Ca' Foscari
University
of Venice

**Department
of Economics**

Working Paper

Andrea Pastore and Stefano F.
Tonellato

On the comparison of model-
based clustering solutions

ISSN: 1827-3580
No. 05/WP/2013

On the comparison of model-based clustering solutions

A. Pastore and S. F. Tonellato

Department of Economics, University Ca' Foscari of Venice

March 8, 2013

Abstract

In this paper we propose a new similarity index, which can be used to compare model-based clustering solutions. We define also an adjusted-for-chance version, although we advise that, whenever feasible, bootstrap replications should be preferred to chance-corrected similarity indices. We describe the properties of the proposed index and of its chance-corrected version. Finally, we present some applications on simulated and real data.

Keywords similarity indices, Rand index, mixture models, bootstrap.

JEL Codes C39, C46, C49

Address for correspondence:

Stefano Tonellato

Department of Economics
Ca' Foscari University of Venice
Cannaregio 873, Fondamenta S. Giobbe
30121 Venezia - Italy
Phone: (+39) 041 2347422
Fax: (+39) 041 2347444
e-mail: stone@unive.it

This Working Paper is published under the auspices of the Department of Economics of the Ca' Foscari University of Venice. Opinions expressed herein are those of the authors and not those of the Department. The Working Paper series is designed to divulge preliminary or incomplete work, circulated to favour discussion and comments. Citation of this paper should consider its provisional character.

1 Introduction

Many indices have been introduced in the literature in order to compare two partitions of the same set of objects, determined by different clustering procedures. The most popular ones among them are, perhaps, the Rand index (Rand, 1971) and its adjusted-for-chance version (Hubert and Arabie, 1985; Steinley, 2004). These indices have been used within consensus clustering methodologies (Krieger and Green, 1999), in the selection of the number of groups (Chae et al., 2006; Albatineh and Niewiadomska-Bugaj, 2011b) and in variable selection for clustering (Carmone et al., 1999; Brusco and Cradit, 2001). The adjusted Rand index is particularly important, since it is equivalent to many other chance-adjusted indices, as pointed out by Albatineh et al. (2006), Albatineh and Niewiadomska-Bugaj (2011a), Warrens (2008a,b). Such indices are generally based on the number of pairs of objects placed in the same or in different groups by both procedures under comparison. Hence, they depend exclusively on the partitions and do not take into account the degree of uncertainty of clustering solutions, even when it can be quantified. This happens with various model-based clustering methods (McLachlan and Basford, 1988; Fraley and Raftery, 2002; James and Sugar, 2003) which provide a matrix of estimated membership probabilities, i.e. the probabilities that each of n individuals belongs to each of g groups. In the following, we shall call probabilistic clustering solutions, any clustering solution providing such estimates. When the aim is to compare two probabilistic solutions, the use of indices that depend only on the associated partitions is not appropriate, since the uncertainty in the objects-to-group allocation is completely lost.

Since we shall consider mainly clustering based on mixtures models, we recall briefly some theory underlying this methodology. Let $\mathcal{U} = \{u_1, \dots, u_n\}$ be a set of n objects and x_i is the measurement of a d -dimensional variable, $\mathbf{X}$, taken on u_i . Assume we want to cluster the elements of $\mathcal{U}$ in homogeneous groups (homogeneity is intended, of course, in terms of the behaviour of $\mathbf{X}$). When $\mathcal{U}$ is assumed to be an iid sample of size n , we can assume that the observations $\mathbf{x}_i$, $i = 1, \dots, n$, come from n iid random variables, $\mathbf{X}_i$, that are modelled as g -component mixtures, with pdf

$$f(\mathbf{x}; \boldsymbol{\pi}, \boldsymbol{\theta}) = \prod_{k=1}^g \pi_k f_k(\mathbf{x}; \boldsymbol{\theta}_k), \quad (1)$$

where $\mathbf{x} \in \mathbb{R}^d$, $d \geq 1$, and $f_k(\mathbf{x}, \boldsymbol{\theta}_k)$, $k = 1, \dots, g$ are probability density functions.

For the sake of simplicity, assume that $\boldsymbol{\pi}$ and $\boldsymbol{\theta}$ are known, with $\boldsymbol{\theta} = [\boldsymbol{\theta}'_1, \dots, \boldsymbol{\theta}'_g]'$, $\boldsymbol{\pi} = [\pi_1, \dots, \pi_g]$ such that $\pi_k \geq 0$, $k = 1, \dots, g$, and $\sum_{k=1}^g \pi_k = 1$. The k -th mixture component, $f_k(\mathbf{x}; \boldsymbol{\theta}_k)$, represents the probabilistic model for the k -th cluster in the population, whereas π_k can be seen as the population proportion of individuals in the k -th cluster. When $\mathcal{U}$ is assumed to be an iid sample of size n from (1), a straightforward application of Bayes theorem leads to the computation of $\mathbf{p}_i$, $i = 1, \dots, n$, with

$$p_{i,k} = \frac{\pi_k f_k(\mathbf{x}; \boldsymbol{\theta}_k)}{\sum_{k=1}^g \pi_k f_k(\mathbf{x}; \boldsymbol{\theta}_k)}, \quad k = 1, \dots, g. \quad (2)$$

A probabilistic solution, A , can then be defined as a set of n vectors such that the elements of the i -th one, $i = 1, \dots, n$, are the membership probabilities referred to the i -th individual in $\mathcal{U}$.

Given a probabilistic clustering solution, say A , for sample $\mathcal{U}$, a partition can be obtained following the maximum a posteriori (MAP) rule, i.e. by assigning each individual to the group with the highest membership probability. In so doing, a partition of $\mathcal{U}$, $\mathcal{C} = \{C_1, \dots, C_g\}$ can be obtained. Let $z_{i,A} \in \{1, \dots, g\}$ denote the label of the cluster $C_{z_{i,A}} \in \mathcal{C}$. For each u_i , we can define a g -dimensional vector $\bar{\mathbf{p}}_i$, whose elements are all equal to zero except the one in position $z_{i,A}$, which is equal to 1. Let $\bar{A} = \{\bar{\mathbf{p}}_i, i = 1, \dots, n\}$. Clearly, $\bar{A}$ represents the partition associated with solution A . It is worth to note that a deterministic partition can always be represented as a degenerate probabilistic solution where each individual is assigned, with probability 1, to one among g groups, i.e. $A = \bar{A}$.

In this paper, we propose an index capable to deal with partitions provided by both model-based and model-free clustering methods. Notice that, even though all parameters in (1) were known, all the $\mathbf{p}_i$'s and $\bar{\mathbf{p}}_i$'s would be random vectors. This fact makes the use of descriptive similarity indices rather arguable, unless their randomness is explicitly taken into account.

