matrix whose elements measure the distance between the rankings of two subjects. The chosen distance measure dramatically affects the final results. When dealing with ordinal data, the issue lies in computing an appropriate distance matrix. Several distance measures have been proposed for ranking data [1], the most important of which are Kendall’s \( \tau \), Spearman’s \( \rho \), and Cayley distances [1,10,13].

These distance measures can be written in terms of similarity functions based on standard association measures (ie, Spearman \( \rho \) rank correlation coefficient, Kendall’s \( \tau \), and so on).

When the aim is to emphasize top ranks, weighted distances for ranking data are proposed [5,14] by introducing weights in the formulas of \( \rho \) and \( \tau \). These measures are called weighted rank correlation (WRC) measures. In this context, since the most important dependence measures can be expressed by copula functions, ref. [3] propose a distance measure for ranking data based on copula functions with (lower) tail dependence for emphasizing the agreement of top ranks when they are considered more important than lower ranks.

In this paper, focusing on untied rankings, we propose a generalization of these kinds of distance measures using mixtures of copulae. This creates a flexible instrument for modeling different types of data dependence structures and considering various circumstances in the classification process. For example, by using a mixture of copulae with lower tail dependence, we can emphasize the agreement between subjects on the top ranks, or, by using a mixture of copulae with upper tail dependence, we can emphasize the low ranks. A mixture of copulae with both lower and upper tail dependence permits us to simultaneously assign more weight to both extremes (the top and lower ranks). In addition, a mixture of copulae permits us to model particular circumstances to meet different segmentation goals and emphasize atypical agreement rankings in the classification process. For example, we can emphasize agreement on the extreme ranks (lower and top ranks), or on the top and central ranks but not the lower ones, or on the central and lower ranks but not the top ones, and so on.

Thus, we employ an appropriate flexible measure to evaluate the dissimilarity between rankings in hierarchical cluster analysis to segment consumers’ ranking preferences (see also [14]). The type of dependence structure one chooses to emphasize determines the appropriate mixture of copulae.
In this manner, mixtures of copulae are a generalization of the previous approaches presented in ref. [3], which involve the selection of only one copula family to emphasize a particular dependence structure.

In this way, mixture of copulae is a generalization of the previous approaches presented in ref. [3], where there was the choice of only one family of copulae, in order to emphasize only a particular dependence structure.

There are several advantages to our approach. First, we propose a measure of distance for ranking data. Second, by using copulae, we can consider more general dependence structures than the linear ones measured by the Pearson’s correlation coefficient. Moreover, we generalize the traditional approach based on the Spearman’s rank correlation coefficient. It is possible to prove that, using a particular copula function (Gaussian copula), the two approaches produce the same classification. Finally, the use of a mixture of copulae permits one to generalize any dependence structure and emphasize particular rankings of interest without the specific a priori choice of weights performed in classical WRC.

The remainder of this paper is organized as follows. In Section 2, we present a definition of a bivariate copula, the main families of copulae, and the definition of a mixture of copulae. In particular, the Spearman’s grade correlation coefficient’s relationship with specific copula families is considered. In Section 3, our proposal of a new measure to evaluate the distance between rankings in hierarchical cluster analysis is presented. This proposal involves the use of a mixture of copulae that permits the emphasis of various agreements. In Section 4, we introduce a mixture of copulae into the classification process and discuss its properties. In Section 5, we evaluate the performance of our proposal on simulated data by comparing a classical approach, a copula method, and a mixture of copulae. Moreover, a comparison with a WRC measure approach is discussed. The final section presents a discussion of the proposed approach.

2 | MIXTURE OF COPULAE

A bivariate copula is a function that links a bivariate distribution function to its marginal distributions. Bivariate copulae are a class of bivariate distributions whose marginals are uniform on the unit interval. A bivariate copula describes the dependence of the structure across pairwise marginal random variables (rvs).

Sklar’s theorem [11] shows that every bivariate/multivariate distribution can be defined via a copula representation. Let \((Y_1, Y_2)\) be a bivariate rv with marginal cumulative distribution functions (cdfs) \(F_{Y_1}(y_1)\) and \(F_{Y_2}(y_2)\), and joint cdf \(F_{Y_1,Y_2}(y_1,y_2; \lambda)\). There exists a copula function \(C(\cdot, \cdot; \lambda)\) with \(C : \mathbb{R}^2 \rightarrow [1]\) such that

\[
F_{Y_1,Y_2}(y_1,y_2; \lambda) = C(F_{Y_1}(y_1), F_{Y_2}(y_2); \lambda), \quad y_1, y_2 \in \mathbb{R}. \tag{1}
\]

If the marginal cdfs are absolutely continuous, the copula \(C(\cdot, \cdot; \lambda)\) is unique, otherwise \(C(\cdot, \cdot; \lambda)\) is uniquely determined on \(\text{Range}F_{Y_1} \times \text{Range}F_{Y_2}\). Moreover, if \(F_{Y_1}(y_1)\) and \(F_{Y_2}(y_2)\) are continuous, the copula can be found by the inverse of Equation (1):

\[
C(u,v) = F_{Y_1}^{-1}(F_{Y_1}^{-1}(u), F_{Y_2}^{-1}(v); \lambda) \tag{2}
\]

with \(u = F_{Y_1}(y_1)\) and \(v = F_{Y_2}(y_2)\).

