TESTS FOR RANDOM AGREEMENT IN CLUSTER ANALYSIS

Monjed H. Samuh[1]

Applied Mathematics Department, Palestine Polytechnic University, Hebron, Palestine

Mathematics and Statistics Department, King Fahd University of Petroleum & Minerals, Dhahran, Saudi Arabia

Friedrich Leisch

Institute of Applied Statistics and Computing, University of Natural Resources and Life Sciences, Vienna, Austria

Livio Finos

Department of Statistical Sciences, University of Padua, Padua, Italy

Abstract *The adjusted Rand index is a measure of similarity or agreement between two clusterings of the same dataset. It is based on counting pairs of points and comparing the agreement and the disagreement between two clusterings or two classification rules. In this paper, the adjusted Rand index is proposed as a test statistic. Two testing methods are proposed. The first method is based on the χ^2 distribution assuming the cluster sizes within each set of clusters are equal. The second method is based on the permutation approach. Comparison between these methods is carried out in terms of empirical level of significance.*

Keywords: *Adjusted Rand index; Cluster analysis; Permutation test; Random agreement; Similarity measures.*

1. INTRODUCTION

Measuring the similarity between two clusterings (two sets of clusters) for the same dataset have received strong interest in the literature. This is due to the existence of many different clustering algorithms (Leonard and Peter, 1990; Theodoridis and Koutroumbas, 2006) or to the fact that different researchers may use the same clustering algorithm but different starting points which yield different clusterings (Brennan and Light, 1974). Therefore, measuring this agreement is a fundamental problem in the cluster analysis field.

In order to clarify ideas and to avoid misunderstanding of what we mean by similarity or agreement between two clusterings, it is helpful to refer to an

[1] Monjed H. Samuh, monjedsamuh@ppu.edu

example. Suppose two researchers are asked independently to cluster or partition a dataset into several clusters, so we have two clusterings. The specific criterion for partitioning is left up to each researcher. Thus the number of clusters within each clustering could be different. Moreover, each researcher may use different labels for his clusters. An important question to be asked is whether the two researchers agree or disagree. For example, consider a two-dimensional dataset of size 100. In Figure 1(a) the two researchers agree completely. In Figure 1(b) they also agree completely although different labels are used. There is a strong agreement in Figure 1(c) although different number of clusters are used. Finally, Figure 1(d) depicts a random agreement. Note that the random agreement occurred when at least one of the researchers partition the dataset into clusters randomly.

It is worthwhile to observe that the problem of measuring agreement between two (or more) researchers, given that the categories or the cluster labels are predefined and imposed on researchers, is investigated in the literature. Cohen (1960) introduced the coefficient kappa to measure the degree of agreement between two researchers who cluster the observations among predefined categories. This measure has been extended to three or more researchers by Light (1971) and Fleiss (1971). See also Cohen (1968), Everitt (1968) and Fleiss et al. (1969).