The paper is organized as follows: section 2 reviews the definitions and properties of Rand and adjusted Rand index; in section 3 we define a new similarity index and show its properties; in section 4 we present some simulation studies and some applications to real data; finally, in section 5, we draw some conclusions.

2 Rand and adjusted Rand indices

In this section we shall revise two of the most commonly used indices for the comparison of partitions: the Rand index and its corrected for chance version.

A general scheme for the definition of the Rand index is the following. Let us consider all the $m = n(n - 1)/2$ pairs we can form using the objects in $\mathcal{U}$ and two partitions $\bar{A}$ and $\bar{B}$. Furthermore, let a, b, c, d be non negative real numbers such that $a + b + c + d = 1$. The m pairs can be classified as follows:

- i. $m \cdot a$ pairs are in the same group in both $\bar{A}$ and $\bar{B}$;
- ii. $m \cdot b$ pairs are in the same group in $\bar{A}$ but in different groups in $\bar{B}$;
- iii. $m \cdot c$ pairs are in different groups in $\bar{A}$ but in the same group in $\bar{B}$;
- iv. $m \cdot d$ pairs are in different groups in both $\bar{A}$ and $\bar{B}$.

Table 1 reports the similarity table which summarizes this classification. The Rand index is simply defined as the sum of two relative frequencies:

$$R = R(\bar{A}, \bar{B}) = a + d. \quad (3)$$

From (3) it is clear that R is the relative frequency of pairs from $\mathcal{U}$ that are classified similarly by the two partitions. The Rand index is a bounded symmetric similarity index since $0 \leq R \leq 1$, and $R = 1$ if and only if $\bar{A} = \bar{B}$. A value of R close to 1 indicates that $\bar{A}$ and $\bar{B}$ are very similar partitions, whereas a value close to 0 means that $\bar{A}$ and $\bar{B}$ differ noticeably. Notice that, if $\mathcal{U}$ is a random sample, the Rand index can be seen as an estimate of the probability that two randomly selected individuals are allocated to the same group or to different groups by both $\bar{A}$ and $\bar{B}$.

It is also to note that the frequencies appearing in table 1 can be computed from the representation of the partitions as degenerate probabilistic clustering solutions. Consider two partitions of $\mathcal{U}$, $\bar{A}$ and $\bar{B}$, with (degenerate) membership probability vectors $\bar{\mathbf{p}}_i$ and $\bar{\mathbf{q}}_i$ respectively, $i = 1, \dots, n$. Then we obtain:

$$a = \frac{1}{m} \sum_{i>j} (\bar{\mathbf{p}}'_i \bar{\mathbf{p}}_j) (\bar{\mathbf{q}}'_i \bar{\mathbf{q}}_j), \quad (4)$$

where the sum is extended to all m combinations of pairs from $\mathcal{U}$. Moreover:

$$\begin{aligned} b &= \frac{1}{m} \sum_{i>j} (\bar{\mathbf{p}}'_i \bar{\mathbf{p}}_j) (1 - \bar{\mathbf{q}}'_i \bar{\mathbf{q}}_j) \\ c &= \frac{1}{m} \sum_{i>j} (1 - \bar{\mathbf{p}}'_i \bar{\mathbf{p}}_j) (\bar{\mathbf{q}}'_i \bar{\mathbf{q}}_j) \\ d &= \frac{1}{m} \sum_{i>j} (1 - \bar{\mathbf{p}}'_i \bar{\mathbf{p}}_j) (1 - \bar{\mathbf{q}}'_i \bar{\mathbf{q}}_j). \end{aligned} \tag{5}$$

The interpretation of R might be misleading, since a positive value might be observed due to a completely random agreement between two partitions. For this reason, Hubert and Arabie (1985) proposed a modification of the Rand index, the so called correction for chance, which gives the adjusted Rand index:

$$R_a = \frac{R - E(R)}{1 - E(R)},$$

where $E(R)$ is the expected value of R conditionally on the number of clusters and the number of individuals in each cluster of both partitions. The idea is to provide an index such that the value 0 can be interpreted as the level of similarity of two randomly selected partitions.

3 Comparison of model based clustering solutions

The comparison of two probabilistic clustering solutions is commonly done by comparing their associated partitions using the indices described in the previous section (Yeung et al., 2001; Scrucca, 2010). Such practice is rather arguable, since it does not take into account the uncertainty in the objects-to-group allocation, which is a relevant feature of model based clustering.

3.1 A Rand-like index for the comparison of model based solutions

In this section we propose an index for the comparison of clustering solutions, assuming, for the moment that all the parameters in (1) are known. The main idea is

to define a similarity table, for two probabilistic clustering solutions, analogous to Table 1. Let us consider the events

$$J_A = \{\text{two randomly chosen individuals belong to the same group in } A\}$$

and

$$J_B = \{\text{two randomly chosen individuals belong to the same group in } B\}$$

each referred to a generic pair of individuals randomly selected from the population. Clearly, the probabilities of these events are related with the distribution of the membership probability vectors $\mathbf{p}_i$ and $\mathbf{q}_i$, $i = 1, \dots, n$, which characterise A and B .

Table 2 is in some sense analogous to Table 1 in the context of probabilistic clustering. The probabilities appearing there can be estimated as follows. Let A be a clustering solution and $S_{A,k}$, $k = 1, \dots, g$, be the clusters induced by A in the sample. We shall assume that the following condition holds:

C1. the events $\{u_i \in S_{A,k}\}$ and $\{u_j \in S_{A,k'}\}$ are independent for all i, j , k and k' .

Consider the event

$$U_{ij} = \{\text{the selected pair is } (u_i, u_j)\}.$$

From condition C1. it is easy to verify that $\Pr(J_A|U_{ij}) = \mathbf{p}'_i \mathbf{p}_j$. Hence,

$$\Pr(J_A) = \mathbb{E}[\Pr(J_A|U_{ij})] = \boldsymbol{\pi}' \boldsymbol{\pi} \quad \text{and} \quad \Pr(\bar{J}_A) = \mathbb{E}[1 - \Pr(J_A|U_{ij})] = 1 - \boldsymbol{\pi}' \boldsymbol{\pi},$$

with $\boldsymbol{\pi}$ defined as in (1). Although we are assuming that all parameters in (1) are known, for reasons that will be clear later we define the following unbiased and consistent estimators for $\Pr(J_A)$, $\Pr(J_B)$, $\Pr(\bar{J}_B)$ and $\Pr(\bar{J}_A)$:

$$\widehat{\Pr}(J_A) = \frac{1}{m} \sum_{i>j} \mathbf{p}'_i \mathbf{p}_j \tag{6}$$

$$\widehat{\Pr}(J_B) = \frac{1}{m} \sum_{i>j} \mathbf{q}'_i \mathbf{q}_j \tag{7}$$

$$\widehat{\Pr}(\bar{J}_A) = \frac{1}{m} \sum_{i>j} (1 - \mathbf{p}'_i \mathbf{p}_j) \tag{8}$$

$$\widehat{\Pr}(\bar{J}_B) = \frac{1}{m} \sum_{i>j} (1 - \mathbf{q}'_i \mathbf{q}_j) \tag{9}$$

