Noname manuscript No.
(will be inserted by the editor)

sARI: A *soft* agreement measure for class partitions incorporating assignment probabilities

Abstract Agreement indices are commonly used to summarize the performance of both classification and clustering methods. The easy interpretation/intuition and desirable properties that result from the Rand and adjusted Rand indices, has led to their popularity over other available indices. While more algorithmic clustering approaches like k-means and hierarchical clustering produce hard partition assignments (assigning observations to a single cluster), other techniques like model-based clustering include information about the certainty of allocation of objects through class membership probabilities (soft partitions). To assess performance using traditional indices, e.g. the adjusted Rand index (ARI), the soft partition is mapped to a hard set of assignments, which commonly overstates the certainty of correct assignments. This paper proposes an extension of the ARI, the soft adjusted Rand index (sARI), with similar intuition and interpretation but also incorporating information from one or two soft partitions. It can be used in conjunction with the ARI, comparing the similarities of hard to soft, or soft to soft, partitions to the similarities of the mapped hard partitions. Simulation study results support the intuition that, in general, mapping to hard partitions tends to increase the measure of similarity between partitions. In applications, the sARI more accurately reflects the cluster boundary overlap commonly seen in real data.

Keywords adjusted Rand index · model-based clustering · mixture models · soft partition · posterior probabilities · class membership probabilities

Mathematics Subject Classification (2000) 62H30 · 91C20 · 62H86

1 Introduction

The rise of statistical clustering and classification techniques from the mid twentieth century since Hartigan (1975), produced the need for a way to measure the similarity between two partitions of the same data set. For both

Address(es) of author(s) should be given

clustering and classification, we might compare a set of estimated cluster or class labels to the true labels (if known), to characterize methodology performance. We might also compare two sets of labels, each generated by different methods, for comparison purposes or to assess performance consistency.

As an example, in Figure 1, we introduce data from two simulated Gaussians, 50 observations from $\mathcal{N}((3,2)', \Sigma)$ and 50 observations from $\mathcal{N}((5,2)', \Sigma)$, where $\Sigma = \begin{pmatrix} 1.20 & 0.15 \\ 0.15 & 1.20 \end{pmatrix}$. In the left scatterplot (Partition 1), the observations are labeled by their true group partition: black circles versus green diamonds. In the right scatterplot (Partition 2), observations are labeled by their estimated group partition from fitting a mixture of Gaussians to the data. For small data sets of low dimensionality, we are able to visualize a general sense of how often and which observations are misclassified or mislabeled; however, in general, we often rely on the use of a corresponding numerical summary that measures how similar two partitions are to each other, an *agreement index*.

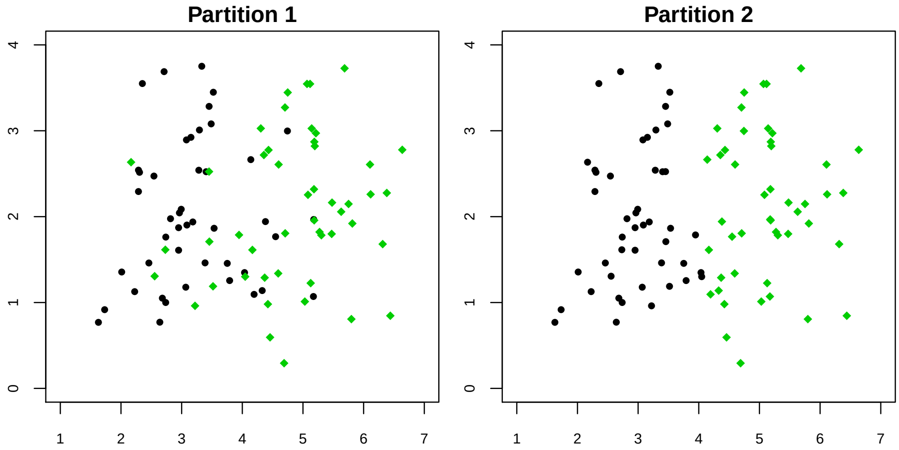

Fig. 1 Comparing two partitions: true vs estimated labels (via Gaussian mixture model).

Many agreement indices have been suggested, such as the Jaccard index (Downton and Brennan, 1980), Rand index (RI, Rand, 1971), Fowlkes-Mallows index (Fowlkes and Mallows, 1983), and adjusted Rand index (ARI, Hubert and Arabie, 1985; Morey and Agresti, 1984). The Hubert-Arabie adjusted Rand index is the most heavily used as it has several desirable properties, such as invariance to parameter differences across clusters, the number of objects clustered, and the number of clusters, see for example Steinley (2004). One commonality across these indices is their use of *hard* partitions, or a set of labels such that each observation is assigned to only one class or cluster with complete certainty. This feature matches nicely to clustering techniques such as hierarchical clustering (Ward, 1963) and k-means clustering (MacQueen et al, 1967) that produce hard assignments.

However, hard partitions or class assignments naturally imply that every observation in a cluster has the same certainty of belonging to their respective cluster or class assignment. Observations near the middle or bulk of their cluster are viewed similarly as observations near the boundary between clusters or classes. In practice, assignment certainty is likely to be very different for these two location types. In the example, Partition 2 has 17 misclassified observations, denoted by an "X" in Figure 2 (left). There are several misclassified observations that are understandably misclassified because of their location well within the "wrong" class. There are others, however, that lie within the overlap between the two simulated classes.

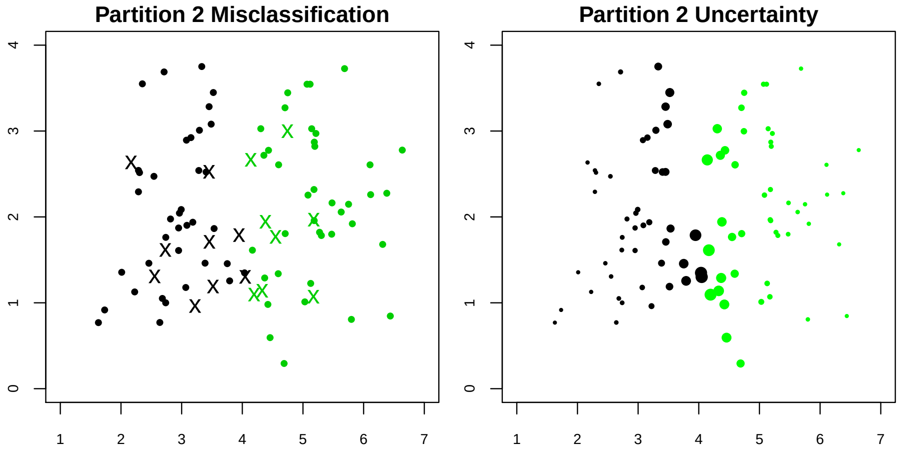

Fig. 2 Partition 2 misclassifications (indicated by "X"); Partition 2 uncertainty (observations with larger symbols have higher assignment uncertainty).

More recently, new clustering techniques (e.g. model-based clustering or Gaussian mixture model clustering: Wolfe (1963); McLachlan and Peel (2004); McLachlan and Krishnan (2007); Fraley and Raftery (2002); McNicholas (2016), fuzzy c-means: Dunn (1973); Bezdek (1981)) have been gaining popularity in practice. Model-based clustering methods produce a *soft* partition, defined here as a set of cluster assignment probabilities for each observation that (most commonly) sum to one. A soft partition can also be obtained with some fuzzy clustering methods, provided the fuzziness can be re-interpreted as a probability (Miyamoto et al, 2008). In Figure 2, the right scatterplot shows a measure of *uncertainty* associated with Partition 2, where uncertainty is defined as 1 minus the maximum cluster assignment probability for each observation. Note that larger observations with higher uncertainty are primarily near the boundary between the two clusters (as expected). Comparing the misclassified observations (left) to their uncertainties (right), we see that some misclassified observations were also some of the most uncertain assignments. Incorporating this uncertainty into agreement indices may be more representative of the partition's performance and any overlapping cluster structure that may exist.