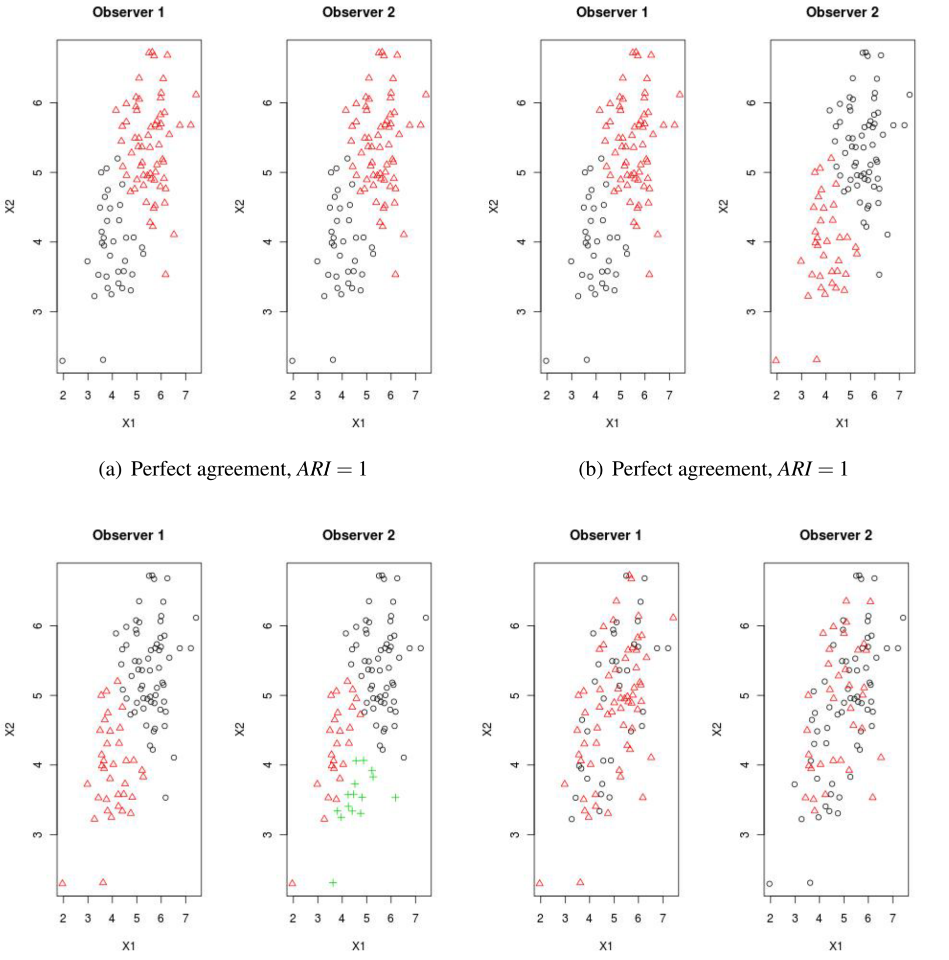
The problem considered in this paper is somewhat different. The two researchers are asked to cluster the observations into several clusters. The specific criterion for clustering is left up to each researcher. Thus the two researchers may develop different number of clusters. Moreover, since no precise set of clusters have been labelled in advance, each researcher may use different clustering criteria resulting in categories with different labels.

A large number of agreement measures have been proposed in the literature, which can be classified into three types of measures:

1. Pair counting, which are based on counting pairs of points and comparing the *agreement* and the *disagreement* between two clusterings. Jaccard index (Jaccard, 1901), Rand index (Rand, 1971), Fowlkes and Mallows index (Fowlkes and Mallows, 1983) and adjusted Rand index (Hubert and Arabie, 1985) are examples of this group of measures.

2. Set matching, which are based on measuring the shared set cardinality between two clusterings. F-measures (Rijsbergen, 1979) and misclassification rate (Meilă, 2005) are examples of this group of measures.

3. Information theoretic, which are based on the conditional probabilities resulting from the number of points shared between clusters of the two clus-



(a) Perfect agreement, $ARI = 1$

(b) Perfect agreement, $ARI = 1$

(c) Strong agreement, $ARI = 0.75$

(d) Random agreement, $ARI \approx 0$

Figure 1: Agreement between two clusterings of a dataset obtained independently by two different researchers

terings. Mutual information (Strehl and Ghosh, 2003) and variation of information (Meilă, 2005) are examples of this group of measures.

For more details see Hubálek (1982), Albatineh et al. (2006), Milligan and Cooper (1986) and Warrens (2008a,b).

Few publications are found in the literature concerning distributional properties of agreement measures. Janson and Vegelius (1981) derived the mean and the variance of Jaccard index. McCormick et al. (1992) derived the exact distribution of the Jaccard index assuming an underlying multinomial distribution with all categories equally likely except one. Hubert and Arabie (1985) derived the mean of the Rand index under the hypergeometric distribution assumption. Fowlkes and Mallows (1983) derived the mean and variance for Rand index. Albatineh (2010) generalized the derivation of Fowlkes and Mallows (1983) for the mean and the variance to a large number of similarity measures. Finally, Shuweihdi and Taylor (2007) showed that the Rand index is linearly related to the Pearson statistic given that the cluster sizes (i.e. the number of observations within each cluster) within each clustering are equal.

In this paper, the adjusted Rand index (*ARI*) is used as a test statistic for testing the null hypothesis of random agreement. The concept of the *ARI* and its properties are reviewed in Section 2. Tests for the null hypothesis of random agreement using χ^2 distribution and permutation approaches are investigated in Section 3. A simulation study to investigate the empirical level of significance is carried out in Section 4. Finally, concluding remarks are presented in Section 6.

2. ADJUSTED RAND INDEX

2.1. DEFINITION AND NOTATION

Consider a dataset with n items denoted by $\mathbf{X} = \{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$. Let $\mathscr{U}$ with r clusters and $\mathscr{V}$ with c clusters be two clusterings to be compared. $\mathscr{U}$ and $\mathscr{V}$ are obtained independently by two researchers, or by the same researcher but in different occasions or different starting points, or by applying two different clustering algorithms. The information on the overlap between $\mathscr{U}$ and $\mathscr{V}$ can be summarized by considering one of the following representations.