The probability of joint events in Table 2 can be estimated introducing the following conditional independence assumption:

$$\text{C2. } J_A \perp J_B | U_{ij} \quad \forall i, j.$$

This assumption is correct when A and B are two solutions based on two random vectors, say $\mathbf{X}$ and $\mathbf{Y}$. Notice that, since we are assuming a conditional independence between J_A and J_B , it is not necessary that $\mathbf{X}$ and $\mathbf{Y}$ are themselves independent. Assumption C2 seems reasonable also when we compare two clustering algorithms for which we cannot conjecture any kind of interaction.

From assumption C2, $\Pr(J_A \cap J_B) = \mathbb{E} [\Pr(J_A | U_{ij}) \Pr(J_B | U_{ij})]$. It follows immediately that an unbiased estimate of $\Pr(J_A \cap J_B)$ is

$$\widehat{\Pr}(J_A \cap J_B) = \frac{1}{m} \sum_{i>j} \mathbf{p}'_i \mathbf{p}_j \mathbf{q}'_i \mathbf{q}_j \quad (10)$$

and, similarly:

$$\widehat{\Pr}(\bar{J}_A \cap J_B) = \frac{1}{m} \sum_{i>j} (1 - \mathbf{p}'_i \mathbf{p}_j) \mathbf{q}'_i \mathbf{q}_j \quad (11)$$

$$\widehat{\Pr}(J_A \cap \bar{J}_B) = \frac{1}{m} \sum_{i>j} \mathbf{p}'_i \mathbf{p}_j (1 - \mathbf{q}'_i \mathbf{q}_j) \quad (12)$$

$$\widehat{\Pr}(\bar{J}_A \cap \bar{J}_B) = \frac{1}{m} \sum_{i>j} (1 - \mathbf{p}'_i \mathbf{p}_j) (1 - \mathbf{q}'_i \mathbf{q}_j) \quad (13)$$

Now, we can introduce a probabilistic analogue of the Rand index defined as follows.

Definition 1. *Let A and B be two clustering solutions, then $I(A, B)$ is defined as:*

$$I(A, B) = \widehat{\Pr}(J_A \cap J_B) + \widehat{\Pr}(\bar{J}_A \cap \bar{J}_B). \quad (14)$$

An immediate consequence of Definition 1 is that $0 \leq I(A, B) \leq 1$. It is also straightforward to verify that, if $\bar{A}$ and $\bar{B}$ are two partitions, i.e. two solutions such that, respectively, $\mathbf{p}_i = \bar{\mathbf{p}}_i$ and $\mathbf{q}_i = \bar{\mathbf{q}}_i$, for $i = 1, \dots, n$, then

$$R = a + d = \widehat{\Pr}(J_A \cap J_B) + \widehat{\Pr}(\bar{J}_A \cap \bar{J}_B) = I(\bar{A}, \bar{B}). \quad (15)$$

It is also easy to verify that, similarly to Rand index, $I(A, B) = 1$ if and only if $A = B = \bar{A}$, i.e. $\mathbf{p}_i = \mathbf{q}_i = \bar{\mathbf{p}}_i$, $i = 1, \dots, n$, and $I(A, B) = 0$ if and only if $\widehat{\text{Pr}}(J_A \cap J_B) = \widehat{\text{Pr}}(\bar{J}_A \cap \bar{J}_B) = 0$. Moreover, from Definition 1, it follows immediately that $I(A, B) = I(B, A)$.

When $A = B$, in the sense that $\mathbf{p}_i = \mathbf{q}_i$ for all i , it generally happens that $I(A, B) \neq 1$. This means that $I(A, B)$ is not a similarity index. The following lemma will allow us to define a bounded symmetric similarity index also under the latter conditions.

Lemma 1. *Let A and B be two clustering solutions and*

$$U(A, B) = \frac{1}{m} \left(\sqrt{\sum_{i>j} (\mathbf{p}'_i \mathbf{p}_j)^2 \sum_{i>j} (\mathbf{q}'_i \mathbf{q}_j)^2} + \sqrt{\sum_{i>j} (1 - \mathbf{p}'_i \mathbf{p}_j)^2 \sum_{i>j} (1 - \mathbf{q}'_i \mathbf{q}_j)^2} \right), \quad (16)$$

then $I(A, B) \leq U(A, B)$.

Proof. Let $\tilde{\mathbf{p}}, \tilde{\mathbf{q}} \in \mathbb{R}^m$, $m = n(n-1)/2$, such that their u -th elements are $\tilde{p}_u = \mathbf{p}'_i \mathbf{p}_j$ and $\tilde{q}_u = \mathbf{q}'_i \mathbf{q}_j$, $u = n(j-1) - j(j+1)/2 + i$, $i = 2, \dots, n$ and $j < i$. Using the Cauchy-Schwartz inequality, we obtain:

$$\begin{aligned} \widehat{\text{Pr}}(J_A \cap J_B) &= \frac{1}{m} \sum_{i>j} \mathbf{p}'_i \mathbf{p}_j \mathbf{q}'_i \mathbf{q}_j = \frac{1}{m} \sum_{u=1}^m \tilde{p}_u \tilde{q}_u \\ &\leq \frac{1}{m} \sqrt{\sum_{u=1}^m \tilde{p}_u^2 \sum_{u=1}^m \tilde{q}_u^2} = \frac{1}{m} \sqrt{\sum_{i>j} (\mathbf{p}'_i \mathbf{p}_j)^2 \sum_{i>j} (\mathbf{q}'_i \mathbf{q}_j)^2} \end{aligned}$$

and, similarly,

$$\begin{aligned} \widehat{\text{Pr}}(\bar{J}_A \cap \bar{J}_B) &= \frac{1}{m} \sum_{i>j} (1 - \mathbf{p}'_i \mathbf{p}_j)(1 - \mathbf{q}'_i \mathbf{q}_j) \\ &\leq \frac{1}{m} \sqrt{\sum_{i>j} (1 - \mathbf{p}'_i \mathbf{p}_j)^2 \sum_{i>j} (1 - \mathbf{q}'_i \mathbf{q}_j)^2}. \end{aligned}$$

Hence, the result follows. □

We now define the following similarity index.