Currently, researchers commonly transform soft partitions to hard partitions by, for example, mapping observations to the cluster or class with the maximum probability, which often tends to overstate the observation's individual certainty of class membership and the overall agreement. In addition, multiple soft partitions can lead to the same hard partition and will then be indistinguishable via agreement indices. This paper proposes an adjustment to the commonly used ARI, called the soft ARI (sARI) which allows for the comparison of a hard partition (e.g. the true labels) to a soft partition of class membership probabilities, or the comparison of two soft partitions. We advocate using the sARI in conjunction with the ARI, to more completely summarize agreement in the presence of uncertainty.

Over the last decade, there have been other indices proposed as variations on the RI or ARI, including those by Campello (2007), Huellermeyer et al (2012), and Amodio et al (2015). The first two allow for comparison of one hard and one soft partition, while the latter allows for comparison of two soft partitions. The goal of these papers was primarily to create an alternative for the ARI that incorporated probabilistic information, but was not motivated to necessarily act as a direct comparison to the ARI (or other indices of this type). In contrast, the sARI's development begins directly from the ARI and so benefits from the same understanding and direct comparability. As such, it is motivated as a companion index to the ARI rather than a supplanting one.

In Section 2 we review the Rand and adjusted Rand indices; in Section 3, we propose the soft ARI. We illustrate some properties of the sARI in comparison to the original ARI in Section 4, and present an application to real data in Section 5. The paper concludes with a summary and discussion in Section 6.

2 Similarity Indices for Comparing Partitions of Data

Although our illustrative examples will use clustering methodology, the similarity or agreement indices presented in this paper can be used for both clustering and classification approaches. As such, we use the terms cluster and class interchangeably. Given a set of observations, the goal of clustering is to find a partition such that observations that are more similar or are from the same original class are more likely to be assigned to the same estimated cluster.

2.1 Rand Index

The Rand index (RI), attributed to Rand in 1971, is commonly used to measure the correspondence between two partitions of a set of observations. The idea is based on counting types of pairs of observations. Given a set of N observations $\mathcal{S} = \{O_1, \ldots O_N\}$, suppose that there are two hard partitions of $\mathcal{S}$ namely $P_1 = \{u_1, u_2, \ldots, u_R\}$ and $P_2 = \{v_1, v_2, \ldots, v_C\}$ with R and C classes, u_r and v_c respectively, such that

$$\bigcup_{r=1}^{R} u_r = \mathcal{S} = \bigcup_{c=1}^{C} v_c, \quad u_r \cap u_{r'} = \emptyset = v_c \cap v_{c'} \;\; \forall r \neq r', c \neq c'.$$

We summarize the combinations of cluster membership in the two partitions using a contingency table such as that given in Table 1.

Table 1 Contingency table notation for comparing two hard partitions.

$P_1 \setminus P_2$	v_1	v_2	$\ldots$	v_C	Sums
u_1	n_{11}	n_{12}	$\ldots$	n_{1C}	$n_{1\cdot}$
u_2	n_{21}	n_{22}	$\ldots$	n_{2C}	$n_{2\cdot}$
$\vdots$	$\vdots$				$\vdots$
u_R	n_{R1}	n_{R2}	$\ldots$	n_{RC}	$n_{R\cdot}$
Sums	$n_{\cdot 1}$	$n_{\cdot 2}$	$\ldots$	$n_{\cdot C}$	$n_{\cdot\cdot} = N$

In this table, n_{rc} is the number of observations in both cluster u_r and v_c, and $n_{r\cdot}$ and $n_{\cdot c}$ are the number of objects in cluster u_r and v_c respectively.

There are four different types of classifications of the $\binom{N}{2}$ distinct pairs of observations

type (a): observation pairs are placed in the same cluster in partition P_1 and in the same cluster in partition P_2

type (b): observation pairs are placed in the same cluster in partition P_1 and in different clusters in partition P_2

type (c): observation pairs are placed in different clusters in partition P_1 and in the same cluster in partition P_2

type (d): observation pairs are placed in different clusters in partition P_1 and in different clusters in partition P_2

Types (a) and (d) are considered agreements (A) in classification, while types (b) and (c) are considered disagreements (D) (Hubert and Arabie, 1985). If A and D are defined in this manner, then $A + D = \binom{N}{2}$. The RI, which is interpreted as the probability of agreement, is defined as $\frac{A}{\binom{N}{2}}$ where $0 \leq \text{RI} \leq 1$, with 1 indicating perfect agreement between the 2 partitions.

Comparing the hard clustering assignments generated from Partition 2, to the true classifications from Partition 1 (Figure 1, Table 2), we have an RI of 0.715. This is fairly high agreement - although again we note that this value assumes complete certainty in the clustering assignment.

Unfortunately in practice, the RI does not span its range and only approaches its upper limit as the number of classes increase (Steinley, 2004). Because of these limitations, several variations on the RI have been proposed.

Table 2 Cross-classification table of Partition 2's hard assignment and the true labels.

$P_1 \setminus P_2$	v_1	v_2	Sums
u_1	42	8	50
u_2	9	41	50
Sums	51	49	100

2.2 Adjusted Rand Index

One such variation, the adjusted Rand index (ARI), proposed by Hubert and Arabie in 1985, corrects the RI for chance (Hubert and Arabie, 1985). In general, an index corrected for chance has the form $\frac{Index - E[Index]}{Max[Index] - E[Index]}$. It is assumed that the index follows a generalized hypergeometric distribution for its model of randomness. By assuming this model, in fixing the marginals, the partitions P_1 and P_2 are chosen at random. Hubert and Arabie (1985), using this assumption, showed that the expectation can be calculated in closed form giving an adjusted index of:

$$\mathrm{ARI}(P_1, P_2) = \frac{\sum_{r,c}\binom{n_{rc}}{2} - \left[\sum_r \binom{n_{r\cdot}}{2}\sum_c\binom{n_{\cdot c}}{2}\right] / \binom{N}{2}}{\frac{1}{2}\left[\sum_r\binom{n_{r\cdot}}{2} + \sum_c\binom{n_{\cdot c}}{2}\right] - \left[\sum_r\binom{n_{r\cdot}}{2}\sum_c\binom{n_{\cdot c}}{2}\right] / \binom{N}{2}}. \quad (1)$$

The maximum value of the ARI is 1 (for identical partitions) and under random partitioning, the expected value of the ARI is 0 (unlike the RI). Because the ARI counts pairs of observations, it is not affected by label switching of classes, and as opposed to other classification indices, it gives credit for classes that are split into multiple classes or for multiple classes that are combined into fewer classes. The ARI has been shown to be among the most desirable of the common agreement indices, also used for variable selection, evaluating simulations, and other non-standard uses as detailed in Steinley (2004).

Calculating the ARI for our illustrative example in Figure 1 and Table 2, we obtain a value of 0.430. Note that we would not compare this value directly to the RI; we would only report moderate agreement of Partition 2 with the true labels (again, assuming certainty in Partition 2's labels).

3 Soft Adjusted Rand Index

Both the RI and ARI are calculated using hard partitions which is necessarily limiting. In this section, we extend the ARI to allow for one or more soft partitions, i.e. matrices of class membership probabilities rather than vectors of class memberships.

3.1 Soft Cluster Membership Probabilities

We refer to a *soft* cluster assignment as a set of probabilities, one per cluster, that indicate the likelihood that an observation belongs to each cluster.

While there are several clustering (and classification) approaches that return soft assignments, our illustrative approach here is model-based clustering, also known as Gaussian (finite) mixture model clustering.