- **Representation 1** Each clustering is represented by a string of symbols containing the cluster labels of the corresponding data points. For example, $\mathscr{U} = \{u_1, u_1, u_3, u_4, u_4, \ldots\}$ and $\mathscr{V} = \{v_3, v_3, v_1, v_2, v_4, \ldots\}$ means the first data point $\mathbf{X}_1$ is labeled by u_1 in clustering $\mathscr{U}$ whereas it is labeled by v_3 in clustering $\mathscr{V}$, and so on.

- **Representation 2** Let $\mathscr{U} = \{\mathbf{u}_1, \ldots, \mathbf{u}_r\}$ and $\mathscr{V} = \{\mathbf{v}_1, \ldots, \mathbf{v}_c\}$, where $\mathbf{u}_i$ is the set of all data points clustered into the i^{th} cluster, $i = 1, \ldots, r$, by $\mathscr{U}$, and $\mathbf{v}_j$ is the set of all data points clustered into the j^{th} cluster, $j = 1, \ldots, c$, by $\mathscr{V}$. Then the information on cluster overlap between $\mathscr{U}$ and $\mathscr{V}$ can be summarized in the form of a $r \times c$ contingency table as illustrated in Table 1, where n_{ij} is the number of items classified into cluster $\mathbf{u}_i$ according to $\mathscr{U}$ and into cluster $\mathbf{v}_j$ according to $\mathscr{V}$. The cluster sizes in the two clusterings are the row and column totals of the contingency table given by $n_{i+} = \sum_j n_{ij}$ and $n_{+j} = \sum_i n_{ij}$.

Table 1: The contingency table, $n_{ij} = \mathbf{u}_i \bigcap \mathbf{v}_j$

$\mathscr{U} \downarrow \quad \mathscr{V} \rightarrow$	$\mathbf{v}_1$	$\mathbf{v}_2$	$\ldots$	$\mathbf{v}_c$	n_{i+}
$\mathbf{u}_1$	n_{11}	n_{12}	$\ldots$	n_{1c}	n_{1+}
$\mathbf{u}_2$	n_{21}	n_{22}	$\ldots$	n_{2c}	n_{2+}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
$\mathbf{u}_r$	n_{r1}	n_{r2}	$\ldots$	n_{rc}	n_{r+}
n_{+j}	n_{+1}	n_{+2}	$\ldots$	n_{+c}	n

- **Representation 3** Any pair of data points from the total of $N = \binom{n}{2}$ different pairs in the dataset $\mathbf{X}$ falls into one of the following four types of pairs:

  1. N_{11}: the number of pairs that are in the same cluster in both $\mathscr{U}$ and $\mathscr{V}$;
  2. N_{00}: the number of pairs that are in different clusters in both $\mathscr{U}$ and $\mathscr{V}$;
  3. N_{01}: the number of pairs that are in the same cluster in $\mathscr{U}$ but in different clusters in $\mathscr{V}$;
  4. N_{10}: the number of pairs that are in different clusters in $\mathscr{U}$ but in the same cluster in $\mathscr{V}$.

  These quantities can be calculated using the n_{ij}'s (Hubert and Arabie, 1985). Intuitively, N_{00} and N_{11} are typically interpreted as agreements in the classification of the items whereas N_{01} and N_{10} represent disagreements. The information on cluster overlap between $\mathscr{U}$ and $\mathscr{V}$ can be summarized in the form of a 2×2 contingency table as illustrated in Table 2. The Rand index (Rand, 1971) is simply defined as the probability of agreement:

$$RI = \frac{N_{00} + N_{11}}{N}.$$

Table 2: The contingency table, definitions of agreement and disagreement

$\mathscr{U} \downarrow \quad \mathscr{V} \rightarrow$	Pairs in same cluster	Pairs in different clusters
Pairs in same cluster	N_{11}	N_{01}
Pairs in different clusters	N_{10}	N_{00}

