

A Separability Index for Distance-based Clustering and Classification Algorithms

Arka P. Ghosh, Ranjan Maitra and Anna D. Peterson

Abstract—We propose a separability index that quantifies the degree of difficulty in a hard clustering problem under assumptions of a multivariate Gaussian distribution for each cluster. A preliminary index is first defined and several of its properties are explored both theoretically and numerically. Adjustments are then made to this index so that the final refinement is also interpretable in terms of the Adjusted Rand Index between a true grouping and its hypothetical idealized clustering, taken as a surrogate of clustering complexity. Our derived index is used to develop a data-simulation algorithm that generates samples according to the prescribed value of the index. This algorithm is particularly useful for systematically generating datasets with varying degrees of clustering difficulty which can be used to evaluate performance of different clustering algorithms. The index is also shown to be useful in providing a summary of the distinctiveness of classes in grouped datasets.

Index Terms—Separation index, multivariate Jaccard index, clustering complexity, exact- c -separation, radial visualization plots



1 INTRODUCTION

There is a large body of literature [1], [2], [3], [4], [5], [6], [7], [8], [9] dedicated to clustering observations in a dataset into homogeneous groups, but no method uniformly outperforms the others. Many algorithms perform well in some settings but not so otherwise. Further, the settings where algorithms work well or poorly is very often not quite understood. Thus there is need for a systematic study of performance of any clustering algorithm, and also for evaluating effectiveness of new methods using the same objective criterion.

Many researchers evaluate the performance of a proposed clustering technique by comparing its performance on classification datasets like textures [10], wine [11], Iris [12], crabs [13], image [14], *E. coli* [15] or [16]’s dataset. While evaluating a clustering algorithm through its performance on select classification datasets is useful, it does not provide a systematic and comprehensive understanding of its strengths and weaknesses over many scenarios. Thus, there is need for ways to simulate datasets of different clustering difficulty and calibrating the performance of a clustering algorithm under different conditions. In order to do so, we need to index the clustering complexity of a dataset appropriately.

There have been some attempts at generating clustered data of different clustering complexity in terms of “separability indices”. [17] proposed a much-used algorithm [18], [19], [20], [21], [22] that generates “well-separated” clusters from

normal distributions over bounded ranges, with provisions for including “scatter” [23], non-informative dimensions, outliers. However, [24] observed that both increasing the variance and adding outliers increases the degree of overlap in unpredictable and differing ways, and thus, this method is incapable of accurately generating indexed clustered data. [25] proposed the OCLUS algorithm for generating clusters based on known (asymptotic) overlap by having the user provide a “design matrix” and an “order matrix” – the former indicates the (at most) triplets of clusters that are desired to be overlapping with each other while the latter dictates the ordering of clusters in each dimension. Although, the idea of using overlap in generating clustered data is appealing, the algorithm has constraints beyond the structure of the design matrix above: for instance, independence between dimensions is also required.

A separation index between any two univariate Gaussian clusters was proposed by [26]. For higher dimensions, they also used the same index on the 1-D transformation obtained after optimally projecting the two multivariate normal clusters onto 1-D space. For multiple clusters, their overall separation index (also the basis for their cluster generation algorithm [27]) is the maximum of all $\binom{n}{2}$ pairwise indices and thus, quite impervious to variations between other groups that are not in this maximizing pair. Additionally, characterizing separation between multi-dimensional clusters by means of the best 1-D projection loses substantial information: thus, resulting statements on cluster overlap can be very misleading. Finding the optimal 1-D projection is also computationally intensive and impractical for very high dimensions.

[28], [29], [30] demonstrated performance of their clustering algorithms using simulation datasets generated using the concept (or a variant) of c -separation between clusters proposed by [31], which defines two Gaussian distributions $N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and $N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$ in \mathbb{R}^n as c -separated if $\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\| \geq c\sqrt{n \max(\lambda_{max}(\boldsymbol{\Sigma}_1), \lambda_{max}(\boldsymbol{\Sigma}_2))}$ where $\lambda_{max}(\boldsymbol{\Sigma}_i)$ is the largest eigenvalue of $\boldsymbol{\Sigma}_i$. [32] formalized the concept to

• The authors are with the Department of Statistics, Iowa State University, Ames, IA 50011-1210, USA.

©2010 IEEE. Personal use of this material is permitted. However, permission to use this material for any other purposes must be obtained from the IEEE by sending a request to pubs-permissions@ieee.org.

Manuscript received June 21, 2010; revised October 21, 2010. First published xxxxxxxx x, xxxx, current version published yyyyyyyy y, yyyy

R. Maitra’s and A. D. Peterson’s research was partially supported by the National Science Foundation under Grant Nos. CAREER DMS-0437555 and VIGRE 0091953, respectively.