This theorem states that each joint distribution can be expressed in terms of two separate but related issues: the marginal distributions and the dependence structures between them. The dependence structures are described by the copula function. Note that the Equation (1) provides a general mechanism for constructing new bivariate models in a straightforward manner. Thus, we can build new bivariate distributions with different dependence structures by changing the copula function.

2.1 | Mixture of copulae

We consider a new family of copulae defined via finite mixtures [11]. The idea is to create a new extremely flexible copula by combining two copulae as follows:

\[
C_M(u,v) = aC_1(u,v; \lambda_1) + (1 - a)C_2(u,v; \lambda_2) \tag{3}
\]

where \(0 \leq a \leq 1\) and \(C_1\) and \(C_2\) are two copulae, and \(\lambda_1\) and \(\lambda_2\) are the parameters associated with the two copulae, respectively. Notably, the two components cannot be from the same copula family.

It can be easily verified that a finite mixture of copulae is also a copula that satisfies all the properties of copulae described below.

We can express the most important dependence measures by using copula functions [11]. For example, the Spearman’s grade correlation coefficient measures the association between two variables. More precisely, if two rvs are continuous and have copula \(C\) with parameter \(\lambda\), then the Spearman’s grade correlation is the correlation between the marginal distributions:

\[
\rho_s(C_\lambda) = 12 \int_{\mathbb{R}^2} C(u_1,u_2)d\mu(u_1)d\mu(u_2) - 3
\]

\[
= \frac{\text{Cov}(U_1, U_2)}{\sqrt{\text{Var}(U_1)}\sqrt{\text{Var}(U_2)}}. \tag{4}
\]

Among all copulas, three especially noteworthy ones are \(W(u,v) = \max(u + v - 1, 0)\), \(\Pi(u,v) = uv\), and \(M(u,v) = \min(u,v)\). These copulae correspond to a perfect negative association \((\rho_s(C_\lambda) = -1)\), independence (that implies \(\rho_s(C_\lambda) = 0\)), and a perfect positive association \((\rho_s(C_\lambda) = 1)\) between the two rvs, respectively. For all \((u,v) \in \mathbb{R}^2\), it holds that \(W(u,v) \leq \Pi(u,v) \leq M(u,v)\).

Another important aspect of a copula function is its tail dependence, which measures the association between the marginal cdfs \(F_X(x)\) and \(F_Y(y)\) in the tails of the distributions, that is, the high/low values of one variable are associated with
the high/low values of the other one. In particular, the parameter used to measure the upper tail dependence is defined as
\[
\chi_u = \lim_{u \to 1} P[Y > F_Y^{-1}(u)|X > F_X^{-1}(u)]. \quad (5)
\]
Similarly, the lower tail dependence parameter is given by
\[
\chi_l = \lim_{u \to 0} P[Y \leq F_Y^{-1}(u)|X \leq F_X^{-1}(u)]. \quad (6)
\]
Both parameters \(\chi_u\) and \(\chi_l\) assume values over the interval [0, 1] where the higher the value of the parameter is, the higher the intensity of the tail dependence. The Spearman’s grade correlation coefficient and both tail dependence parameters are directly associated with the parameters of some copula family [11].

In this paper, we employ a mixture of three copula functions typically used in classification problems: the Clayton, Gumbel, and Gaussian copulae. Table 1 shows the main characteristics of the cited copula functions.

The three copulae present several important characteristics. The Gaussian copula has radial symmetry and is the copula of the bivariate normal distribution. \(\Phi_2\) is the cdf of a bivariate standard Gaussian random variable, and \(\Phi_1\) and \(\Phi_2\) represent the cdf of a univariate standard Gaussian random variable. The \(\lambda\) parameter of the Gaussian copula represents a linear correlation coefficient that assumes positive and negative values. Therefore, assuming a Gaussian copula, the marginal probabilities have the same level of correlation (positive or negative) below and above their mean, and there is no higher association in the tails of the distribution. In contrast, the Clayton and Gumbel copulae are asymmetric (exchangeable) and exhibit tail dependencies. Both the Clayton and Gumbel copulae solely allow for positive association between variables (\(\tau \geq 0\)), yet they exhibit strong \(\text{left} \) and strong \(\text{right} \) tail dependence, respectively. Table 2 shows dependence coefficients, \(\chi_l\) and \(\chi_u\). Note that, for the Gaussian copula, the relationship between Spearman’s \(\rho_S\) and the copula parameter (corresponding to the Pearson’s correlation coefficient) is \(\lambda = 2 \sin(\rho_S)\). The expression of the equation converting \(\lambda\) to Spearman’s \(\rho_S\) is not reported because either it has no closed form expression or is expressed in a complicated form (see [6] for a more detailed explanation). In the former case of having no closed form expression, \(\rho_S\) is calculated by numerical approximation techniques as implemented in standard copula packages (e.g. the Copula R package).

In Equation (3), by combining the considered copulae with different dependence structures, we can create new, extremely flexible copulae enabling several combinations of association types that differ from the linear correlation of the normal distribution. Note that we have selected the most well-known and widely used of these in the applications to illustrate our proposal. The characteristics of the newly created copulae and their impact on our proposal in clustering ranking data are illustrated in Section 4. Extensions with several other copulae are straightforward to obtain.