Definition 2. *The p -Rand index is defined as*

$$\tilde{I}(A, B) = \frac{I(A, B)}{U(A, B)}. \quad (17)$$

The following proposition shows that $\tilde{I}(A, B)$ is a bounded symmetric similarity index.

Proposition 1. *Let A and B be two clustering solutions and assume that conditions C1. and C2. hold. Then, $\tilde{I}(A, B)$ is a bounded symmetric similarity index for A and B . Moreover, if $A = \bar{A}$ and $B = \bar{B}$, then:*

$$\tilde{I}(\bar{A}, \bar{B}) = \frac{a + d}{\sqrt{(a + b)(a + c)} + \sqrt{(d + b)(d + c)}} \quad (18)$$

Proof. The first part of the proposition follows from Definition 1 and Lemma 1.

To prove (18), recall that $\bar{\mathbf{p}}'_i \bar{\mathbf{p}}_j$ is equal to 0 or 1, so $(\bar{\mathbf{p}}'_i \bar{\mathbf{p}}_j)^2 = \bar{\mathbf{p}}'_i \bar{\mathbf{p}}_j$. Hence, from (4) and (5), it follows that:

$$U(\bar{A}, \bar{B}) = \sqrt{(a + b)(a + c)} + \sqrt{(d + b)(d + c)}. \quad (19)$$

Then, (18) follows from (15), (17) and (19) $\square$

3.2 Correction for chance

In the correction of the Rand index proposed by Hubert and Arabie (1985), a generalised hypergeometric distribution is assumed as the model for randomness, i.e. it is assumed that partitions $\bar{A}$ and $\bar{B}$ are picked at random among those partitions sharing the same number of groups and the same number of objects in each group that have been identified for A and B . When model-based solutions are considered, a null model can be defined similarly, assuming that A and B are picked at random among all those pairs of solutions identified respectively by $\mathbf{p}_{\sigma_i}$ and $\mathbf{q}_{\sigma_i}$, $i = 1, \dots, n$, where $\{\sigma_1, \dots, \sigma_n\}$ represents an arbitrary permutation of $\{1, \dots, n\}$. Assuming such null model, the estimates of the marginal probabilities in Table 2 are given by (6)-(9), while the joint probabilities can be estimated using the corresponding marginal estimates, exploiting the assumption of independence between the solutions. Moreover,

$U(A, B)$ is invariant, conditionally on the $\mathbf{p}_i$'s and $\mathbf{q}_i$'s. Then, an estimate of the expected value of p -Rand index $\tilde{I}(A, B)$, under the null model is:

$$\hat{E}[\tilde{I}(A, B)] = \frac{\widehat{\Pr}(J_A) \cdot \widehat{\Pr}(J_B) + (1 - \widehat{\Pr}(J_A)) \cdot (1 - \widehat{\Pr}(J_B))}{U(A, B)}$$

Hence, the adjusted p -Rand index can be defined as follows.

Definition 3. *The adjusted p -Rand index is defined as:*

$$\tilde{I}_a(A, B) = \frac{\tilde{I}(A, B) - \hat{E}[\tilde{I}(A, B)]}{1 - \hat{E}[\tilde{I}(A, B)]}. \quad (20)$$

It is worth to note that Warrens (2008b) showed that

$$R_a = \frac{R - E(R)}{1 - E(R)} = \frac{2(ab - bc)}{(a + b)(a + c) + (b + d)(c + d)}.$$

Hence, the adjusted p -Rand index is not equivalent to the adjusted Rand index.

For the sake of simplicity, hereafter we shall denote $I(A, B)$, $\tilde{I}(A, B)$ and $\tilde{I}_a(A, B)$ as I , $\tilde{I}$ and $\tilde{I}_a$.

3.3 A bootstrap alternative to correction for chance

The use of chance corrected indices rests on some arguable and restrictive assumptions concerning the null model, which we have recalled in the previous subsection. In dealing with the comparison of model based solutions, however, we might consider the use of a naive bootstrap test, in order to assess whether the agreement between A and B is random or not. Assume we collected n observations $\mathbf{z}_i = (\mathbf{x}_i, \mathbf{y}_i)$ from $(\mathbf{X}, \mathbf{Y})$. We might build L n -dimensional samples $(\mathbf{z}_1^l, \dots, \mathbf{z}_n^l)$, $l = 1, \dots, L$, with $\mathbf{z}_i^l = (\mathbf{x}_i^l, \mathbf{y}_i^l)'$ and $i = 1, \dots, n$, by resampling $\mathbf{x}_i^l$ from $(\mathbf{x}_1, \dots, \mathbf{x}_n)$ and $\mathbf{y}_i^l$ from $(\mathbf{y}_1, \dots, \mathbf{y}_n)$ independently and computing $\tilde{I}^l$. The bootstrap sample $(\tilde{I}^1, \dots, \tilde{I}^L)$ mimics the behaviour of $\tilde{I}$ under H_0 , the hypothesis of random agreement between A and B (which means that A and B are independent solutions and their observed agreement is only due to chance). Let q_α^{boot} denote the bootstrap estimate of the α -quantile of $\tilde{I}$ under H_0 . Such null hypothesis might be compared with the alternative H_1 ,

H_1 : A and B agree and their agreement is not random (i.e. their observed agreement is not due to chance). The null hypothesis should be rejected whenever $\tilde{I}^{obs} > q_{1-\alpha}^{boot}$.

Some examples of the use of bootstrap will be given in next section.

4 Examples and applications

The main advantage of the indices $\tilde{I}$ and $\tilde{I}_a$ is that they allow us to take into account the uncertainty in objects-to-group allocation when we compare alternative solutions. In this section, we shall highlight some implications of this feature, by considering some simulations and applications to real data.