In finite mixture model clustering, we assume that the observations are a sample from a population with density $f(\mathbf{x})$. We further assume that $f(\mathbf{x})$ can be modeled as a weighted mixture of component (cluster) densities, i.e.

$$f(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \cdot f_k(\mathbf{x}; \theta_k)$$

where $\pi_k > 0$, $\sum_{k=1}^{K} \pi_k = 1$, and K is the number of mixture components in the population. In model-based clustering all f_k are assumed to be Gaussian. This model is usually estimated through an EM algorithm (Dempster et al, 1977) iterating between cluster assignments and parameter estimation with the final model being chosen by some criterion, such as the Bayesian Information Criterion. The model can also be estimated through standard Bayesian inference using MCMC approaches. Each component f_k is assumed to represent a single cluster. Each observation $\mathbf{x}_i$ is then assigned a vector of cluster membership posterior probabilities: for $i = 1, 2, ..., N$ and $k = 1, 2, ..., K$,

$$z_{ki} = \hat{P}(\mathbf{x}_i \in \text{cluster } k) = \hat{\pi}_k \hat{f}_k(\mathbf{x}_i) / \hat{f}(\mathbf{x}_i).$$

We used model-based clustering to generate Partition 2 in Figure 1 (specifying $K = 2$); hard assignments were generated by assigning each observation to the cluster with the highest probability. Figure 3 and Table 3 show a sample of 10 labeled observations (5 from each simulated class: $u_1,\ u_2$), providing their location and posterior probabilities for belonging to v_1 or v_2 in the model-based clustering solution. Bolded numbers correspond to the hard cluster assignment, and italicized numbers indicate incorrect final assignments (given that u_1 matches closest to v_1 and u_2 to v_2). We see a large degree of variation across pairs of posterior probabilities. Observations like 0 and 5 have a high degree of certainty about their (correctly) assigned class. Observations like 1, 2, 6, 7 show less certainty, but still have a higher probability of belonging to the correct class. We see some observations that are close to an even split, but have a slightly higher probability of belonging to the incorrect class (e.g. 3, 8). Lastly, observations like 4 and 9 have high probabilities of being assigned to the incorrect class. These soft posterior probabilities are as expected given their locations relative to fitting two overlapping Gaussians; note that the hard assignment procedure forces essentially a linear boundary between two elliptical clusters (see Figure 1 right).

3.2 Calculating the Soft Adjusted Rand Index

Let p_{rci} be the product of the (posterior) probabilities of assignment of observation i to u_r in P_1 and v_c in P_2. That is, $p_{rci} = z_{ri} \cdot z_{ci}$, given posterior probabilities z_{ri} and z_{ci}. Note that this assumes independence between the

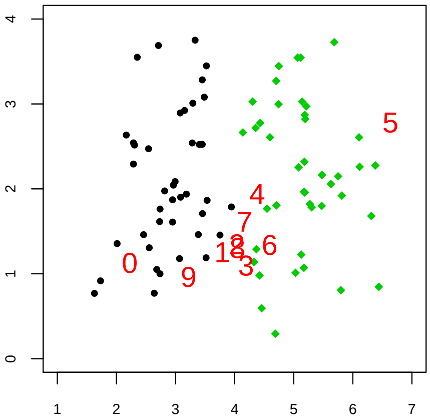

Fig. 3 Example observations from Table 3, remaining points colored by estimated cluster.

Table 3 Posterior probabilities of example observations shown in Figure 3.

		Posteriors	
	Label	v_1	v_2
u_1	0	**0.987**	0.013
	1	**0.671**	0.329
	2	**0.533**	0.467
	3	0.456	***0.544***
	4	0.315	***0.685***
u_2	5	0.002	**0.998**
	6	0.243	**0.757**
	7	0.446	**0.554**
	8	***0.529***	0.471
	9	***0.889***	0.111

two partitions. While it is likely that the classes resulting from the two methods will be dependent, it is reasonable to assume that one partition does not influence the solution of the second partition. If comparing the true labels to a model-based clustering solution, $p_{rci} = 1 \cdot z_{ci}$, if observation i comes from true class r. Note that probabilities, z_{ki} were defined previously in a model-based clustering context, however the definition of p_{rci} holds for any algorithm that provides cluster membership probabilities. Given the posteriors, we then have:

1. $p_{rc\cdot} = \sum_{i=1}^{N} p_{rci}$ is the sum of the product of the (posterior) probabilities of assignment to u_r in P_1 and v_c in P_2 over all observations.
2. $p_{\cdot c\cdot} = \sum_{r=1}^{R} p_{rc\cdot} = \sum_{r=1}^{R} \sum_{i=1}^{N} p_{rci}$ is the sum of the product of the (posterior) probabilities for belonging to group v_c in P_2 for all observations.
3. $p_{r\cdot\cdot} = \sum_{c=1}^{C} p_{rc\cdot} = \sum_{c=1}^{C} \sum_{i=1}^{N} p_{rci}$ is the sum of the product of the (posterior) probabilities for belonging to group u_r in P_1 for all observations.

These probabilities can be summarized in a contingency table (Table 4), similar to Table 1, except that this table allows fractional values.

Table 4 Contingency table notation for comparing two soft partitions.

$P_1 \setminus P_2$	v_1	v_2	$\dots$	v_C	Sums
u_1	$p_{11\cdot}$	$p_{12\cdot}$	$\dots$	$p_{1C\cdot}$	$p_{1\cdot\cdot}$
u_2	$p_{21\cdot}$	$p_{22\cdot}$	$\dots$	$p_{2C\cdot}$	$p_{2\cdot\cdot}$
$\vdots$	$\vdots$				$\vdots$
u_R	$p_{R1\cdot}$	$p_{R2\cdot}$	$\dots$	$p_{RC\cdot}$	$p_{R\cdot\cdot}$
Sums	$p_{\cdot 1\cdot}$	$p_{\cdot 2\cdot}$	$\dots$	$p_{\cdot C\cdot}$	$p_{\cdots} = N$

Replacing the n_{rc}'s in Equation (1) with the appropriate $p_{rc\cdot}$'s, gives us

$$\text{sARI}(P_1, P_2) = \frac{\sum_{r,c}\binom{p_{rc\cdot}}{2} - \left[\sum_r \binom{p_{r\cdot\cdot}}{2} \sum_c \binom{p_{\cdot c\cdot}}{2}\right] / \binom{N}{2}}{\frac{1}{2}\left[\sum_r \binom{p_{r\cdot\cdot}}{2} + \sum_c \binom{p_{\cdot c\cdot}}{2}\right] - \left[\sum_r \binom{p_{r\cdot\cdot}}{2} \sum_c \binom{p_{\cdot c\cdot}}{2}\right] / \binom{N}{2}}. \tag{2}$$

When there is certainty in the assignment of observations to classes, i.e. all (posterior) probabilities are either 0 or 1, then sARI will be identical to ARI as $p_{rc\cdot} = n_{rc}$. This is true in particular when there is a perfectly classified set of observations. There will be near agreement between sARI and ARI at the other extreme, when there is random partitioning of the observations into the classes. In this case, both n_{rc} and $p_{rc\cdot}$ are tending to N/RC.