### 3 DISSIMILARITY MEASURE VIA MIXTURE OF COPULAE

We propose using a mixture of copulae and the related Spearman’s \(\rho_S\) correlation coefficient to define a dissimilarity measure between subjects in hierarchical cluster analysis for ranking data.

Following the idea of polychoric correlation [12], we propose a method to evaluate the agreement between two subjects, \(S_a\) and \(S_b\), expressing their preferences on \(k\) objects by rankings.

We consider the mixture of copula \(C_M\) in Equation (3) to describe the dependence structure of each pair of latent continuous variables \((Y^a_i, Y^b_j)\) underlying the pair \((Y^a, Y^b)\) for \(i, j = 1, 2, \ldots, k\). \((Y^a, Y^b)\) is a bivariate ordinal variable where \(i\) and \(j\) represent the rank denoting the decreasing level of importance assigned by subjects \(S_a\) and \(S_b\) on \(k\) objects, and \(p_{ij}\) is the joint relative frequency with value \(1/k\) if the pair \((i, j)\) is observed, and 0 otherwise. The following example better clarifies the definition and meaning of the bivariate variable \((Y^a, Y^b)\) and correspondent probabilities \(p_{ij}\). An association measure calculated on this table reflects the degree of agreement between the two subjects.

**Example** We consider two subjects, \(S_a\) and \(S_b\), who order three objects \((Ob_1, Ob_2,\) and \(Ob_3)\) according to their preferences. The rankings of the two subjects are:

\[
S_a: [Ob_2, Ob_1, Ob_3]
\]
\[
S_b: [Ob_3, Ob_2, Ob_1]
\]

### TABLE 1 Copula functions and corresponding lower and upper tail dependence parameters (\(\chi_l, \chi_u\))

| Copula   | \(C(u,v)\)                                      | \(\lambda \in \Lambda\) |
|----------|-------------------------------------------------|--------------------------|
| Gaussian | \(\Phi_2(\Phi_1^{-1}(u), \Phi_1^{-1}(v))\)      | \((-1, 1)\)              |
| Clayton  | \((u^{-\lambda} + v^{-\lambda} - 1)^{-\lambda/\lambda}\) | \((0, \infty)\)         |
| Gumbel   | \(\exp\left[-((-\ln(u))^\lambda + (-\ln(v))^\lambda)^{\lambda/\lambda}\right]\) | \([1, \infty)\)         |

### TABLE 2 Copula parameter values and the related dependence and tail dependence coefficients

| Copula   | \(\lambda\) | \(\rho_S\) | \(\chi_l = 0\) | \(\chi_u = 0\) |
|----------|-------------|------------|----------------|----------------|
| Gaussian | 0.5         | 0.483      | 0.000          | 0.000          |
|          | 0.7         | 0.682      | 0.000          | 0.000          |
|          | 0.9         | 0.891      | 0.000          | 0.000          |
| Clayton  | 2           | 0.683      | 0.707          | 0.000          |
|          | 8           | 0.941      | 0.917          | 0.000          |
|          | 18          | 0.984      | 0.962          | 0.000          |
| Gumbel   | 2           | 0.683      | 0.000          | 0.586          |
|          | 5           | 0.943      | 0.000          | 0.851          |
|          | 10          | 0.986      | 0.000          | 0.928          |
Therefore, the positions (ranks) of the three objects are:

- $S_a : [2,1,3]
- S_b : [1,2,3]

The bivariate variable $(Y_a, Y_b)$ is defined in Table 3.

Now, let $F_1$ and $F_2$ be the cdfs of $Y_a$ and $Y_b$. We assume that each pair $(Y_a, Y_b)$ corresponds to the bivariate discrete random variable obtained by discretization of the continuous uniform latent variable $(U = F(Y_a), V = F(Y_b))$ with support on $I^2$ and the cdf given by $C_M$.

Let $A_{ij} = [u_{i-1}, u_i] \times [v_{j-1}, v_j]$, $i, j = 1, 2, \ldots, k$, be the rectangles defining the discretization. Let $p_{11}, \ldots, p_{kk}$ be the joint probabilities of the ordinal variables corresponding to rectangles $A_{11}, \ldots, A_{kk}$.

Let $V_{C_a}(A_{11}), \ldots, V_{C_a}(A_{kk})$ be the volumes of the rectangles under copula $C_a$. There exists a unique element in the family of the mixture of copulae that satisfies the following relationship:

$$(V_{C_a}(A_{11}), \ldots, V_{C_a}(A_{ij}), \ldots, V_{C_a}(A_{kk})) = (p_{11}, \ldots, p_{ij}, \ldots, p_{kk}).$$  \hfill (7)

Given the mixture of copulae $C_M$ that satisfies Equation (7), we define the Spearman’s grade correlation coefficients for the pair $(Y_a, Y_b)$, where $a \neq b$, that performs well in measuring the agreement between two rankings:

$$\rho_S(C_M) = 12 \int_{I^2} [aC_1(u, v; \lambda_1) + (1 - a)C_2(u, v; \lambda_2)] dudv - 3$$  \hfill (8)