The Rand index lies between 0 and 1. It takes the value of 1 when the two clusterings are identical and 0 when the two clusterings have no agreement at all. In fact, the latter happens if and only if one clustering consists of a single cluster and the other only of clusters containing single points. However, the unique case where $RI = 0$ is quite extreme and has little practical value. In most situations the Rand index often lies within the narrower range of $[0.5, 1]$. Therefore, the Rand index possibly gives high values to pairs of randomly generated clusterings, e.g. 0.5, and this baseline value does not take on the same value in different scenarios. In fact, it is desirable for the similarity measure between two random clusterings to take values close to zero, or at least a constant value. A further problem with the Rand index is that its expected value between two random clusterings does not even take a constant value. Hubert and Arabie (1985), by taking the generalized hypergeometric distribution as the model of randomness, i.e. the two clusterings are picked at random subject to having the original number of classes and objects in each, found the expected value for $N_{00} + N_{11}$. They suggested using a corrected version of the Rand index of the form:

$$Adjusted\_Index = \frac{Index - \mathbb{E}(Index)}{Max(Index) - \mathbb{E}(Index)}$$

thus giving rise to the adjusted Rand index given by:

$$ARI(\mathscr{U}, \mathscr{V}) = \frac{\sum_i \sum_j \binom{n_{ij}}{2} - \sum_i \binom{n_{i+}}{2} \sum_j \binom{n_{+j}}{2} / \binom{n}{2}}{0.5\left(\sum_i \binom{n_{i+}}{2} + \sum_j \binom{n_{+j}}{2}\right) - \sum_i \binom{n_{i+}}{2} \sum_j \binom{n_{+j}}{2} / \binom{n}{2}}. \tag{1}$$

The ARI is bounded above by 1 and takes on the value 0 when the index equals its expected value (under the generalized hypergeometric distribution assumption for randomness). For more details see Hubert and Arabie (1985); Yeung and Ruzzo (2001).

Using Representation 3, Warrens (2008b) showed that the *ARI* can be rewritten as follows:

$$ARI(\mathscr{U}, \mathscr{V}) = \frac{2(N_{11}N_{00} - N_{01}N_{10})}{(N_{11} + N_{01})(N_{00} + N_{01}) + (N_{00} + N_{10})(N_{10} + N_{11})}.$$

Albatineh et al. (2006) introduced a family of similarity measures which can be written in the form $\beta_0 + \beta_1 \sum_i \sum_j n_{ij}^2$, where β_0 and β_1 are unique for each measure. The *ARI* can be exactly treated the same way. Given Equation 1, after simple algebra, the *ARI* is written in the following form:

$$ARI(\mathscr{U},\mathscr{V}) = \beta_0 + \beta_1 \sum_i \sum_j n_{ij}^2, \tag{2}$$

where

$$\beta_0 = \frac{-n - \frac{PQ}{n(n-1)}}{0.5(P+Q) - \frac{PQ}{n(n-1)}}$$

and

$$\beta_1 = \frac{1}{0.5(P+Q) - \frac{PQ}{n(n-1)}}$$

with $P = \sum_i n_{i+}^2 - n$ and $Q = \sum_j n_{+j}^2 - n$.

2.2. *ARI* AND PEARSON STATISTIC

Let the totals within each marginal be equal, that is,

$$n_{i+} = \frac{n}{r}, \forall i = 1, \ldots, r \tag{3}$$

and

$$n_{+j} = \frac{n}{c}, \forall j = 1, \ldots, c. \tag{4}$$

Shuweihdi and Taylor (2007) showed that the Rand index is linearly related with the Pearson statistic. By the same way, the relationship between *ARI* and Pearson statistic can be derived. The Pearson statistic is given by

$$X^2 = \sum_i \sum_j \frac{\left(n_{ij} - \frac{n_{i+}n_{+j}}{n}\right)^2}{\frac{n_{i+}n_{+j}}{n}}.$$

Under restrictions 3 and 4, the Pearson statistic becomes

$$X^2 = \frac{rc}{n} \sum_i \sum_j n_{ij}^2 - n.$$

Therefore, after simple algebra,

$$ARI = \gamma_0 + \gamma_1 X^2, \tag{5}$$

where $\gamma_0 = \frac{c+r-rc-1}{d}$ and $\gamma_1 = \frac{n-1}{nd}$ with $d = 0.5nc - rc + 0.5c + 0.5nr - n + 0.5r$.

3. TESTS FOR RANDOM AGREEMENT

Consider two independent clusterings $\mathscr{U}$ and $\mathscr{V}$. We wish to compare the observed number of agreements with the number expected from "chance" agreement. Thus, we test the null hypothesis:

$$H_0 : \{\text{There is a random agreement between } \mathscr{U} \text{ and } \mathscr{V}\}$$

or equivalently

$$H_0 : A \leq 0.5,$$

against the one-tailed alternative hypothesis:

$$H_1 : \{\mathscr{U} \text{ and } \mathscr{V} \text{ are not random}\}$$

or equivalently

$$H_1 : A > 0.5,$$

where A is the population parameter (the true value of the statistic *ARI*).