Digital Object Identifier

exact- c -separation by requiring equality for at least one pair (i, j) of clusters and used it for generating datasets to calibrate some partitional initialization algorithms. He also pointed out some inherent shortcomings that originate from ignoring the relative orientations of the cluster dispersions.

More recently, [33] proposed a method for generating Gaussian mixture distributions according to some summary measure of overlap between every component pair, defined as the unweighted sum of the probabilities of their individual misclassification rates. They also provided open-source C software (C-MixSim) and a R package (MixSim) for generating clusters corresponding to desired overlap characteristics. In contrast to many of the existing indices and simulation algorithms, their methodology does not impose any restriction on the parameters of the distributions but was derived entirely in the context of mixture models and model-based clustering algorithms. Thus, their methods were specifically geared toward soft clustering and model-based clustering scenarios.

In this paper, we complement [33]'s scenario and derive a separation index (Section 2) for distance-based partitioning and hard clustering algorithms. Our index is motivated by the intuition that for any two well-separated groups, the majority of observations should be closer to their own center than to the other. We use Gaussian-distributed clusters and Euclidean and Mahalanobis distances to simplify our theoretical calculations. The preliminary index is investigated and fine-tuned in the context of homogeneous spherical clusters in Section 3.1 and then extended to the case for multiple groups. The methodology is then studied for the general case in Section 3.2. Our derived index can be used to quantify class distinctions in grouped data, and we illustrate this application in Section 4 in the context of several classification datasets. The main paper concludes with some discussion. The paper also has an appendix detailing the algorithm that uses our index to generate datasets of desired clustering complexity.

2 METHODOLOGICAL DEVELOPMENT

Consider a dataset $\mathcal{S} = \{\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n\}$. The objective of hard clustering or fixed-partitioning algorithms is to group the observations into hard categories $\mathcal{C}_1, \mathcal{C}_2, \dots, \mathcal{C}_K$ such that some objective function measuring the quality of fit is minimized. If the objective function is specified in terms of minimizing the total distance of each observation from some characteristic of the assigned partition, then it is defined as

$$\mathcal{O}_K = \sum_{i=1}^n \sum_{k=1}^K I(\mathbf{X}_i \in \mathcal{C}_k) \mathcal{D}_k(\mathbf{X}_i) \quad (1)$$

where $\mathcal{D}_k(\mathbf{X}_i)$ is the distance of \mathbf{X}_i from the center of the k -th partition, and assumed to be of the form

$$\mathcal{D}_k(\mathbf{X}_i) = (\mathbf{X}_i - \boldsymbol{\mu}_k)' \boldsymbol{\Delta}_k (\mathbf{X}_i - \boldsymbol{\mu}_k) \quad (2)$$

where $\boldsymbol{\Delta}_k$ is a non-negative definite matrix of dimension $p \times p$. We consider two special cases here: in the first case, $\boldsymbol{\Delta}_k = \mathbf{I}_p$, the identity matrix of order $p \times p$ where $\mathcal{D}_k(\mathbf{X}_i)$ reduces to the squared Euclidean distance and solving for (1) involves finding partitions $\mathcal{C}_1, \mathcal{C}_2, \dots, \mathcal{C}_K$ in \mathcal{S} such that the sum of the squared Euclidean distance of each observation to the center

of its assigned partition is minimized. The popular k -means algorithm [34], [35] provides locally optimal solutions to this minimization problem. In the second scenario, $\boldsymbol{\Delta}_k = \boldsymbol{\Sigma}_k^{-1}$, where $\boldsymbol{\Sigma}_k$ is the dispersion matrix of the k th partition. Then $\mathcal{D}_k(\mathbf{X}_i)$ is the Mahalanobis distance between \mathbf{X}_i and $\boldsymbol{\mu}_k$.