The Spearman’s grade correlation coefficients of the convex combination of copulae correspond to the convex combination of the individual Spearman’s rho of the two copulae:

$$\rho_S(C_M) = a \rho_{S1} + (1 - a)\rho_{S2}$$  \hfill (9)

Finally, the distance $d^C_{a,b}$ between the rankings of subjects $a$ and $b$ is defined as:

$$d^C_{a,b} = 1 - \frac{\rho_S(C_M) + 1}{2}$$  \hfill (10)

The measure proposed in Equation (10) satisfies all the properties of distance for ranking data [1,9]. Let $\pi_a$ and $\pi_b$ be two different rankings of $k$ objects, then:

- $d(\pi_a, \pi_a) = 0$,
- $d(\pi_a, \pi_b) > 0$ if $\pi_a \neq \pi_b$,
- $d(\pi_a, \pi_b) = d(\pi_b, \pi_a)$, if the copulae in the mixture are exchangeable, the distance must also be right invariant: the relabeling of the objects has no effect on the distance.

We calculate the distances in Equation (10) for each pair of $n$ subjects. We propose to use the obtained $n \times n$ matrix as a dissimilarity matrix for hierarchical cluster analysis.

The values of the copula dependence parameters $\lambda_1$ and $\lambda_2$ establish the level of association/agreement between the rankings of the two subjects. If we consider two extreme value copulae with lower and upper tail dependence, the two parameters indicate the strength of the lower and upper tail dependence as well as the subjects’ agreement on the top and lower ranks. The parameters $\lambda_1$ and $\lambda_2$ in Equation (9) are data driven rather than chosen a priori (as are the weights in WRC indices). In particular, they are estimated by the maximum likelihood method. In this manner, the degree of association and the “weight” of agreement in the top/lower ranks are estimated by the data.

## 4 | CHOICE OF THE MIXTURE IN THE CLASSIFICATION PROCESS AND ITS PROPERTIES

Using a mixture of copulae permits the flexible modeling of different types of data dependence structures and the consideration of different circumstances in the classification process. For example, consider the three families of Archimedean copulae described in Section 2: the Gaussian, Clayton, and Gumbel copula. Recall that the Gaussian copula permits positive and negative correlation between the variables and does not allow dependence in the tails. In contrast, the Clayton and Gumbel copulae solely permit positive association and exhibit strong left (lower) and right (upper) tail dependencies, respectively. Figure 1 shows contour plots for the considered copulae using several values of the parameters according to Table 2.

By mixing two copulae in several ways, we can create several families of copulae with different characteristics, especially in the tails. The tail dependence measures of such a mixture are quite interesting. If tail dependence measures exist for mixing copula $C_M$, then the lower/upper tail dependence measures are a convex combination of the lower/upper tail dependence measures of the two copulae [2]:

$$\chi_{1M} = a \chi_{1l} + (1 - a) \chi_{1u}$$
$$\chi_{0M} = a \chi_{0l} + (1 - a) \chi_{0u}$$

If we choose two copulae in the mixture with opposite tail dependence structures, that is, the Clayton and Gumbel copulae, we create a new copula that accounts for both upper and lower tail dependencies:

$$\chi_{1M} = a \chi_{1l}^{1/\lambda_1}$$
$$\chi_{0M} = (1 - a)2 - \chi_{1M}^{1/\lambda_2}$$  \hfill (11)
where $\chi_{LM}$ is an increasing function of $\alpha$ and $\chi_{uM}$ is a decreasing function of $\alpha$.

A discussion of the properties and disadvantages of these indices in the classification of extreme events is presented in refs. [7,8].

Figures 2–5 show contour plots for several combinations of the considered copulae with uniform marginal distributions. The figures reveal the wide range of shapes that the mixture can take. Combining copulae with different structures, we can construct dependence structures that are very different.
from those of multivariate normal models. Additional pairs of copulae can also be used.

The use of a different mixture in Equation (10) leads to different results in the classification process for ranking data and permits us to alternatively emphasize the agreement between the rankings. For example, we can assign more “weight” to extreme ranks. Using only Gaussian copulae in the mixture, we must assign the same “weight” to all ranks. By choosing to use a mixture of Gaussian and Clayton or Gumbel copulae, we can assign different “weight” to the observations and emphasize the association of one lower or upper tail and the agreement of the top or lower ranks. Finally, by using a mixture of Clayton and Gumbel copulae, we can emphasize the agreement of both the lower and upper ranks. The $\alpha$ parameter in the mixture can be selected in accordance with a researcher’s goals. If he/she wishes to emphasize the agreement on the extreme ranks, with more importance given to the top than lower ranks, a high $\alpha$ to the copula must be assigned that enables lower tail dependence and, consequently, a low weight must be assigned to the copula that allows upper tail dependence. For example, if a mixture of Gumbel and Clayton copulae with weights equal to $\alpha = 0.2$ and $1 - \alpha = 0.8$ is chosen, the researcher wishes to emphasize the top ranks over the lower ones.