Performing the test based on the statistic *ARI* requires the knowledge of its probability distribution under the null hypothesis which is tedious to find in closed form. To overcome this problem, two approaches are proposed; χ^2 distribution approach (Section 3.1) and permutation approach (Section 3.2).

3.1. χ^2 DISTRIBUTION APPROACH

When the clusterings $\mathscr{U}$ and $\mathscr{V}$ have equal cluster sizes, it is shown in Section 2.2 that the *ARI* can be written as a linear function of Pearson statistic (see Equation 5).

Since X^2 has an asymptotic χ^2 distribution with $\nu = (r-1)(c-1)$ degrees of freedom, then the probability distribution of *ARI* is given by

$$f_{ARI}(x) = \frac{1}{2^{\nu/2}\Gamma(\nu/2)\gamma_1}\left(\frac{x-\gamma_0}{\gamma_1}\right)^{\nu/2-1}\exp\left\{\frac{-(x-\gamma_0)}{2\gamma_1}\right\}, \text{ where } x \geq \gamma_0.$$

with mean

$$\mathbb{E}\left(ARI(\mathscr{U},\mathscr{V})\right) = \gamma_0 + \gamma_1\nu,$$

and variance

$$\mathbb{V}\left(ARI(\mathscr{U},\mathscr{V})\right) = 2\nu\gamma_1^2.$$

To test the null hypothesis of random agreement, the following test statistic is used.

$$X^2_{ARI}(\mathscr{U},\mathscr{V}) = \frac{ARI - \gamma_0}{\gamma_1},$$

which has an asymptotic χ^2 distribution with $\nu = (r-1)(c-1)$ degrees of freedom. Therefore, the asymptotic p-value is given by

$$\lambda_1 = 1 - F_{X^2}(X^{2o}_{ARI}) = \int_{X^{2o}_{ARI}}^{\infty} f_{ARI}(x)\,dx,$$

where X^{2o}_{ARI} is the observed test statistic and $F_{X^2}(\cdot)$ is the asymptotic cdf of χ^2 random variable.

The size of the test has the correct nominal level α in the sense that $\int_{X^2_\alpha}^{\infty} f_{ARI}(x)\,dx = \alpha$.

3.2 PERMUTATION APPROACH

χ^2 distribution approach, discussed in Section 3.1, is valid when the cluster sizes within each clustering are equal and the expected frequency of each cell is at least 5. In practice, these restrictions may not be attained. Therefore, an alternative approach is required. In this section, a permutation test is proposed.

The idea of permutation test dates back to Fisher (1934/1935), and Pitman (1937/1938) was next to consider permutation tests. Fisher (1934, 1935) introduced the permutation approach for exact inference within the conditionality and sufficiency principles of inference. He introduced the permutation test as the exact test for the association between two binary variables when some cells have expected frequencies less than 5; that is, when the chi-square test fails. Also it is useful for one sided testing if at least one variable is ordered categorical. In addition, Fisher introduced the exact test for testing differences between means of two populations when the assumptions of the two-sample t-test were not met. He pointed out that the probability of a type I error for the two-sample permutation test is closely approximated by the normal theory probability of a type I error for the particular problem dealt with.

Pitman (1937a,b, 1938) developed exact permutation methods consistent with the Neyman-Pearson approach for the comparison of $k \geq 2$-samples and for bivariate correlation. For two-sample design, Pitman introduced a test statistic which is a monotonic increasing function of the square of the t-test statistic.

Permutation tests are considered a subclass of nonparametric tests (Lehmann and Romano, 2005; Pesarin and Salmaso, 2010). They are computationally intensive, but modern computational power makes permutation tests feasible. Nonparametric test statistics do not rely on a specific probability distribution that describes the underlying population. In fact, permutation tests are always distribution free since observed data are sufficient statistics in the null hypothesis (see Pesarin and Salmaso, 2010, Sec. 2.1.3). However, some tests (two-sample design, ANOVA, etc.) require assumptions upon to the samples rather than the underlying distributions or parameters. An important assumption is that the observations are exchangeable under the null hypothesis. The exchangeability is generally assured by random allocation of treatments to units in experimental work. In observational studies, exchangeability in the null hypothesis shall be assumed in order to obtain exact testing solutions. If this assumption cannot be justified, then approximate permutation solutions are obtained in accordance, for instance, with the nonparametric Behrens-Fisher testing.