In this paper, we adopt a convenient model-based formulation for the setup above. In this formulation, $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n$ are independent p -variate observations with $\mathbf{X}_i \sim N_p(\boldsymbol{\mu}_{\zeta_i}, \boldsymbol{\Sigma}_{\zeta_i})$, where $\zeta_i \in \{1, 2, \dots, K\}$ for $i = 1, 2, \dots, n$. Here we assume that $\boldsymbol{\mu}_k$'s are all distinct and that n_k is the number of observations in cluster k . Then the density for the \mathbf{X}_i 's is given by $f(\mathbf{X}) = \sum_{k=1}^K I(\mathbf{X} \in \mathcal{C}_k) \phi(\mathbf{X}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, where \mathcal{C}_k is the subpopulation indexed by the $N_p(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ density and $I(\mathbf{X} \in \mathcal{C}_k)$ is an indicator function specifying whether observation \mathbf{X} belongs to the k th group having a p -dimensional multivariate normal density $\phi(\mathbf{X}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \propto |\boldsymbol{\Sigma}_k|^{-\frac{p}{2}} \exp(-\frac{1}{2}(\mathbf{X} - \boldsymbol{\mu}_k)' \boldsymbol{\Sigma}_k^{-1} (\mathbf{X} - \boldsymbol{\mu}_k))$, $k = 1, \dots, K$. When $\boldsymbol{\Sigma}_k = \mathbf{I}_p$, maximizing the loglikelihood with respect to the parameters $\zeta_i, i = 1, 2, \dots, n$ and $\boldsymbol{\mu}_k$'s is equivalent to solving for (1) with $\mathcal{D}_k(\mathbf{X}_i)$ as Euclidean distance. Using this formalism, we develop, in theoretical terms, a separation index that quantifies separation between any two clusters and relates it to the difficulty in recovering the true partition $\mathcal{C}_1, \mathcal{C}_2, \dots, \mathcal{C}_K$ of a dataset. We begin by defining a preliminary index.

2.1 A preliminary separation index between two clusters

Consider the case with $K = 2$ groups, labeled \mathcal{C}_j and \mathcal{C}_l for $(j \neq l \in \{1, 2\})$. Define

$$Y^{j,l}(\mathbf{X}) = \mathcal{D}_j(\mathbf{X}) - \mathcal{D}_l(\mathbf{X}), \quad \text{where } \mathbf{X} \in \mathcal{C}_l, \quad (3)$$

and $Y^{l,j}(\mathbf{X})$, similarly. Using the modeling formulation above, $Y^{j,l}(\mathbf{X})$ is a random variable which represents the difference in squared distances of $\mathbf{X} \in \mathcal{C}_l$ to the center of \mathcal{C}_j and to the center of \mathcal{C}_l . For distance-based classification methods, $\Pr[Y^{j,l}(\mathbf{X}) < 0]$ is the probability that an observation is classified into \mathcal{C}_j when in fact the observation is from \mathcal{C}_l . Intuitively, one expects that since $\mathbf{X}_{\zeta_i} : \zeta_i = l, i = 1, \dots, n$, belong to \mathcal{C}_l , then most of these n_l observations will be closer to the mean of \mathcal{C}_l , compared to the mean of \mathcal{C}_j . Based on this observation, we define the index in terms of the probability that α fraction of the observations are closer to the incorrect cluster center. In other words, we find the probability that an order statistic (say, $\lfloor n_l \alpha \rfloor$ -th, for $\alpha \in (0, 1)$) of $\{Y_{\zeta_i}^{j,l}(\mathbf{X}) : \zeta_i = l, i = 1, \dots, n\}$ is less than 0. (Here $\lfloor x \rfloor$ is the greatest integer smaller than x). We specify this probability as $p^{j,l}$. To simplify notation we will assume that $n_l \alpha$ is an integer. Therefore,

$$p^{j,l} = \sum_{i=n_l \alpha}^{n_l} \binom{n_l}{i} \Pr[Y^{j,l}(\mathbf{X}) < 0]^i \Pr[Y^{j,l}(\mathbf{X}) > 0]^{n_l - i}, \quad (4)$$

and $p^{l,j}$ is defined similarly. Since both these probabilities can be extremely small we take the average to an inverse power of a function of n_j and n_l . We define the pairwise index as

$$\mathcal{I}^{j,l} = 1 - \left(\frac{1}{2} (p^{l,j} + p^{j,l}) \right)^{\frac{1}{n_{j,l}}}, \quad (5)$$

where $n_{j,l} = n_{l,j} = \sqrt{(n_l^2 + n_j^2)/2}$. Note that this index incorporates class sizes into it, which is desirable since misclassification rates are affected by the relative sizes of groups. Also, when $n_j = n_l = n$ then $n_{j,l} = n$. The index $\mathcal{I}^{j,l}$ takes values in $[0, 1]$, with a value close to unity indicating a well-separated dataset and values closer to zero indicating that the two groups have substantial overlap. We call the index in (5) the preliminary index, because we will make suitable adjustments to this obtained through simulations in Section 3. The calculation of \mathcal{I} requires the knowledge of the distribution of the random variables defined in (3), which is summarized in the Theorem below for different cases. Here and for the remainder of this paper, the notation $\mathbf{X} \stackrel{d}{=} \mathbf{Y}$ means that \mathbf{X} has the same distribution as \mathbf{Y} .