5 | EXAMPLE

In this section, we assess the proposed model with hierarchical cluster analysis using simulated data. We compare the results obtained with our distance measure based on a mixture of copulae with those based on classical Spearman’s
correlation and Spearman’s grade correlation of Clayton copula [3]. In Table 4 we consider 10 rankings representing the judgments of 10 consumers on six aspects of a product, attributing “1” to the most important aspect and “6” to the least important one. No ties are allowed.

Our aim is to emphasize the extreme ranks (top ranks, lower ranks, or both simultaneously, but with varying emphasis) to develop a more flexible classification than the classical one obtained by the Spearman correlation coefficient.

Referring to Table 4, let us consider consumers 1, 2, 3, 4, and 8.

If our primary focus is to emphasize the top ranks, the preferences of consumers 1, 3, and 8 are more similar than those of consumers 1, 2, and 4. On the one hand, 1 and 3 exchange the lowest ranks and 1 and 8 exchange the central ranks. On the other hand, 1 and 2 exchange the top ranks, while 1 and 4 exchange the first two top and last two ranks.

Ideally, a classification that emphasizes the top ranks approaches that of 1, 3, and 8 and separates them from consumers 2 and 4.

If we emphasize the top and lower ranks simultaneously, the preferences of consumers 1 and 8 are more similar than those of 1 and 3. Thus, a classification that emphasizes both the top and the lower ranks approaches that of 1 and 8 and separates consumer 3.

We performed several hierarchical cluster analyses on the selected data by using the classical Spearman’s correlation coefficient $\rho$ (see Figure 6), the Spearman’s grade correlation coefficient $\rho_s^C$ obtained by Clayton copula (see Figure 7), and a mixture of Gumbel and Clayton copulae with weight $\alpha = 0.5$ (see Figure 8). We performed our cluster analysis using a complete linkage clustering method.

The results are visualized with a dendrogram and an evaluation of distance matrices. Tables 5–7 show the distance matrices in the hierarchical cluster analysis conducted with
Spearman’s correlation coefficient $\rho$, the Spearman’s grade correlation coefficient $\rho^C$ using Clayton copula, and the Spearman grade correlation coefficient $\rho^{GC}$ using a mixture of Gumbel and Clayton copulae with weight $\alpha = 0.5$, respectively. The results obtained using other mixtures are available to interested authors upon request.

The classical approach assigns the same importance (weights) at every rank. Clayton copulae emphasize the top ranks (ranks 1 and 2 in our example) due to strong lower tail association.

The mixture of Gumbel and Clayton copulae with weight $\alpha = 0.5$ (equal weight for every copula) assigns different weights to the ranks and only emphasizes agreement on the extreme ranks (top and lower ranks). The mixture of copulae generalizes the previous two approaches.

The three methods lead to very different results, both in terms of graphical visualization by dendrogram and evaluation of distance matrices.

Consider consumers 1, 2, 3, and 8. If our aim is to emphasize only the top ranks, the distance between 1 and 2 should be greater than the distance between 1 and 3 or 1 and 8. The hierarchical clustering obtained by classical Spearman’s $\rho$ gives the same value of the distance ($d = 0.042$) computed between 1 and 2, 1 and 3, and 1 and 8, while Spearman’s grade correlation coefficient using Clayton copula captures the similarity between 1 and 3 with $d^C = 0.014$, which is greater than that between 1 and 2, $d^C = 0.070$. Consumers 1 and 3, whose preferences differ only on the two lowest ranks, are grouped together at a very low height in Figure 7, while they are grouped at a greater height in Figure 6.
If our aim is to emphasize extreme ranks (both top and lower ranks), the distance between 1 and 8 should be less than that between 1 and 2. The Spearman’s grade correlation coefficient using a mixture of Gumbel and Clayton copula gives $d_{GC} = 0.005$ between 1 and 8 while it gives $d_{GC} = 0.076$ between 1 and 2. As a matter of fact, the extreme ranks between 1 and 8 are equal (only the two central ranks are different), while only the central ranks are equal for consumers 1 and 2.

If we wish to consider both the top and lower ranks by comparing the distances in 5 and 6 the distances between 1 and 8 are equal to 0.022 and 0.005, respectively, since 1 and 8 have the same top and lower ranks. Consumers 1 and 8, whose preferences differ only for the two central ranks, are grouped together at a very low height in Figure 8, while they are grouped at a greater height in Figure 7. Moreover, our classification approaches that of 1 and 8 and separates consumer 3, which have different lower ranks.

### TABLE 4  Example: rankings of six products given by 10 consumers

| Consumer | Rankings | Consumer | Rankings |
|----------|----------|----------|----------|
| 1        | 1 2 3 4 5 6 6 | 1 2 3 6 4 5 | 1 2 3 6 5 4 |
| 2        | 2 1 3 4 5 6 7 | 1 2 3 6 4 5 | 1 2 4 3 5 6 |
| 3        | 1 2 3 4 6 5 8 | 1 2 4 3 5 6 | 1 2 3 4 5 6 |
| 4        | 2 1 3 4 6 5 9 | 3 1 2 4 5 6 | 3 1 2 4 5 6 |
| 5        | 3 2 1 4 5 6 10 | 1 2 3 5 4 6 | 1 2 3 5 4 6 |

### FIGURE 5  Plots of a mixture copula of Gaussian ($\lambda_1 = 0.8$) and Clayton($\lambda_2 = 1$) with $\alpha = 0.5$ on the left side and $\alpha = 0.2$ on the right side
DISCUSSION

Comparison with WRC measure

In conclusion, we suggest using the classical approach when one wishes to assign equal weights to all ranks in the calculation of the distance between the rankings of two subjects. Spearman’s grade correlation coefficient \( \rho _{s} \) using Clayton copula gives greater importance to the top ranks, which emphasizes the similarity of consumers with similar top ranks. Finally, the use of a mixture of Gumbel and Clayton copulae gives greater importance to the top and lower ranks simultaneously.