Permutation tests are widely used in many research fields such as agriculture, clinical trials, educational statistics, business statistics and industrial statistics. For more works on permutation test and its variations see Edgington (1995), Pesarin (2001), Salmaso (2003), Good (2005), Basso et al. (2009) and Pesarin and Salmaso (2010) and the references therein.

The goal of using a permutation test in our problem is the computation of the conditional probability distribution of the *ARI*. For the purpose of finding the permutation sample space, Representation 1 of the two clusterings (discussed in Section 2.1) is considered. The cluster labels within each clustering are permuted then *ARI* is calculated using $\mathscr{U}^*$ and $\mathscr{V}^*$. Algorithm 1 is used to obtain the permutation (conditional) p-value for testing the null hypothesis of random agreement.

Note that the permutation mid p-value (Lancaster, 1961) is calculated due to the discreteness of the permutation distribution of the test statistic.

4. SIMULATION STUDY

In this section, the empirical level of significance of the proposed tests is investigated. To assess the empirical level of significance, the tests are performed on a two random clusterings. A random clustering can be created by assigning data points to clusters randomly. As an example, two clusterings each with three categories ($r = c = 3$) are created under the null hypothesis and three different configurations are considered: (a) $n_{i+} = 50, \forall i = 1,2,3$ and $n_{+j} = 50, \forall j = 1,2,3$; (b) $n_{1+} = n_{+1} = 5$, $n_{i+} = 50$, $i = 2,3$ and $n_{+j} = 50$, $j = 2,3$; (c) $n_{1+} = 5, n_{2+} =$

Algorithm 1 Conditional p-value of the ARI

1. For two given clusterings $\mathscr{U}$ and $\mathscr{V}$, calculate the observed test statistic $ARI(\mathscr{U},\mathscr{V})$, denoted by ARI^o.

2. Take a random permutation $\mathscr{U}^*$ of $\mathscr{U}$ and $\mathscr{V}^*$ of $\mathscr{V}$.

3. Calculate the test statistic $ARI^* = ARI(\mathscr{U}^*,\mathscr{V}^*)$.

4. Independently repeat Steps 2 and 3 many times, say B times, obtaining B test statistics, say $\{ARI_b^*, b = 1,\dots,B\}$.

5. The permutation mid p-value is estimated as

   $$\lambda_2 = \frac{\sum_{b=1}^{B} \mathbb{I}(ARI_b^* > ARI^o)}{B} + \frac{\sum_{b=1}^{B} \mathbb{I}(ARI_b^* = ARI^o)}{2B},$$

   where $\mathbb{I}(\cdot)$ is the indicator function.

$3, n_{3+} = 7$ and $n_{+1} = 1, n_{+2} = 10, n_{+3} = 4$. Steps for assessing the empirical significance level are summarized in Algorithm 2. A simulation study based on $R = 5000$ datasets is performed. The number of permutations on each dataset is $B = 1000$.

Algorithm 2 Empirical level of significance

1. For the given dataset, randomly create two clusterings $\mathscr{U}$ and $\mathscr{V}$.

2. Use the aforementioned approaches to obtain the p-values, λ_1 and λ_2.

3. Independently repeat Steps 1 and 2 many times, say R times, giving R p-values for each approach, say $\{\lambda_{ir}, r = 1,\dots,R\}, i = 1,2$.

4. For a preassigned nominal level of significance α, the empirical level of significance is given by

   $$\hat{\alpha}_i = \frac{\sum_{r=1}^{R} \mathbb{I}(\lambda_{ir} \le \alpha)}{R}, i = 1,2.$$

The simulation results are reported in Tables 3-5 for each configuration. It is clear that the empirical level of significance for the proposed tests in configuration (a) is closed to the nominal one; that is, the p-values under the null hypothesis are uniformly distributed over its support, $[0,1]$. While in configurations (b) and (c) the proposed permutation test is still valid but not the χ^2 distribution.