Combinations cannot be computed on fractional values; however, notice that we can rewrite combinations in Equation (1) using gamma functions, e.g.

$$\sum_{r,c} \binom{n_{rc}}{2} = \sum_{r,c} \left(\frac{n_{rc}!}{2!(n_{rc} - 2)!} \right) = \sum_{r,c} \frac{\Gamma(n_{rc} + 1)}{2 \times \Gamma(n_{rc} - 1)}.$$

With this, we can rewrite Equation (2) using gamma functions, which allow for computation on non-integer table values. Define

$$\Lambda_{rc} = \sum_r \frac{\Gamma(p_{r\cdot\cdot} + 1)}{\Gamma(p_{r\cdot\cdot} - 1)} + \sum_c \frac{\Gamma(p_{\cdot c\cdot} + 1)}{\Gamma(p_{\cdot c\cdot} - 1)}.$$

Then the sARI for two partitions P_1, P_2 is defined as

$$\text{sARI}(P_1, P_2) = \frac{\sum_{r,c} \frac{\Gamma(p_{rc\cdot}+1)}{\Gamma(p_{rc\cdot}-1)} - \frac{1}{N(N-1)}\Lambda_{rc}}{\frac{1}{2}\Lambda_{rc} - \frac{1}{N(N-1)}\Lambda_{rc}}. \tag{3}$$

The gamma function has asymptotes at 0 and each negative integer. If any cells in Table 4 are ≤ 1, we run the risk of hitting an asymptote of the gamma function for the denominator of each term in Equation (3). To ensure sARI calculation feasibility, we present an alternative ARI calculation which we adapt for use with the sARI.

3.3 Alternative Calculation of Soft Adjusted Rand Index

As an alternative to Equation (1), Steinley (2004), showed that the classifications, type (a) - type (d) can be written as

$$a = \tfrac{1}{2}\left[\sum_r \sum_c n_{rc}^2 - N\right],\; b = \tfrac{1}{2}\left[\sum_r n_{r\cdot}^2 - \sum_r \sum_c n_{rc}^2\right],$$
$$c = \tfrac{1}{2}\left[\sum_c n_{\cdot c}^2 - \sum_r \sum_c n_{rc}^2\right],\; d = \tfrac{1}{2}\left[\sum_r \sum_c n_{rc}^2 + N^2 - \sum_r n_{r\cdot}^2 - \sum_c n_{\cdot c}^2\right],$$

and the ARI is given by

$$\mathrm{ARI}(P_1, P_2) = \frac{\binom{N}{2}(a+d) - [(a+b)(a+c) + (c+d)(b+d)]}{\binom{N}{2}^2 - [(a+b)(a+c) + (c+d)(b+d)]}. \tag{4}$$

We use this computation to calculate the sARI, replacing the integer table values from Table 1 with the posterior probabilities from Table 4 as follows

$$\tilde{a} = \tfrac{1}{2}\left[\sum_r \sum_c p_{rc\cdot}^2 - N\right],\; \tilde{b} = \tfrac{1}{2}\left[\sum_r p_{r\cdot\cdot}^2 - \sum_r \sum_c p_{rc\cdot}^2\right],$$
$$\tilde{c} = \tfrac{1}{2}\left[\sum_c p_{\cdot c\cdot}^2 - \sum_r \sum_c p_{rc\cdot}^2\right],\; \tilde{d} = \tfrac{1}{2}\left[\sum_r \sum_c p_{rc\cdot}^2 + N^2 - \sum_r p_{r\cdot\cdot}^2 - \sum_c p_{\cdot c\cdot}^2\right],$$

We use this equivalent form of the sARI for our calculations

$$\mathrm{sARI}(P_1, P_2) = \frac{\binom{N}{2}(\tilde{a}+\tilde{d}) - [(\tilde{a}+\tilde{b})(\tilde{a}+\tilde{c}) + (\tilde{c}+\tilde{d})(\tilde{b}+\tilde{d})]}{\binom{N}{2}^2 - [(\tilde{a}+\tilde{b})(\tilde{a}+\tilde{c}) + (\tilde{c}+\tilde{d})(\tilde{b}+\tilde{d})]}. \tag{5}$$

3.4 Example

Returning to the illustrative example, rather than forcing a hard assignment from model-based clustering, we use soft cluster membership probabilities (see Section 3.1) to create Table 5, resulting in a sARI of 0.301 for agreement with the true labels. This value is in comparison to an ARI of 0.430 for the hard partition agreement. In comparing Table 5 and Table 2, we see that the diagonal cells, representing agreement, have lost magnitude. The upper right off-diagonal has increased from 8 to 10.431 and the lower left off-diagonal has increased from 9 to 11.825. Since, in this simple two-group example, off-diagonals represent disagreements, an increase in these cells represents a decrease in agreement between the true labels and the clustering. A smaller sARI value relative to the corresponding ARI reflects the uncertainty in the model-based clustering solution.

Table 5 Probability contingency table for illustrative example.

$P_1 \setminus P_2$	v_1	v_2	Sums
u_1	39.569	10.431	50
u_2	11.825	38.175	50
Sums	51.394	48.606	100

Further investigating this difference, Table 6 reproduces the ten selected observations from Table 3, italicized if misclassified. If we sum entries 0 through 2 in the first column and 5 through 7 in the second column for the original data columns, we see the contribution of the correct classifications for the soft partition is $0.987 + 0.671 + 0.533 + 0.998 + 0.757 + 0.554 = 4.5$ while the contribution for the hard partition is 6. In contrast, for the incorrect classifications, the contribution for the soft partition is the sum of entries 3 and 4 in the second column and the last two entries in the first column, giving a total of $0.544 + 0.685 + 0.529 + 0.889 = 2.647$ versus the penalty of 4 for the hard partition. So the soft partition here has lost $6 - 4.5 = 1.5$ and gained $4 - 2.647 = 1.353$ for a net loss of 0.147.

If we do this over all observations, we find that compared to the hard partition, the correct classifications in the soft partition contributed 9.087 less and the misclassifications 3.831 less for a net loss of 5.256 in overall correct classification terms, resulting in a lower sARI relative to the ARI. The overconfidence of the hard correct classifications outweighs the under-confidence in the hard misclassifications.

Small variations in the posterior probabilities will result in little to no difference in the hard classification and ARI, while they could however have a significant impact on the sARI. For example, if in Figure 1, the green multivariate Gaussian group is shifted to the left by 0.25, increasing overlap in the two classes, the ARI decreases slightly from 0.430 to 0.404 due to observation 2 additionally being misclassified. The posterior probabilities are affected by the added uncertainty from the increased cluster overlap, and the resulting sARI decreases from 0.301 to 0.212. Conversely, we can shift the green group to the right by 0.25, decreasing the overlap in the two classes, resulting in more certainty in the classification. This is reflected in both the ARI and sARI increasing to 0.513 and 0.388 respectively. The original posterior probabilities for the 10 example observations and the updated posterior probabilities, based on the two shifts are given in Table 6.

4 Simulation Study

Here we look at a variety of simulations to illustrate the difference in ARI and sARI, and how that difference is likely to change depending on varying data structures. We first consider the comparison of a soft partition to a hard partition (as with our illustrative example for model-based clustering versus true labels); and then consider the comparison of two soft partitions.

The simulated datasets are generated using the `clusterGeneration` package (Qiu and Joe, 2015) in the `R` (R Core Team, 2017) software language, which allows control of multiple aspects of mixtures of Gaussians. Qiu and Joe (2006) introduce a separation index, which measures the magnitude of the gaps between any two clusters by looking at their maximum separation on a one-dimensional projection. A specified level of separation is used pairwise for a cluster and its nearest neighboring cluster. As such, the chosen separation

Table 6 Posterior probabilities of example observations from Table 3. The partitions u_1 and u_2 are the true assignment (Figure 1 - Partition 1) while v_1 and v_2 are the model-based clustering assignment (Figure 1 - Partition 2). The different columns represent when the data are shifted together by 0.25, the original data and when the data are shifted apart by 0.25. Probabilities are italicized for misclassified observations.