Figures 9–11 compare each pair of dendrograms and show the changes of position of the subjects using the three different approaches.

5.1 | Comparison with WRC measure

Among the distance-type measures based on rank correlation coefficients, WRC measures are remarkably useful for evaluating the agreement between two rankings with an emphasis on the concordance of top ranks. WRCs are properly used in the literature for evaluating the distance between rankings in hierarchical cluster analysis. In the simplified formula of Spearman’s \( \rho \) (based on square Euclidean distance between rankings), these indices adopt weights that are a function of both ranks. Dancelli et al [5] investigate five existing WRC indices that introduce these weights in an unsimplified formula of Spearman’s \( \rho \) (ie, by using the Pearson’s product-moment correlation index between ranks).

WRC approaches are very useful for emphasizing top ranks. However, WRC carries an important limitation due to the a priori (ie, subjective) selection of related weights, which can affect final results. The copula approach overcomes this limitation because the weights are data driven rather than chosen a priori. Finally, we compare the distances between subjects obtained by our approach with those obtained by applying a WRC measure. The results of a previous study [4,5] lead us to employ the following WRC measure (attributed to Salama and Quade in ref. [14]) for comparison with our proposal:

\[
\rho _{w} = 1 - 2 \frac{\sum_{i=1}^{k} (r_{i} - q_{i})^{2}((1/r_{i}) + (1/q_{i}))}{(k + 1) \sum_{i=1}^{k} [2i - (k + 1)]^{2}/(k - i + 1)}
\]  

(12)

where \( R : r_{1}, r_{2}, \ldots, r_{k} \) and \( Q : q_{1}, q_{2}, \ldots, q_{k} \) are the two rankings.

We applied the WRC measure to the data in Table 4 and visualized the results by dendrogram in Figure 12 and an evaluation of the distance matrix (see Table 8).

Referring to Table 4, let us consider consumers 1, 3, and 8. As expected, since the WRC measure emphasizes top ranks, consumers 1, 3, and 8 are close to each other, which is consistent with Figure 7. If we emphasize the top and lower ranks simultaneously, the preferences of consumers 1 and 8 are more similar than those of 1 and 3. Thus, a classification that emphasizes both the top and the lower ranks approaches that of 1 and 8 and separates consumer 3.

This is not verified by the WRC measure approach; however, it is verified by our approach using a mixture of Clayton and Gumbel copulae. In conclusion, our proposal is a generalization of the several distance measures discussed in this paper and offers a more flexible method to achieve different goals.

6 | DISCUSSION

We have proposed a new measure to evaluate distance in hierarchical cluster analysis for subject expressing their preferences by rankings. In the literature, several distance measures are presented for ranking data, but each approach carries certain limitations. The most important of these distance measures are based on Spearman’s rank correlation coefficient. Such indices assign the same weight to all ranks to determine the distance between the rankings of a pair of subjects. When the aim is to emphasize the agreement on top/lower ranks, weighted distances for ranking data are proposed, by introducing weights in the formula of \( \rho \). These measures are called WRC measures. In general, these approaches evaluate the agreement between two rankings with a priori choice...
TABLE 5  Example: Distance matrix performed by the classical Spearman’s correlation coefficient $\rho$

|   | 1   | 2   | 3   | 4   | 5   | 6   | 7   | 8   | 9   |
|---|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| 2 | 0.042 |     |     |     |     |     |     |     |     |
| 3 | 0.042 | 0.069 |     |     |     |     |     |     |     |
| 4 | 0.069 | 0.042 | 0.042 |     |     |     |     |     |     |
| 5 | 0.125 | 0.097 | 0.153 | 0.125 |     |     |     |     |     |
| 6 | 0.125 | 0.153 | 0.097 | 0.125 | 0.236 |     |     |     |     |
| 7 | 0.097 | 0.125 | 0.125 | 0.153 | 0.208 | 0.042 |     |     |     |
| 8 | 0.042 | 0.069 | 0.069 | 0.097 | 0.208 | 0.208 | 0.181 |     |     |
| 9 | 0.097 | 0.042 | 0.125 | 0.069 | 0.042 | 0.208 | 0.181 | 0.153 |     |
|10 | 0.042 | 0.069 | 0.097 | 0.125 | 0.153 | 0.097 | 0.042 | 0.097 | 0.125|