4.1 Simulation experiments

4.1.1 Comparison of a probabilistic solution with its associated partition

Let A be a probabilistic clustering solution for the set $\mathcal{U} = \{u_1, \dots, u_n\}$ and let $\bar{A}$ be its associated partition, as defined in section 1. The comparison of A with $\bar{A}$ through $\tilde{I}$ or $\tilde{I}_a$ would provide a misleading informations, since condition C2 in Section 3 does not hold. However, we might interpret the values of such indices as the level of similarity between A and a partition A^* identical to $\bar{A}$ but independent of A (think of A^* as a classification provided by an oracle and accidentally equal to $\bar{A}$). In this simulation study, we shall consider the behaviour of the p -Rand index on a sample of 500 iid random pairs (A_s, A_s^*) , $s = 1, \dots, 500$. A random probabilistic solution A_s is obtained by generating n g -dimensional independent vectors $\mathbf{p}_{i,s}$ as follows. The maximum value of each $\mathbf{p}_{i,s}$ is fixed at an arbitrary value, α , independent of i and s . The position of such maximum is allocated randomly to one of the g elements of $\mathbf{p}_{i,s}$. When $g > 2$, the remaining elements of $\mathbf{p}_{i,s}$ are generated from a Dirichlet distribution $\mathcal{D}_{g-1} \left(\frac{1000}{g-1}, \dots, \frac{1000}{g-1} \right)$ and rescaled by a factor equal to $1 - \tau$, with $\tau \in (0, 1)$. Simulation experiments have been performed with $n = 150$, $\frac{1}{g} < \tau \leq .9$ (with steps equal to 0.1), $g = 2, \dots, 9$. For each simulation, the mean and the standard deviation of $\tilde{I}_a$ have been computed in 500 replications. The results are given in Table 3. Apparently, as g increases and α decreases, i.e. when the uncertainty of individual-to-group allocation increases, the

agreement between A and A^* decreases. It is also worth to notice that the value taken by $\tilde{I}_a$ is not particularly informative about the randomness of the agreement between two clusterings. In Table 3, we can see that, on average, $\tilde{I}_a = 0.096$, when $\tau = 0.6$ and $g = 2$. In such circumstance the value of the index seems to denote a low agreement between the two solutions, but its distribution (s.d.=0.002) actually says the opposite: the expected value of I_a is significantly apart from zero. We can then claim that:

- a) a value of $\tilde{I}_a$ next to 0 might mean that there is a high uncertainty in the probabilistic solution;
- b) the randomness of the agreement between two solutions should be assessed by considering the probability distribution of $\tilde{I}_a$ (or of any other index used for this purpose, being it corrected for chance or not), not its sample value only.

Table 3 also shows that, if g increases, then $\tilde{I}_a$ increases for any fixed value of τ , the value of $\tilde{I}_a$ increases with g , which seems to indicate that the agreement between A and A^* might be simply due to the fact that pairs of individuals are classified similarly in A and A^* just because of the high number of clusters. Again, this limitation would be removed if we could consider the distribution of $\tilde{I}_a$ and not simply the sample value of this statistic. Notice also that, for any fixed g , $\tilde{I}_a$ decreases as τ decreases. This is a direct consequence of the fact that, $\tilde{I}_a$ takes explicitly into account the level of uncertainty of A . Similar conclusions can be drawn by examining Figure 1. The adjusted Rand index $R_a(\bar{A}, A^*)$, instead, would always be equal to one, since $\bar{A} = A^*$ ($\bar{A}$ denotes the MAP partition) in all the simulations synthesized in Table 3.

4.1.2 Comparison of two probabilistic solutions

In the following simulation studies, we shall compare the clustering solutions obtained by fitting mixture models like (1) to different variables measured on the same individuals. We shall also emphasize the usefulness of bootstrap in such situations.

Clusterings based on two independent bivariate normal mixtures Let $\mathbf{X}$ and $\mathbf{Y}$ be two independent and identically distributed bivariate random vectors, with pdf

$$f(\cdot) = \sum_{k=1}^2 \frac{1}{2} N(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}), \quad (21)$$

with

$$\boldsymbol{\mu}_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \boldsymbol{\mu}_2 = \begin{bmatrix} 2.5 \\ 2.5 \end{bmatrix} \text{ and } \boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix} \quad (22)$$

We simulated 2000 samples of size $n = 200$. On each sample we fitted a two-component Gaussian mixture model for $\mathbf{X}$ and $\mathbf{Y}$ separately and obtained two clustering solutions (Fraley and Raftery, 2002, 2006). For each solution, the $\mathbf{p}_i$'s were computed by plugging the estimates of unknown parameters in (2). The black curves in Fig. 2(a) and 2(b) represent the estimated pdf and the empirical cumulative distribution function of $\tilde{I}$.

If we knew the probability distribution of $\tilde{I}$, we might compare the sample value of p -Rand index with some critical values in order to assess whether the agreement between two clustering solutions is purely random or not. In other words, we might build a “significance test” for the null hypothesis of purely random agreement between two solutions. Unfortunately, the distribution of $\tilde{I}$ is not known under such null hypothesis. We then tried to explore the possibility of exploiting the non parametric bootstrap in order to approximate such null distribution. We simulated a sample of size 200 from (21), using the parametrization in (22), and we obtained the value $\tilde{I} = 0.597$. We then resampled $\mathbf{X}$ and $\mathbf{Y}$ independently, obtaining a bootstrap sample $\tilde{I}^{(l)}$, $l = 1, \dots, 2000$. The red curves in Fig. 2(a) and 2(b) represent the estimated pdf and the empirical cumulative distribution based on such bootstrap sample. From Fig. 2, we can see that there is a substantial agreement between the estimated pdf (ecdf) obtained from 2000 independent realisations of $\tilde{I}$ and the one obtained via bootstrap. We can also notice that at $\tilde{I}$ the value of the bootstrap estimate of the pdf of $\tilde{I}$ is rather high, which supports the null hypothesis. An approximate α -size test might be built using $\tilde{I}$ as test statistic and $[q_{\alpha/2}^{boot}, q_{1-\alpha/2}^{boot}]$ as acceptance region, where $[q_{\alpha/2}^{boot}$ and $q_{1-\alpha/2}^{boot}]$ are the bootstrap estimates of the $\alpha/2$ and $1 - \alpha/2$ quantiles of $\tilde{I}$. This means that the comparison of two solutions based on the p -Rand index and the bootstrap seems much more informative than the comparison based on any chance-corrected index. Figure 3 shows the results obtained by the same experiment under the assumption that the number of components in (21) is unknown (it was estimated by selecting the best model according with BIC -statistic). We can notice that the estimated pdf (black curve in Fig. 3(a)) has two minor modes in correspondence of about 0.8 and 1, which are due to the uncertainty in the number of components. Such modes are not caught exactly by the bootstrap estimate of the pdf (ecdf), but precision might be improved by increasing the size of the bootstrap sample. Such imprecision, however, does not seem to dramatically

affect the estimates of the quantiles of interest.