		Shifted Closer		Original Data		Shifted Apart	
	Label	v_1	v_2	v_1	v_2	v_1	v_2
	0	0.969	0.031	0.987	0.013	0.995	0.005
	1	0.574	0.426	0.671	0.329	0.758	0.242
u_1	2	0.447	*0.553*	0.533	0.467	0.619	0.381
	3	0.386	*0.614*	0.456	*0.544*	0.531	0.469
	4	0.262	*0.738*	0.315	*0.685*	0.376	*0.624*
	5	0.005	0.995	0.002	0.998	0.001	0.999
	6	0.307	0.693	0.243	0.757	0.168	0.832
u_2	7	0.492	0.508	0.446	0.554	0.368	0.632
	8	*0.569*	0.431	*0.529*	0.471	0.455	0.545
	9	*0.881*	0.119	*0.889*	0.111	*0.883*	0.117

index is more of a bound for the amount of separation across all the clusters. Separation values range from -1 (no separation) to 1 (extremely well-separated) where zero indicates clusters that are just touching. Figure 4 provides example scatterplots of 2-dimensional projected clusters for 3 simulated data sets of 3 clusters and 4 variables, for separation levels -0.5, 0, 0.5.

4.1 Simulation study for one soft partition versus one hard partition

In this subsection, 100 mixtures are generated from the `clusterGeneration` package for each of 39 different levels of separation from -0.95 to 0.95. Three sets of simulations are presented with mixtures generated for either 2, 3, or 8 clusters over 3, 4, or 5 dimensions. The true labels from the generated mixtures are compared to soft partitions, produced by fitting mixtures of Gaussians using the `mclust` package (Fraley and Raftery, 2002; Scrucca et al, 2016) and choosing the solution with the optimal BIC.

Results of the three simulations can be seen in Figures 5, 6, and 7. The ARI is calculated between the true partition and the hard partition obtained by a maximum a posterior mapping of the `mclust` solution. The sARI is calculated between the true hard partition and the posterior probability matrix of cluster membership from the `mclust` solution. In all four figures, results are shown over the 39 levels of separation: the upper left (a) plot shows the average ARI and average sARI for the 100 mixtures; the upper right (b) plot illustrates the distributions of the sARI values; the lower left (c) plot has the distributions of the ARI values; and the lower right (d) plot displays the distribution of the differences between the ARI and the sARI. The general results are similar regardless of the number of clusters or variables simulated (similar results were found for other combinations of clusters and dimensions not shown in

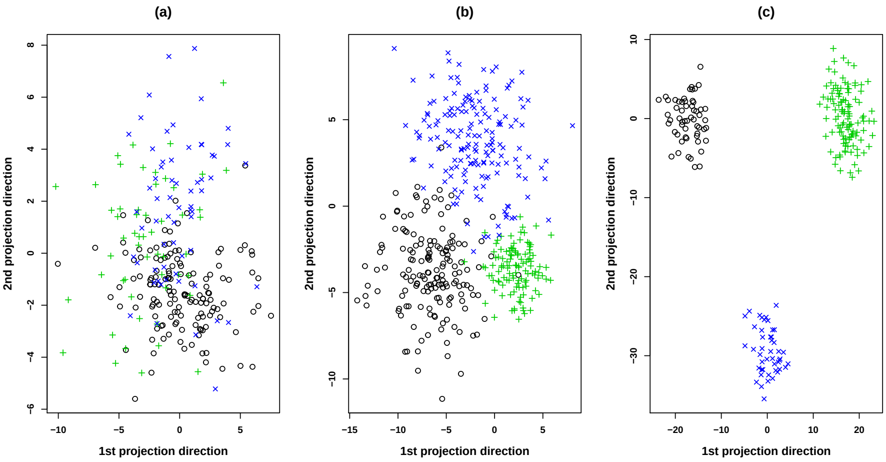

Fig. 4 Example scatterplots of 2-dimensional projected clusters (3 clusters, 4 variables): (a) separation of -0.5; (b) separation of 0; (c) separation of 0.5.

this paper). For highly overlapping clusters (separation level near -1) as well as clusters with little overlap (separation level near 1), there is high agreement between partitions with both the ARI and sARI. The non-trivial differences occur with separation levels between -0.6 and 0.05. In this range, we can expect little separation between clusters, and observations with less certain classifications. This results in sARI values that are on average less than the ARI values. Each of the simulation sets demonstrate that on average, using a hard partition with these separation levels results in a measure of agreement that is overconfident.

There are however, instances where the sARI is slightly greater than the ARI. This happens more frequently, on average in the simulations where there are fewer clusters (2 or 3 rather than 8). Fewer clusters are also associated with larger differences between the two measures of agreement. For example, in the set of simulations shown in Figure 5, across the 100 × 39 replications, approximately 3.3% of them give a negative difference between the ARI and sARI; the distribution summary of these negative differences, as well as for those from the other simulation scenarios are given in Table 7. We expect to see negative differences when there are enough observations with largest (posterior) probability on the wrong cluster, but the correct cluster still has non-trivial assignment probability. For example, when there are two clusters and observations have near 50/50 chance of being in either cluster, but are classified incorrectly more often than they are classified correctly.

Figure 8 gives two such examples, one for a separation index of 0.15 and one for a separation index of -0.30. The two simulated classes are represented with different symbol types. An observation is black if correctly classified and red if misclassified. An observation is large if the uncertainty associated with the classification is greater than 0.4 (where choosing between the two groups

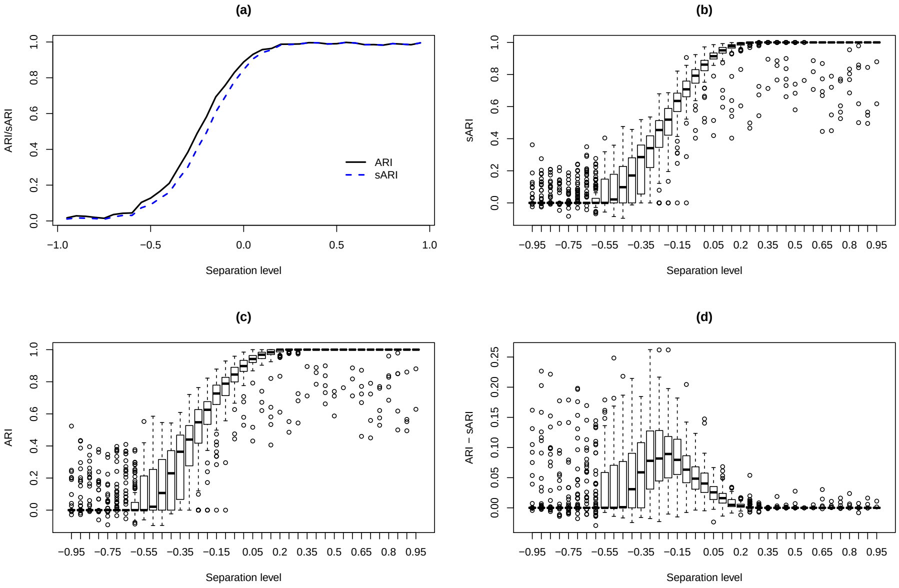

Fig. 5 Plots for varying levels of cluster separation (2 clusters, 3 variables): (a) Lines of average ARI, sARI for one hard partition and one soft partition; (b) boxplots of sARI; (c) boxplots of ARI; (d) boxplots for difference of ARI versus sARI.

is close to a "coin flip"). We would expect the sARI to be greater than the ARI generally when there are more large red observations than large black observations. For example, in the left plot, there are two misclassified observations; the smaller observation has posterior probability vector (0.847, 0.153), indicating that the (mis)classification was quite certain. However, the larger observation has posterior probability vector (0.504, 0.496), indicating that the (mis)classification was very uncertain. While the ARI considers the 2 observations fully misclassified, the sARI only considers 1.486 observations misclassified (1.351 of which comes from these two observations), thus producing a sARI (0.970) that is larger than the ARI (0.960). Similarly, in the right plot, the ARI considers 99 observations misclassified, and the sARI considers 94.823 observations misclassified resulting in an ARI of 0.173 and a sARI of 0.195.