TABLE 6  Example: Distance matrix performed by Spearman’s grade correlation coefficient $\rho_C$ by Clayton copula

|   | 1   | 2   | 3   | 4   | 5   | 6   | 7   | 8   | 9   |
|---|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| 2 | 0.070 |     |     |     |     |     |     |     |     |
| 3 | 0.014 | 0.070 |     |     |     |     |     |     |     |
| 4 | 0.070 | 0.014 | 0.070 |     |     |     |     |     |     |
| 5 | 0.247 | 0.188 | 0.269 | 0.212 |     |     |     |     |     |
| 6 | 0.064 | 0.155 | 0.050 | 0.131 | 0.331 |     |     |     |     |
| 7 | 0.050 | 0.130 | 0.064 | 0.155 | 0.312 | 0.017 |     |     |     |
| 8 | 0.022 | 0.088 | 0.045 | 0.130 | 0.359 | 0.130 | 0.105 |     |     |
| 9 | 0.188 | 0.026 | 0.220 | 0.058 | 0.022 | 0.274 | 0.255 | 0.255 |     |
|10 | 0.017 | 0.071 | 0.050 | 0.130 | 0.271 | 0.049 | 0.014 | 0.064 | 0.220|

TABLE 7  Example: Distance matrix performed by Spearman’s grade correlation coefficient $\rho_{GC}$ by mixture of Clayton and Gumbel copulae with weight $\alpha = 0.5$

|   | 1   | 2   | 3   | 4   | 5   | 6   | 7   | 8   | 9   |
|---|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| 2 | 0.076 |     |     |     |     |     |     |     |     |
| 3 | 0.009 | 0.077 |     |     |     |     |     |     |     |
| 4 | 0.077 | 0.009 | 0.076 |     |     |     |     |     |     |
| 5 | 0.374 | 0.107 | 0.158 | 0.132 |     |     |     |     |     |
| 6 | 0.023 | 0.088 | 0.027 | 0.080 | 0.376 |     |     |     |     |
| 7 | 0.027 | 0.080 | 0.023 | 0.088 | 0.126 | 0.066 |     |     |     |
| 8 | 0.005 | 0.091 | 0.032 | 0.081 | 0.386 | 0.047 | 0.049 |     |     |
| 9 | 0.107 | 0.054 | 0.132 | 0.041 | 0.076 | 0.113 | 0.111 | 0.292 |     |
|10 | 0.066 | 0.092 | 0.027 | 0.080 | 0.290 | 0.027 | 0.009 | 0.051 | 0.129|

of the weights of the ranks one wishes to emphasize (ie, the top-ranks). The copula approach avoids the a priori selection of these weights and is suitable for emphasizing agreement on the top ranks when they are considered important, on the lower ranks vice versa, or on both top/lower ranks. The proposed distance builds upon the Spearman’s grade correlation coefficient on a copula transformation of rank denoting the level of importance assigned by subjects classifying $k$ objects.

To achieve this aim, we have presented a generalization of this type of distance using mixtures of copulae in order to achieve a more adaptable method. A mixture of copulae allows for the flexible modeling of different types of dependence structures found in data and the consideration of different circumstances in the classification process. For example, by using mixtures of copulae with lower and upper tail dependence, we can emphasize the subjects’ agreement on the extreme ranks (both the top and lower ranks) when they are considered to be important.

Using a mixture of copulae permits us to model particular circumstances, meet different segmentation goals and emphasize unique agreement rankings in the classification process, that is, agreement on the extreme ranks (lower and top ranks), or on the top and central ranks but not the lower ones, or on the central and lower ranks but not the top ones, and so on. The type of dependence structure that one wishes to emphasize drives the selection of an appropriate mixture of copulae. We evaluated the performance of our proposal in hierarchical cluster analysis using simulated data by comparing a classical
FIGURE 9  Compare two dendrograms: Spearman correlation coefficient (on the left) and Spearman grade correlation coefficient by Clayton copula (on the right).

FIGURE 10  Compare two dendrograms: Spearman grade correlation coefficient by mixture of copulae (Clayton and Gumbel copulae with weight $\alpha = 0.5$) (on the left) and Spearman correlation coefficient (on the right).

FIGURE 11  Compare two dendrograms: Spearman grade correlation coefficient by mixture of copulae (Clayton and Gumbel copulae with weight $\alpha = 0.5$) (on the left) and Spearman grade correlation coefficient by Clayton copula (on the right).
TABLE 8 Example: Distance matrix performed by the WRC measure $\rho_{w}$

| d  | 1  | 2  | 3  | 4  | 5  | 6  | 7  | 8  | 9  |
|----|----|----|----|----|----|----|----|----|----|
| 2  | 0.042 |
| 3  | 0.010 | 0.052 |
| 4  | 0.052 | 0.010 | 0.042 |
| 5  | 0.148 | 0.106 | 0.158 | 0.117 |
| 6  | 0.146 | 0.088 | 0.034 | 0.076 | 0.194 |
| 7  | 0.034 | 0.076 | 0.046 | 0.088 | 0.182 | 0.012 |
| 8  | 0.016 | 0.058 | 0.026 | 0.068 | 0.237 | 0.094 | 0.082 |
| 9  | 0.106 | 0.023 | 0.117 | 0.033 | 0.042 | 0.153 | 0.141 | 0.144 |
|10  | 0.012 | 0.054 | 0.034 | 0.076 | 0.160 | 0.034 | 0.010 | 0.044 | 0.119 |

FIGURE 12 Cluster dendrogram with WRC measure

approach, a copula approach (Clayton copula) and a mixture of copulae (Clayton and Gumbel copulae) in order to give more importance to extreme ranks simultaneously. Our analysis of the resulting dendrograms and distance matrices underlines the advantages using a mixture of copulae when the aim is to emphasize particular ranks (ie, the extreme ranks). Our proposal generalizes the traditional method based on the Spearman’s rank correlation coefficient as it is possible to prove that, using a particular copula function (Gaussian copula), the two approaches produce the same classification. A comparison between the traditional approach, an approach using Clayton copula and our approach using a mixture of Clayton and Gumbel copulae was conducted on simulated data to evaluate the performance of our proposal.