A completely analogous simulation experiment has been run with $\mathbf{X}$ and $\mathbf{Y}$ independent, with pdf

$$f(\cdot) = \sum_{k=1}^5 \frac{1}{5} N(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \quad (23)$$

with

$$\begin{aligned} \boldsymbol{\mu}_1 &= \begin{bmatrix} 0 \\ 0 \end{bmatrix}, & \boldsymbol{\mu}_2 &= \begin{bmatrix} 2.5 \\ 2.5 \end{bmatrix}, & \boldsymbol{\mu}_3 &= \begin{bmatrix} 5 \\ 0 \end{bmatrix}, \\ \boldsymbol{\mu}_4 &= \begin{bmatrix} 0 \\ 5 \end{bmatrix}, & \boldsymbol{\mu}_5 &= \begin{bmatrix} 5 \\ 5 \end{bmatrix}, & \boldsymbol{\Sigma} &= \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}. \end{aligned} \quad (24)$$

Figure 4 shows the estimated pdf (ecdf) based on the determinations of $\tilde{I}$ on 2000 independent samples of size 200. The same figure shows (red curves) the bootstrap estimated pdf (ecdf) obtained from a bootstrap sample of size 2000, generated by resampling from a sample of size 200. Here it is worth to note the pronounced multimodality of the estimated pdf of $\tilde{I}$, due to the higher number of groups, which was again estimated by selecting the optimal model through *BIC*. Notice also that the bootstrap estimate of the pdf of $\tilde{I}$ is very close to the estimate obtained from 2000 independent replicates of $\tilde{I}$.

Probabilistic solutions based on perturbed data Here we are interested in the behaviour of $\tilde{I}$ when we compare the solutions obtained from original and perturbed data in two simulation experiments.

In the first experiment we generated simulated data, $\mathbf{X}$, from (21), with the parameterisation given in (22). Such data were perturbed by adding a bivariate Gaussian noise, with zero mean and covariance matrix $\sigma_\epsilon^2 \mathbf{I}$, with σ_ϵ ranging from 0.2 to 13.0. For each value of σ_ϵ , we generated then 2000 samples of size 200 from $(\mathbf{X}, \mathbf{Y})$, where $\mathbf{Y} = \mathbf{X} + \boldsymbol{\epsilon}$. For each simulated sample, we computed $\tilde{I}$. Fig. 5 reports the boxplots of $\tilde{I}$ for each batch of simulations corresponding to a specific value of σ_ϵ . Those boxplots show that the distribution of $\tilde{I}$ is strongly concentrated next to 1 when perturbations are small, with an asymmetric and multimodal distribution. The variance of $\tilde{I}$ increases with σ_ϵ , whereas multimodality and asymmetry decrease. The distribution of $\tilde{I}$ becomes neatly unimodal and symmetrical only for high values of σ_ϵ , corresponding to substantially uncorrelated $\mathbf{X}$ and $\mathbf{Y}$.

We ran an analogous experiment generating $\mathbf{X}$ from (23), parametrised as in (24). Perturbed data $\mathbf{Y}$ were obtained as above, with σ_ϵ ranging from 0.25 to 5.00. In Fig. 6 we can notice a faster change in the distribution of $\tilde{I}$ for increasing σ_ϵ , with persisting asymmetry and multimodality, due to the higher number of components in the mixture.

In order to have an idea of bootstrap performance in such a situation, we performed a different experiment. For each value of σ_ϵ , we generated a single sample of size 200 from $(\mathbf{X}, \mathbf{Y})$ as described above and, for each such sample, obtained 5000 bootstrap replications of $\tilde{I}$. Fig. 7 shows the sample values of $\tilde{I}$ and some percentiles of the bootstrap samples in the case of perturbed bivariate 2 and 5 component mixtures. These naive bootstrap tests show a decreasing evidence of agreement between clusterings for increasing values of σ_ϵ .

4.2 Applications to real data

4.2.1 Classification of monitoring stations in Venice Lagoon

Pastres et al. (2011) presented an application of model based functional data clustering in order to classify thirty monitoring stations in Venice Lagoon. Five classifications were obtained, based on five distinct variables, i.e. salinity and phosphorus, nitrate, ammonium and chlorophyll-a concentrations. Table 4 shows the solutions based on the five trophic variables. Letters H , L and I represent group labels and correspond to High, Low and Intermediate level of the variable. This terminology was adopted because the different groups were characterised by mean curves with similar shape, located at different levels (see Pastres et al. (2011) for further details). Looking at Table 4 we can notice that all the trophic variables identify two groups, except phosphorus, which provides a solution with three groups. Each solution provides an estimated membership probability distribution for each site. Such solutions were deemed highly uncertain whenever the minimum estimated membership probability resulted greater than or equal to 0.1 and they are highlighted in Table 4. It is worth to notice that the solution based on ammonium is the one giving rise to the highest uncertainty in site-to-group allocation, since 8 (over 30) estimated membership probability distributions are highly uncertain. Each monitoring station was allocated to the group with highest membership probability. Clearly, clusterings based on such partitions waste all the information concerning the uncertainty in site-to-group allocation. The upper triangle of Table 5 provides the values of the p -Rand index computed for each pair of solutions obtained by Pastres et al. (2011),

whereas the lower triangle of the same table shows the values of Rand index for the corresponding partitions. We can notice that the agreement between the solutions is generally increased when we take into account the uncertainty in the classification. In particular, the agreement between all the pairs of solutions encompassing the one based on ammonium concentrations is sensibly increased when estimated membership probabilities are taken explicitly into account.

4.2.2 Fisher’s Iris data

This is a very well known data set consisting of three samples of size 50 of Iris flowers of three species: Iris setosa, Iris virginica and Iris versicolor. The length and width of petals and sepals were measured on each flower. We shall compare the clusterings obtained by using length and width of petals and sepals separately. A partition in two groups is identified using sepal width and length, whereas three groups were obtained using petal width and length. We are interested in quantifying the agreement between two clusterings, rather than understanding which partition best represents the presence of grouping in the data. Figures 8 and 9 show that the group of Iris setosa is rather correctly identified by both clusterings, whereas virginica and versicolor are merged when using sepal measurements. Chance-corrected Rand and p -Rand indices are equal to 0.554 and 0.451. Such values seem to indicate that the apparent agreement between the two partitions is mitigated by the degree of uncertainty of the probabilistic solutions. On the other hand, if we consider the bootstrap distribution of $\tilde{I}_a$ under the assumption of random agreement (see Fig. 10), we can see that the observed value of $\tilde{I}_a$ actually denotes a strong agreement between the two solutions.

5 Conclusions

We introduced a new similarity index capable of taking into account the degree of uncertainty of different classifications and suitable for the comparison of model-based clustering solutions. We also defined a chance-corrected version of such index. We believe that correction for chance should be used with caution, since a similarity index should always be interpreted by taking into account its sampling distribution. Unfortunately, we almost always ignore the distribution of such indices. However, through some simulation studies, we found that, for the indices we propose, the bootstrap can be very helpful, at least when mixture models are fitted to moder-

ately large samples. We found particularly interesting the fact that non-parametric bootstrap can be used to assess the null hypothesis of random agreement against the alternative of non-random (i.e. the observed agreement is not due to chance). We think that further research is needed in order to understand and, perhaps, improve the properties of such tests. Clearly, in particularly complex models, resampling via the bootstrap might be computationally too far expensive or even unfeasible. Also in such circumstances, the index we propose is comparable with the adjusted Rand index.