4.2 Simulation study for two soft partitions

In this subsection, soft partitions are obtained from fitting model-based clustering models(`mclust`) to each half of the variables generated from a single mixture of Gaussians, simulated using `clusterGeneration`. Thus, the observations are the same, but the resulting pairs of soft partitions will likely be somewhat different.

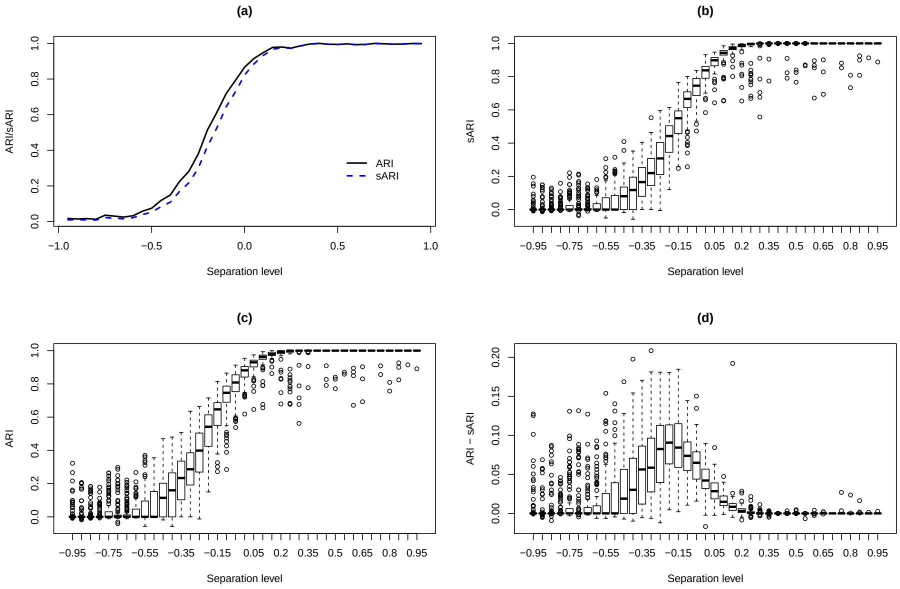

Fig. 6 Plots for varying levels of cluster separation (3 clusters, 4 variables): (a) Lines of average ARI, sARI for one hard partition and one soft partition; (b) boxplots of sARI; (c) boxplots of ARI; (d) boxplots for difference of ARI versus sARI.

Table 7 Summary statistics for the difference in ARI and sARI for replications of each simulation set where the sARI is greater than the ARI.

Simulation	Count	Min.	1st Q.	Median	3rd Q.	Max.
2 clusters:						
3 variables	130	-0.0295	-0.0076	-0.0042	-0.0019	-1.479*e*-5
4 variables	174	-0.0275	-0.0072	-0.0031	-0.0012	-1.377*e*-7
5 variables	143	-0.0279	-0.0079	-0.0035	-0.0013	-2.092*e*-6
3 clusters:						
3 variables	126	-0.0287	-0.0038	-0.0021	-0.0009	-3.883*e*-6
4 variables	126	-0.0169	-0.0036	-0.0020	-0.0007	-4.014*e*-6
5 variables	97	-0.0145	-0.0041	-0.0022	-0.0012	-3.882*e*-6
8 clusters:						
3 variables	67	-0.0058	-0.0019	-0.0011	-0.0005	-4.202*e*-5
4 variables	75	-0.0042	-0.0014	-0.0009	-0.0004	-1.391*e*-5
5 variables	141	-0.0049	-0.0014	-0.0006	-0.0002	-1.462*e*-6

4.2.1 2 cluster, 8 variable model with varying separation

We generate 100 2 cluster, 8 variable datasets for the same 39 levels of separation. We then randomly split each dataset into two sets of 4 variables and fit model-based clustering models to each variable set. There are three

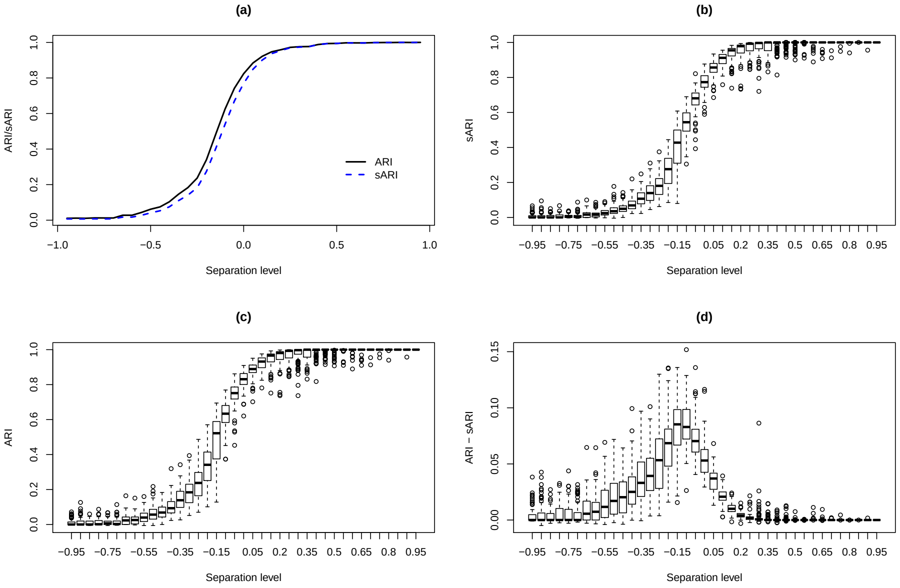

Fig. 7 Plots for varying levels of cluster separation (8 clusters, 5 variables): (a) Lines of average ARI, sARI for one hard partition and one soft partition; (b) boxplots of sARI; (c) boxplots of ARI; (d) boxplots for difference of ARI versus sARI.

possibilities to consider: comparing the two original soft partitions via sARI, transforming one of the partitions to a hard partition and comparing it to the other soft partition via sARI, and finally the standard approach of transforming both partitions to hard partitions and comparing via ARI. Looking at the ARI versus 1 hard/1 soft partition sARI versus 2 soft partitions sARI allows us to assess how each successive transformation affects the level of similarity estimated between clusterings.

We first allow the number of clusters found in each solution to be selected via BIC. In Figure 9, we see in (a) the average levels of the indices as separation changes (increasing from left to right). There is almost perfect overlap between the different indices and hard/soft partitions from around -0.4 and lower. The line is also trending back up to the left which seems odd at first glance, given the increasing level of overlap, until you look at the plots in (b) and (c). Investigating further, we find that as the overlap increases between the two clusters, both solutions find 1 cluster (ARI/sARI of 1), or one solutions finds 1 cluster, while the other finds 2 (ARI/sARI of 0). The proportion of times both clusterings find only 1 cluster increases as separation decreases.

In order to properly contrast the behavior between sARI and ARI without this issue, Figure 10 presents results where we fix the number of clusters found by `mclust` to 2. In Figure 10 (a), the lines are mostly monotonically increasing from left to right as expected.