Table 3: The empirical level of significance, $n_{i+}= 50$, $\forall i = 1, 2, 3$ **and** $n_{+j} = 50$; $\forall\, j = 1, 2, 3$

	Nominal level α						
Method	0.05	0.10	0.20	0.40	0.60	0.80	0.90
χ^2 distribution	0.049	0.104	0.215	0.427	0.604	0.813	0.906
permutation	0.051	0.105	0.208	0.410	0.600	0.805	0.905

Table 4: The empirical level of significance, $n_{1+}= n_{+1}= 5$, $n_{i+}= 50$; $i = 2; 3$ **and** $n_{+\,j} = 50$, $j = 2, 3$

	Nominal level α						
Method	0.05	0.10	0.20	0.40	0.60	0.80	0.90
χ^2 distribution	0.049	0.103	0.184	0.409	0.550	0.804	0.999
permutation	0.049	0.098	0.200	0.408	0.596	0.800	0.898

Table 5: The empirical level of significance, $n_{1+}= 5$, $n_{2+}= 3$; $n_{3+}= 7$ **and** $n_{+1}= 1$, $n_{+2}= 10$, $n_{+3}= 4$

	Nominal level α						
Method	0.05	0.10	0.20	0.40	0.60	0.80	0.90
χ^2 distribution	0.040	0.049	0.182	0.4100	0.828	0.999	0.999
permutation	0.048	0.103	0.190	0.4100	0.575	0.828	0.871

5. AN APPLICATION

For a practical application of the test, blood clots detection study (Vanbelle, 2009) is used. The study was conducted on 50 patients to measure the efficacy of two new methods (by two medical raters) with respect to a standard method (reference) in the detection of blood clots in the legs. Each patient was classified as having (1) or not having (0) blood clot(s) in the legs with respect to a reference method called "Standard" and 2 new methods "Method 1" and "Method 2". The study aimed at comparing the agreement between the standard method and each of the new methods in order to make a choice between them. The classification of the patients according to the presence of blood clots is given in Table 6.

Table 6: Blood clots detection (0 =No, 1 =Yes) in the legs of 50 patients with a standard method and two new methods

		Method 1			Method 1		
		0	1	Total	0	1	Total
Standard	0	18	11	29	26	3	29
	1	4	17	21	4	17	21
	Total	22	28	50	30	20	50

To test which method gives better agreement with the standard method, the two proposed testing methods are applied and the results are given in Table 7. It is clear that Method 2 gives better agreement with the Standard method than Method 1. Our results agree with the outcomes in Vanbelle (2009).

Table 7: Inference on blood clots detection data under χ^2 approach and permutation approach

		χ^2 Approach	Permutation Approach
Method 1 – Standard	Test Statistic	1.11	0.01
	p-value	0.177	0.223
Method 2 – Standard	Test Statistic	22.44	0.51
	p-value	0.000	0.000

6. CONCLUDING REMARKS

Testing for random agreement between two clusterings of a dataset is investigated in this paper. Two methods are proposed; χ^2 approach using Pearson χ^2 as a test statistic, and the permutation approach using the adjusted Rand index as a test statistic. The two methods are compared in terms of type I error control and it is found that the permutation approach preserves the test size. Moreover, the permutation approach is valid for any sample size while the χ^2 approach is valid asymptotically.

Acknowledgment The authors would like to thank the Editor/referee whose comments and suggestions were very useful in improving a first draft of the paper.

REFERENCES

Albatineh, A.N. (2010). Means and variances for a family of similarity indices used in cluster analysis. In *Journal of Statistical Planning and Inference*, 140 (10): 2828–2838.

Albatineh, A.N., Niewiadomska-Bugaj, M. and Mihalko, D. (2006). On similarity indices and correction for chance agreement. In *Journal of Classification*, 23 (2): 301–313.

Basso, D., Pesarin, F., Salmaso, L. and Solari, A. (2009). *Permutation Tests for Stochastic Ordering and ANOVA: Theory and Applications with R*. Springer, New York.

Brennan, R.L. and Light, R.J. (1974). Measuring agreement when two observers classify people into categories not defined in advance. In *British Journal of Mathematical and Statistical Psychology*, 27 (2): 154–163.

Cohen, J. (1960). A coefficient of agreement for nominal scales. In *Educational and Psychosocial Measurement*, 20: 37–46.

Cohen, J. (1968). Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit. In *Psychological Bulletin*, 70 (4): 213–220.

Edgington, E.S. (1995). *Randomization Tests*. Marcel Dekker, Inc., New York, NY, USA.

Everitt, B. (1968). Moments of the statistics kappa and weighted kappa. In *British Journal of Mathematical and Statistical Psychology*, 21 (1): 97–103.

Fisher, R.A. (1934). *Statistical Methods for Research Workers*. Edinburgh: Oliver & Boyd.

Fisher, R.A. (1935). *The Design of Experiments*. Edinburgh: Oliver & Boyd.

Fleiss, J.L. (1971). Measuring nominal scale agreement among many raters. In *Psychological bulletin*, 76 (5): 378–382.