References

- Albatineh, A. and M. Niewiadomska-Bugaj (2011a). Correcting Jaccard and other similarity indices for chance agreement in cluster analysis. *Advances in Data Analysis and Classification* 5, 179–200. 10.1007/s11634-011-0090-y.
- Albatineh, A. and M. Niewiadomska-Bugaj (2011b). Mcs: A method for finding the number of clusters. *Journal of Classification* 28, 184–209. 10.1007/s00357-010-9069-1.
- Albatineh, A., M. Niewiadomska-Bugaj, and D. Mihalko (2006). On similarity indices and correction for chance agreement. *Journal of classification* 23(2), 301–313.
- Brusco, M. and J. Cradit (2001). A variable-selection heuristic for k-means clustering. *Psychometrika* 66, 249–270. 10.1007/BF02294838.
- Carmone, F., A. Kara, and S. Maxwell (1999). Hlnov: a new model to improve market segment definition by identifying noisy variables. *Journal of marketing research* 36(4), 501–509.
- Chae, S. S., J. L. DuBien, and W. D. Warde (2006). A method of predicting the number of clusters using Rand’s statistic. *Computational Statistics and Data Analysis* 50(12), 3531 – 3546.
- Fraley, C. and A. E. Raftery (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association* 97(458), 611–631.

- Fraley, C. and A. E. Raftery (2006). MCLUST version 3 for R: Normal mixture modeling and model-based clustering. Technical Report 504, University of Washington, Department of Statistics. (revised 2009).
- Hubert, L. and P. Arabie (1985). Comparing partitions. *Journal of classification* 2(1), 193–218.
- James, G. M. and C. A. Sugar (2003). Clustering for sparsely sampled functional data. *Journal of the American Statistical Association* 98(462), 397–408.
- Krieger, A. M. and P. E. Green (1999). A generalized rand-index method for consensus clustering of separate partitions of the same data base. *Journal of Classification* 16, 63–89. 10.1007/s003579900043.
- McLachlan, G. and K. Basford (1988). *Mixture Models: Inference and Applications to Clustering*. Statistics, textbooks and monographs. M. Dekker.
- Pastres, R., A. Pastore, and S. F. Tonellato (2011). Looking for similar patterns among monitoring stations. Venice Lagoon application. *Environmetrics* 22(6), 712–724.
- Rand, W. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 846–850.
- Scrucca, L. (2010). Dimension reduction for model-based clustering. *Statistics and Computing* 20, 471–484. 10.1007/s11222-009-9138-7.
- Steinley, D. (2004). Properties of the Hubert-Arabie adjusted rand index. *Psychological Methods* 9(3), 386.
- Warrens, M. (2008a). On similarity coefficients for 2×2 tables and correction for chance. *Psychometrika* 73(3), 487–502.
- Warrens, M. (2008b). On the equivalence of Cohen’s kappa and the Hubert-Arabie adjusted Rand index. *Journal of Classification* 25(2), 177–183.
- Yeung, K. Y., C. Fraley, A. Murua, A. E. Raftery, and W. L. Ruzzo (2001). Model-based clustering and data transformations for gene expression data. *Bioinformatics* 17(10), 977–987.

		Procedure B		
		in the same cluster	in different clusters	total
Procedure A	in the same cluster	a	b	$a + b$
	in different clusters	c	d	$c + d$
total		$a + c$	$b + d$	1

Table 1: Similarity table for two partitions (a , b , c and d denote relative frequencies).

		Procedure B		
		in the same cluster	in different clusters	total
Procedure A	in the same cluster	$\Pr(J_A \cap J_B)$	$\Pr(J_A \cap \bar{J}_B)$	$\Pr(J_A)$
	in different clusters	$\Pr(\bar{J}_A \cap J_B)$	$\Pr(\bar{J}_A \cap \bar{J}_B)$	$\Pr(\bar{J}_A)$
total		$\Pr(J_B)$	$\Pr(\bar{J}_B)$	1

Table 2: A table for the comparison of two probabilistic clustering solutions

g	τ						
	0.9	0.8	0.7	0.6	0.5	0.4	0.3
2	0.942 (0.001)	0.715 (0.003)	0.370 (0.004)	0.096 (0.002)	–	–	–
3	0.970 (0.001)	0.855 (0.002)	0.639 (0.003)	0.375 (0.003)	0.152 (0.002)	0.024 (0.001)	–
4	0.977 (0.000)	0.892 (0.001)	0.726 (0.002)	0.500 (0.003)	0.272 (0.003)	0.100 (0.002)	0.011 (0.001)
5	0.981 (0.000)	0.909 (0.001)	0.768 (0.002)	0.567 (0.003)	0.345 (0.003)	0.159 (0.002)	0.040 (0.001)
6	0.982 (0.000)	0.918 (0.001)	0.793 (0.002)	0.607 (0.003)	0.394 (0.003)	0.201 (0.002)	0.067 (0.001)
7	0.984 (0.000)	0.924 (0.001)	0.809 (0.002)	0.635 (0.003)	0.429 (0.003)	0.234 (0.002)	0.089 (0.001)
8	0.984 (0.000)	0.929 (0.001)	0.820 (0.002)	0.655 (0.003)	0.454 (0.003)	0.259 (0.002)	0.107 (0.001)
9	0.985 (0.000)	0.932 (0.001)	0.828 (0.002)	0.670 (0.002)	0.475 (0.003)	0.279 (0.003)	0.122 (0.002)

Table 3: Results of the simulation experiment illustrated in Section 4.1.1. Mean values of $\tilde{I}_a(A, A^*)$ and standard deviation (in brackets) for different values of the number of groups g and of the parameter τ ; $\tau_0 = 1000$.

Figure 1: Behaviour of $\tilde{I}_a$ as a function of the maximum membership probability, τ , with varying number of groups, g , with mixture weights fixed at $(1 - \tau)/(g - 1)$ and $g = 2, 3, 4, 5, 7, 9$ (from right to left).

Figure 2: Distribution of p-Rand index: comparison of two independent bivariate normal mixtures with two components, known number of components. (a) Black = estimated density on 2000 independent replications; red= bootstrap estimated density, $B = 2000$; blue segment = value of p-Rand on the “observed” sample used for bootstrapping. (b) Black = empirical cumulative distribution function on 2000 independent replications; red = empirical cumulative distribution function on 2000 values of p-Rand index resampled via bootstrap; blue segment = value of p-Rand on the “observed” sample used for bootstrapping.