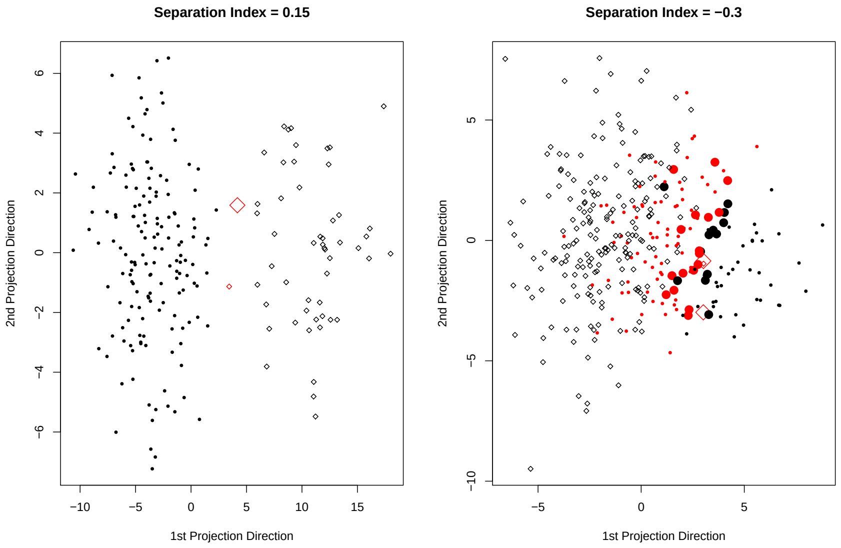

Fig. 8 Example plots demonstrating simulated data from 2 clusters and 3 variables where the ARI is less than the sARI. The left plot is for a separation index of 0.15 and the right plot for a separation index of -0.3. Observation symbols represent the two simulated classes. An observation is colored red if it was misclassified by model-based clustering. An observation is large if it has uncertainty above 0.4.

For both sARI lines, as the ARI tends to 1, the difference between the ARI and sARI approaches 0. The sARI and ARI differ most over a range of reasonable (but not total) cluster overlap values. On average, the ARI is ranked over sARI for one soft and one hard partition, which is itself ranked over the sARI for two soft partitions. Plot (d) in both Figures 9 and 10 show that ARI is almost always above the sARI (with a few exceptions).

5 Application

In this section, we apply the sARI to the diabetes data, a classic clustering benchmarking data set, available in the `mclust` package in `R`. This data has been used as an example in many model-based clustering papers including Banfield and Raftery (1993), Fraley and Raftery (2007), and Scrucca et al (2016). This data set contains diabetes diagnoses (classified as chemical, normal, or overt) on 145 patients based on their blood plasma and insulin levels measured under 3 conditions. A pairs plot of the data can be seen in Figure 11.

The data set is clustered using mixtures of Gaussians fit with the `mclust` package. Results are presented for the clustering solutions from two different parameterized models. We use the 3 cluster VVV (volume, shape and orientation allowed to vary across the ellipsoidal clusters) model, which is the optimal

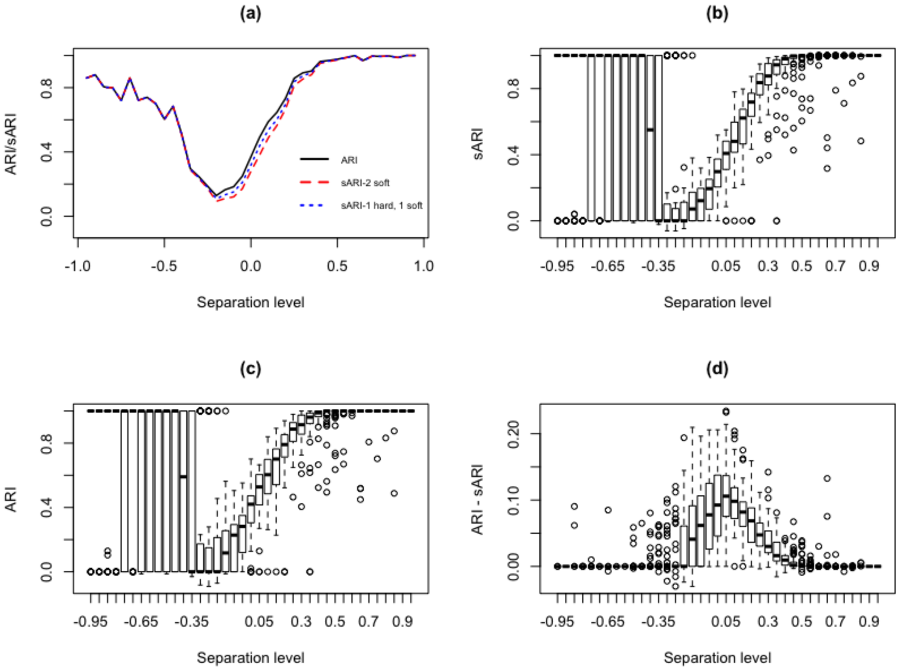

Fig. 9 Plots for varying levels of cluster separation for 2 clusters (`mclust` allowed to use BIC to select the number of clusters in each solution): (a) Lines of average ARI, sARI for 1 hard and 1 soft partition, sARI for 2 soft partitions; (b) boxplots of sARI for 2 soft partitions; (c) boxplots of ARI; (d) boxplots for difference of ARI versus sARI (for 2 soft partitions).

model chosen by the BIC. We also use the solution for an alternative parameterization of EEI (equal volume and shape, with orientation parallel to the coordinate axes) with 9 clusters (optimal BIC for this parameterization). The hard and soft clustering solutions from each model are compared to the diagnosis provided in the data and to each other. The resulting ARI and sARI's are summarized in Table 8.

Comparing the hard 3 cluster solution of Gaussian mixtures with parameterization VVV to the given diagnoses results in an ARI of 0.664 (contingency table given in Table 9). When comparing the soft clustering solution for this same model to the given diagnoses, the sARI drops to 0.602 (contingency table given in Table 10). The comparison of the hard 9 cluster solution for Gaussian mixtures with parameterization EEI to the given diagnoses produces an ARI of 0.564. Comparing the soft clustering solution for this model to the diagnoses results in a sARI of 0.381. The hard partitions produced by the two models are quite similar to each other (ARI = 0.799), whereas the soft partitions only have moderate similarity (sARI = 0.459).

As expected from the simulation results, when clustering the diabetes data, a hard partition from either clustering solution overestimates the similarity

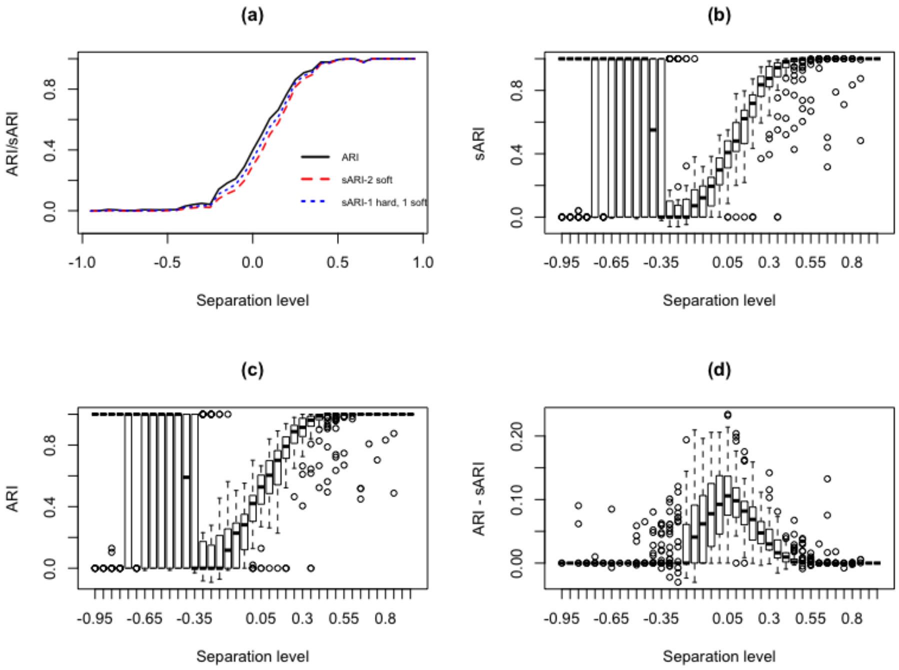

Fig. 10 Plots for varying levels of cluster separation for 2 cluster solutions (where `mclust` was constrained to fit 2 clusters in each solution): (a) Lines of average ARI, sARI for 1 hard and 1 soft partition, sARI for 2 soft partitions; (b) boxplots of sARI for 2 soft partitions; (c) boxplots of ARI; (d) boxplots for difference of ARI versus sARI (for 2 soft partitions).

between the classification and the true diagnosis, as seen by comparing the hard partition and soft partition columns of Table 8. Having a sARI that is substantially less than the ARI indicates that there is non-trivial overlap in the cluster boundaries, producing uncertainty in the classification. This information might also indicate that the three variables do not identify the final diagnoses as well as the ARI might imply, leading us to conclude that either the diagnosis classes are non-separable or that a search for more useful variables is necessary.