Fleiss, J.L., Cohen, J. and Everitt, B. (1969). Large sample standard errors of kappa and weighted kappa. In *Psychological Bulletin*, 72 (5): 323–327.

Fowlkes, E.B. and Mallows, C.L. (1983). A method for comparing two hierarchical clusterings. In *Journal of the American Statistical Association*, 78 (383): 553–569.

Good, P. (2005). *Permutation, Parametric, and Bootstrap Tests of Hypotheses*. Springer-Verlag, New York.

Hubálek, Z. (1982). Coefficients of association and similarity, based on binary (presence-absence) data: An evaluation. In *Biological Reviews*, 57 (4): 669– 689.

Hubert, L. and Arabie, P. (1985). Comparing partitions. In *Journal of classification*, 2 (1): 193–218.

Jaccard, P. (1901). Étude comparative de la distribution florale dans une portion des alpes et du jura. In *Bulletin del la Société Vaudoise des Sciences Naturelles*, 37: 547–579.

Janson, S. and Vegelius, J. (1981). Measures of ecological association. In *Oecologia*, 49 (3): 371–376.

Lancaster, H. (1961). Significance tests in discrete distributions. In *Journal of the American Statistical Association*, 56 (294): 223–234.

Lehmann, E.L. and Romano, J.P. (2005). *Testing Statistical Hypotheses*. Springer, New York.

Leonard, K. and Peter, J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. John While & Sons, New York.

Light, R.J. (1971). Measures of response agreement for qualitative data: Some generalizations and alternatives. In *Psychological Bulletin*, 76 (5): 365–377.

McCormick, W., Lyons, N. and Hutcheson, K. (1992). Distributional properties of Jaccard’s index of similarity. In *Communications in Statistics-Theory and Methods*, 21 (1): 51–68.

Meila, M. (2005). Comparing clusterings: an axiomatic view. In *Proceedings of the 22nd International Conference on Machine Learning*, 577–584. ACM.

Milligan, G.W. and Cooper, M.C. (1986). A study of the comparability of external criteria for hierarchical cluster analysis. In *Multivariate Behavioral Research*, 21 (4): 441–458.

Pesarin, F. (2001). *Multivariate Permutation Tests: With Applications in Biostatistics*. Wiley Chichester.

Pesarin, F. and Salmaso, L. (2010). *Permutation Tests for Complex Data: Theory, Application and Software*. John Wiley & Sons.

Pitman, E.J. (1937a). Significance tests which may be applied to samples from any populations. In *Supplement to the Journal of the Royal Statistical Society*, 4 (1): 119–130.

Pitman, E.J. (1937b). Significance tests which may be applied to samples from any populations. II. the correlation coefficient test. In *Supplement to the Journal of the Royal Statistical Society*, 4 (2): 225–232.

Pitman, E.J. (1938). Significance tests which may be applied to samples from any populations. III. the analysis of variance test. In *Biometrika*, 29 (3/4): 322–335.

Rand, W.M. (1971). Objective criteria for the evaluation of clustering methods. In *Journal of the American Statistical association*, 66 (336): 846–850.

Rijsbergen, C.J.V. (1979). *Information Retrieval*. Butterworth-Heinemann, Newton, MA, USA.

Salmaso, L. (2003). Synchronized permutation tests in 2^k factorial designs. In *Communications in Statistics-Theory and Methods*, 32 (7): 1419–1437.

Shuweihdi, F. and Taylor, C.C. (2007). Inference for similarity indices. In *Systems Biology & Statistical Bioinformatics*, 139–142.

Strehl, A. and Ghosh, J. (2003). Cluster ensembles—a knowledge reuse framework for combining multiple partitions. In *The Journal of Machine Learning Research*, 3: 583–617.

Theodoridis, S. and Koutroumbas, K. (2006). *Pattern Recognition*. Academic Press, Inc., Orlando, FL, USA.

Vanbelle, S. (2009). *Agreement Between Raters and Groups of Raters*. Ph.D. thesis, Université de Liège, Belgique.

Warrens, M.J. (2008a). On similarity coefficients for 2◊2 tables and correction for chance. In *Psychometrika*, 73 (3): 487–502.

Warrens, M.J. (2008b). On the equivalence of Cohen’s kappa and the HubertArabie adjusted Rand index. In *Journal of Classification*, 25 (2): 177–183.

Yeung, K.Y. and Ruzzo, W.L. (2001). Details of the adjusted Rand index and clustering algorithms, supplement to the paper (An empirical study on principal component analysis for clustering gene expression data). In *Bioinformatics*, 17 (9): 763–774.