Figure 3: Distribution of p-Rand index: comparison of two independent bivariate normal mixtures with two components, unknown number of components. (a) Black = estimated density on 2000 independent replications; red= bootstrap estimated density, $B = 2000$; blue segment = value of p-Rand on the “observed” sample used for bootstrapping. (b) Black = empirical cumulative distribution function on 2000 independent replications; red = empirical cumulative distribution function on 2000 values of p-Rand index resampled via bootstrap; blue segment = value of p-Rand on the “observed” sample used for bootstrapping.

Figure 4: Distribution of p-Rand index: comparison of two independent bivariate normal mixtures with 5 components, unknown number of components. (a) Black = estimated density on 2000 independent replications; red= bootstrap estimated density, $L = 2000$; blue segment = value of p-Rand on the “observed” sample used for bootstrapping. (b) Black = empirical cumulative distribution function on 2000 independent replications; red = empirical cumulative distribution function on 2000 values of p-Rand index resampled via bootstrap; blue segment = value of p-Rand on the “observed” sample used for bootstrapping.

Figure 5: Boxplots of $\tilde{T}$ on 2000 replicated samples of size $n = 200$ from a bivariate two component normal mixture and its perturbations with varying σ_ϵ

Figure 6: Boxplots of $\tilde{T}$ on 2000 replicated samples of size $n = 200$ from a bivariate 5-component normal mixture and its perturbations with varying σ_ϵ

(a) Two component perturbed mixture

(b) Five component perturbed mixture

Figure 7: P-Rand index computed on perturbed data with some percentile estimates based on the bootstrap

	Salinity		Chlorophyll-a		Ammonium		Nitrate		Reactive phosphorus		
Site	<i>H</i>	<i>L</i>	<i>H</i>	<i>L</i>	<i>H</i>	<i>L</i>	<i>H</i>	<i>L</i>	<i>H</i>	<i>I</i>	<i>L</i>
B01	0.00	1.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	0.00
B02	0.00	1.00	0.01	0.99	0.99	0.01	1.00	0.00	0.00	1.00	0.00
B03	1.00	0.00	0.00	1.00	0.95	0.05	0.02	0.98	0.00	0.00	1.00
B04	0.00	1.00	0.02	0.98	1.00	0.00	1.00	0.00	0.00	0.95	0.05
B05	0.00	1.00	1.00	0.00	0.95	0.04	0.07	0.93	0.00	0.95	0.05
B06	0.00	1.00	1.00	0.00	0.94	0.06	0.62	0.38	0.00	1.00	0.00
B07	0.00	1.00	1.00	0.00	1.00	0.00	0.99	0.01	0.74	0.26	0.00
B08	0.00	1.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	0.00
B09	0.03	0.97	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	0.00
B10	0.00	1.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	0.00
B11	1.00	0.00	0.00	1.00	0.99	0.01	0.99	0.01	0.00	0.54	0.46
B12	1.00	0.00	0.00	1.00	0.95	0.05	0.86	0.14	0.00	0.01	0.99
B13	1.00	0.00	0.00	1.00	0.87	0.13	0.82	0.18	0.00	0.01	0.99
B14	1.00	0.00	0.00	1.00	0.04	0.96	0.00	1.00	0.00	0.00	1.00
B15	1.00	0.00	0.00	1.00	0.37	0.63	0.00	1.00	0.00	0.00	1.00
B16	0.00	1.00	1.00	0.00	0.87	0.13	0.00	1.00	0.00	0.01	0.99
B17	1.00	0.00	0.96	0.04	0.72	0.28	0.00	1.00	0.00	0.00	1.00
B18	1.00	0.00	0.00	1.00	0.19	0.81	0.00	1.00	0.00	0.00	1.00
B19	1.00	0.00	0.00	1.00	0.46	0.55	0.00	1.00	0.00	0.00	1.00
B20	0.00	1.00	1.00	0.00	1.00	0.00	0.02	0.98	0.00	1.00	0.00
C01	1.00	0.00	0.01	0.99	0.95	0.05	0.02	0.98	0.00	0.01	0.98
C02	0.00	1.00	0.95	0.05	1.00	0.00	1.00	0.00	0.09	0.92	0.00
C03	0.93	0.07	1.00	0.00	1.00	0.00	1.00	0.00	0.89	0.11	0.00
C04	1.00	0.00	0.00	1.00	0.80	0.20	0.99	0.01	0.00	0.00	1.00
C05	1.00	0.00	0.96	0.04	1.00	0.00	1.00	0.00	0.00	1.00	0.00
C06	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	0.01	0.99	0.00
C07	1.00	0.00	0.00	1.00	0.75	0.25	0.28	0.72	0.00	0.00	1.00
C08	1.00	0.00	0.00	1.00	0.94	0.06	0.00	1.00	0.00	0.00	1.00
M01	1.00	0.00	0.00	1.00	0.00	1.00	0.98	0.02	0.00	0.00	1.00
M02	1.00	0.00	0.00	1.00	0.01	0.99	0.10	0.90	0.00	0.00	1.00

Table 4: Estimated membership probabilities for the classification with respect to the trophic variables. Membership probability distributions characterised by high uncertainty (minimum value greater or equal than 0.1) are highlighted in bold.

	Salinity	Phosphorus	Nitrate	Ammonium	Chlorophyll-a
Salinity	1.00	0.73	0.56	0.64	0.69
Phosphorus	0.69	1.00	0.66	0.65	0.72
Nitrate	0.52	0.60	1.00	0.72	0.56
Ammonium	0.54	0.51	0.63	1.00	0.67
Chlorophyll-a	0.67	0.67	0.52	0.54	1.00

Table 5: p-Rand (upper triangle) and Rand (lower triangle) index for the solutions provided by Pastres et al. (2011) and for their corresponding partitions

(a) fig:sepal-a

(b) fig:sepal-b

Figure 8: (a) Black = individuals allocated to group 1; red = individuals allocated to group 2. (b) Uncertainty plot: bubble diameter increases with allocation uncertainty.

(a) fig:petal-a

(b) fig:petal-b

Figure 9: (a) Black = individuals allocated to group 1; red = individuals allocated to group 2; green = individuals allocated to group 3. (b) Uncertainty plot: bubble diameter increases with allocation uncertainty.

(a)

(b)

Figure 10: (a) Bootstrap estimate of the density of p-Rand index. (b) Bootstrap ecdf of p-Rand index.