6 Discussion

The proposal of the soft adjusted Rand index (sARI) is not intended to supplant the adjusted Rand index (ARI) but to be used in conjunction with it. The use of both can help the user understand what effect the transformation of soft partition to hard partition can have on the comparison with other clustering results and more accurately reflects the uncertainty in a set of cluster/class assignments. In the vast majority of cases, sARI will be smaller than

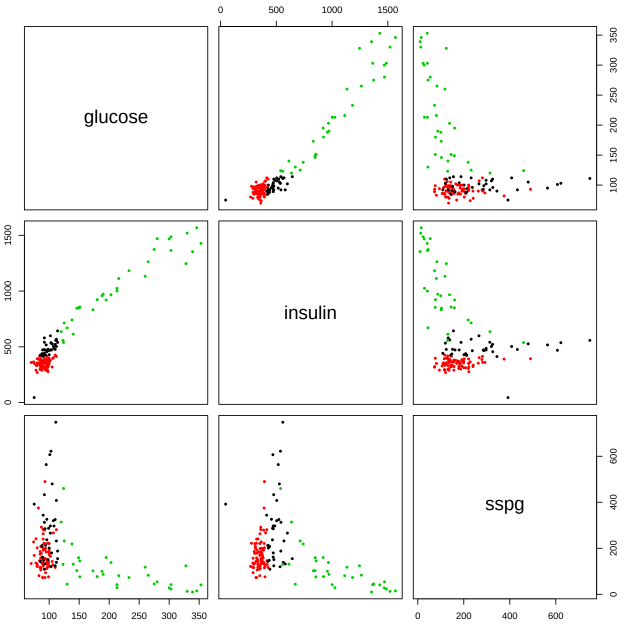

Fig. 11 Pairs plot of the diabetes data colored by diagnosis (chemical - black, normal - red, overt - green).

Table 8 Agreement between different Gaussian mixture model clustering solutions for the diabetes data.

		Model VVV		
			Hard	Soft
		Diagnosis	Partition	Partition
	Diagnosis	1	0.664	0.602
Model EEI	Hard Partition	0.564	0.799	0.695
	Soft Partition	0.381	0.514	0.459

ARI, due to the overconfidence in hard correct classifications outweighing the overconfidence in hard misclassifications. In a small proportion of cases, where the overlap between clusters is such that the assignment is almost random, sARI can occasionally be slightly larger than ARI.

Obviously this idea can be extended to other indices of partition similarity by simple analogue, e.g. Fowlkes-Mallows (Fowlkes and Mallows, 1983), Jaccard (Jaccard, 1901), etc. There is also the potential to extend beyond the idea of independence between two partitions when constructing the proportions in the probability classification tables. Co-occurrence matrices from Bayesian inference for clustering are one area where that might be achieved.

Table 9 Contingency table for diabetes diagnosis and hard clustering solution for 3 component (VVV) mixture of Gaussians.

	1	2	3
Normal	72.00	4.00	0.00
Chemical	9.00	26.00	1.00
Overt	0.00	6.00	27.00

Table 10 Contingency table for diabetes diagnosis and soft clustering solution for 3 component (VVV) mixture of Gaussians.

	1	2	3
Normal	68.89	7.06	0.48
Chemical	8.96	25.36	1.67
Overt	0.00	6.00	27.00

References

Amodio S, D'Ambrosio A, Iorio C, Siciliano R (2015) Adjusted Concordance Index, an extension of the Adjusted Rand index to fuzzy partitions. arXiv preprint arXiv:150900803

Banfield JD, Raftery AE (1993) Model-based Gaussian and non-Gaussian clustering. Biometrics pp 803–821

Bezdek JC (1981) Objective function clustering. In: Pattern Recognition with Fuzzy Objective Function Algorithms, Springer, pp 43–93

Campello RJGB (2007) A fuzzy extension of the Rand index and other related indexes for clustering and classification assessment. Pattern Recognition Letters 28:833 – 841

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society Series B (Methodological) 39(1):1 – 38

Downton M, Brennan T (1980) Comparing classifications: an evaluation of several coefficients of partition agreement. Classification Society Bulletin 4(4):53–54

Dunn JC (1973) A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. Journal of Cybernetics 3(3):32 – 57

Fowlkes EB, Mallows CL (1983) A method for comparing two hierarchical clusterings. Journal of the American Statistical Association 78(383):553–569

Fraley C, Raftery AE (2002) Model-based clustering, discriminant analysis, and density estimation. Journal of the American Statistical Association 97(458):611–631

Fraley C, Raftery AE (2007) Model-based methods of classification: using the mclust software in chemometrics. Journal of Statistical Software 18(6):1–13

Hartigan JA (1975) Clustering Algorithms. Wiley & Sons, New York

Hubert L, Arabie P (1985) Comparing partitions. Journal of Classification 2(1):193–218

Huellermeyer E, Rifqi M, Henzgen S, Senge R (2012) Comparing fuzzy partitions: A generalization of the Rand index and related measures. IEEE Transactions on Fuzzy Systems 20(3):546–556

Jaccard P (1901) Étude comparative de la distribution florale dans une portion des alpes et du jura. Bulletin de la Société Vaudoise des Sciences Naturelles 37(142):547–579

MacQueen J, et al (1967) Some methods for classification and analysis of multivariate observations. In: Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, Oakland, CA, USA., vol 1, pp 281–297

McLachlan G, Krishnan T (2007) The EM Algorithm and Extensions, vol 382. John Wiley & Sons

McLachlan G, Peel D (2004) Finite Mixture Models. John Wiley & Sons

McNicholas PD (2016) Model-based clustering. Journal of Classification 33(3):331–373

Miyamoto S, Ichihashi H, Honda K (2008) Algorithms for Fuzzy Clustering. Springer

Morey LC, Agresti A (1984) The measurement of classification agreement: An adjustment to the Rand statistic for chance agreement. Educational and Psychological Measurement 44(1):33–37

Qiu W, Joe H (2006) Separation index and partial membership for clustering. Computational Statistics and Data Analysis 50(3):585–603

Qiu W, Joe H (2015) clusterGeneration: Random Cluster Generation (with Specified Degree of Separation). R package version 1.3.4, https://CRAN.R-project.org/package=clusterGeneration

R Core Team (2017) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria

Rand WM (1971) Objective criteria for the evaluation of clustering methods. Journal of the American Statistical Association 66(336):846–850

Scrucca L, Fop M, Murphy TB, Raftery AE (2016) mclust 5: Clustering, classification and density estimation using gaussian finite mixture models. The R Journal 8(1):289

Steinley D (2004) Properties of the Hubert-Arabie adjusted Rand index. Psychological methods 9(3):386

Ward JH (1963) Hierarchical grouping to optimize an objective function. Journal of the American Statistical Association 58(301):236–244

Wolfe JH (1963) Object cluster analysis of social areas. PhD thesis, University of California