Finally, we emphasize that other similarity measures based on copula functions could be considered in Equation (10); for example, the upper and lower tail dependence coefficients as described in ref. [15] could be used. Cluster analysis on the simulated ranking data of Section 5 leads to the same classification by both methods. Obviously, the distances between subjects change, but the compositions of the classification subgroups obtained by the two approaches are equivalent. In addition, we are working with ranking data. In this context, the number $k$ of ranks is typically low. The differences between the two approaches would likely become more evident using continuous classification data, that is, in a financial context.

The indices of tail dependence typically used in the literature are defined in terms of copula behavior along the main diagonal path of $C(u, u)$. However, the tail dependence of copulas can be substantially stronger along other paths [7,8]. For this reason, the authors were unable to capture the right degree of tail dependence. Refs. [7,8] present the decisive role of the diagonal path in measuring tail dependence copulas and propose a method to disregard the diagonal section of copulas when measuring tail dependence. This problem is not crucial to a ranking data framework. In fact, in the evaluation of the agreement between rankings assigned by a pair of subjects, two subjects are in perfect agreement when they assign the same rank to a fixed object. Therefore, we are interested in the fact that perfect agreement, and thus the strongest dependence, is only on the main diagonal of $C(u, u)$. For this reason, we consider only either symmetric Archimedean copula (Clayton, Gumbel, and Gaussian copulae) or a convex combination of these symmetric Archimedean copulae where the main diagonal is the unique path of maximal dependence. This permits unambiguous perfect preference agreement among two subjects and relies on the diagonal dependence path.

ORCID
Andrea Bonanomi https://orcid.org/0000-0003-2857-1430

REFERENCES
1. M. Alvo and P. L. H. Yu, Statistical Methods for Ranking Data, Springer, New York, 2014. MR3308929
2. V. Arakelian and D. Karlis, Clustering dependencies via mixtures of copulas, Commun. Stat.-Simul. Comput. 43 (2014), 1644–1661. MR3215798
3. A. Bonanomi, M. Nai Ruscone, and S. A. Osmetti, Defining subjects distance in hierarchical cluster analysis by copula approach, Qual. Quant. 51 (2017), 859–872.
4. E. Brentari, L. Dancelli, and M. Manisera, Clustering ranking data in market segmentation: a case study on the Italian McDonald’s customers’ preferences, J. Appl. Stat. 43(11) (2016), 1959–1976. MR3569704
5. L. Dancelli, M. Manisera, and M. Vezzoli, On two classes of Weighted Rank Correlation measures deriving from the Spearman’s $\rho$, in Statistical Models for Data Analysis, P. Giudici, S. Ingassia, and M. Vichi, Eds., Springer, Heidelberg, 2013, 107–114.
6. E. W. Frees and E. A. Valdez, *Understanding relationships using copulas*, N Amer. Actuar. J. 2 (1998), 1–25. MR1988432
7. E. Furman, J. Su, and R. Zitikis, *Paths and indices of maximal tail dependence*, ASTIN Bull. 45(3) (2015), 661–678. MR3394069
8. E. Furman et al., *Tail dependence of the Gaussian copula revisited*, Insur. Math. Econ. 69 (2016), 97–103. MR3515880
9. P. H. Lee and P. L. H. Yu, *Mixtures of weighted distance based models for ranking data with applications in political studies*, Comput. Stat. Data Anal. 56 (2012), 2486–2500. MR2910064
10. C. L. Mallows, *Non-null ranking models*, Biometrika 44 (1958), 114–130. MR0087267
11. R. B. Nelsen, *An Introduction to Copulas*, Springer, New York, 2013. MR2197664
12. K. Pearson, *Mathematical contributions to the theory of evolution. VIII. On the correlation of characters not quantitatively measurable*, Proc. R. Soc. Lond. 66(424–433) (1900), 241–244.
13. C. Spearman, *The proof and measurement of association between two things*, Am. J. Psychol. 15 (1904), 77–101.
14. A. Tarsitano, *Weighted rank correlation and hierarchical clustering*, Book of short papers, CLADAG, Parma, 2005.
15. H. Wang, R. Pappadà, F. Durante, and E. Foscolo, *A portfolio diversification strategy via tail dependence clustering*, in International Conference on Soft Methods in Probability and Statistics, Springer, Cham, 2016, 511–518.

**How to cite this article:** Bonanomi A, Nai Ruscone M, Osmetti SA. Dissimilarity measure for ranking data via mixture of copulae. *Stat Anal Data Min: The ASA Data Sci Journal* 2019;12:412–425. https://doi.org/10.1002/sam.11402