Theorem 1: For any $l, j \in \{1, \dots, K\}$, let $Y^{j,l}(\mathbf{X})$ be as defined in (3), for some distance metric \mathcal{D}_j and \mathcal{D}_l as defined in (2). Further assume $\mathbf{X} \sim N_p(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$. Then for specific choices of $\boldsymbol{\Delta}_l, \boldsymbol{\Delta}_j$ and for $\boldsymbol{\mu}_{lj} = \boldsymbol{\mu}_l - \boldsymbol{\mu}_j$ the following hold:

- For the Euclidean distance ($\boldsymbol{\Delta}_j = \boldsymbol{\Delta}_l = \mathbf{I}$) $Y^{j,l}(\mathbf{X}) \sim N(\boldsymbol{\mu}'_{lj} \boldsymbol{\mu}_{lj}, 4\boldsymbol{\mu}'_{lj} \boldsymbol{\Sigma}_l \boldsymbol{\mu}_{lj})$.
- For the Mahalanobis distance ($\boldsymbol{\Delta}_j = \boldsymbol{\Sigma}_j^{-1}, \boldsymbol{\Delta}_l = \boldsymbol{\Sigma}_l^{-1}$), when both clusters have identical covariance structures, i.e. $\boldsymbol{\Sigma}_l = \boldsymbol{\Sigma}_j \equiv \boldsymbol{\Sigma}$, then $Y^{j,l}(\mathbf{X}) \sim N(\boldsymbol{\mu}'_{lj} \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_{lj}, 4\boldsymbol{\mu}'_{lj} \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_{lj})$.
- For the Mahalanobis distance ($\boldsymbol{\Delta}_j = \boldsymbol{\Sigma}_j^{-1}, \boldsymbol{\Delta}_l = \boldsymbol{\Sigma}_l^{-1}$), when the two clusters do not have identical covariance structures (i.e. $\boldsymbol{\Sigma}_l \neq \boldsymbol{\Sigma}_j$), let $\lambda_1, \lambda_2, \dots, \lambda_p$ be the eigenvalues of $\boldsymbol{\Sigma}_{j|l} \equiv \boldsymbol{\Sigma}_l^{-\frac{1}{2}} \boldsymbol{\Sigma}_j^{-1} \boldsymbol{\Sigma}_l^{\frac{1}{2}}$ and the corresponding eigenvectors be given by $\gamma_1, \gamma_2, \dots, \gamma_p$. Then $Y^{j,l}(\mathbf{X})$ has the distribution of $\sum_{i=1:p, \lambda_i \neq 1} \frac{(\lambda_i - 1)^2 U_i - \lambda_i \delta_i^2}{\lambda_i - 1} + \sum_{i=1:p, \lambda_i = 1} \lambda_i \delta_i (2W_i + \delta_i)$, where U_i 's are independent non-central χ^2 random variables with one degree of freedom and non-centrality parameter given by $\lambda_i^2 \delta_i^2 / (\lambda_i - 1)^2$ with $\delta_i = \boldsymbol{\gamma}'_i \boldsymbol{\Sigma}_l^{-\frac{1}{2}} (\boldsymbol{\mu}_l - \boldsymbol{\mu}_j)$ for $i \in \{1, 2, \dots, p\} \cap \{i : \lambda_i \neq 1\}$, independent of W_i 's, which are independent $N_p(0, 1)$ random variables, for $i \in \{1, 2, \dots, p\} \cap \{i : \lambda_i = 1\}$.

Proof: For any p -variate vector \mathbf{X} ,

$$\begin{aligned} & \mathcal{D}_j(\mathbf{X}) - \mathcal{D}_l(\mathbf{X}) \\ &= (\mathbf{X} - \boldsymbol{\mu}_j)' \boldsymbol{\Delta}_j (\mathbf{X} - \boldsymbol{\mu}_j) - (\mathbf{X} - \boldsymbol{\mu}_l)' \boldsymbol{\Delta}_l (\mathbf{X} - \boldsymbol{\mu}_l) \\ &= \mathbf{X}' (\boldsymbol{\Delta}_j - \boldsymbol{\Delta}_l) \mathbf{X} + 2\mathbf{X}' (\boldsymbol{\Delta}_l \boldsymbol{\mu}_l - \boldsymbol{\Delta}_j \boldsymbol{\mu}_j) \\ & \quad + (\boldsymbol{\mu}'_j \boldsymbol{\Delta}_j \boldsymbol{\mu}_j - \boldsymbol{\mu}'_l \boldsymbol{\Delta}_l \boldsymbol{\mu}_l). \end{aligned} \quad (6)$$

Therefore, when $\boldsymbol{\Delta}_j = \boldsymbol{\Delta}_l = \mathbf{I}$, $Y^{j,l}(\mathbf{X}) = \mathcal{D}_j(\mathbf{X}) - \mathcal{D}_l(\mathbf{X}) = 2\mathbf{X}' (\boldsymbol{\mu}_l - \boldsymbol{\mu}_j) + \boldsymbol{\mu}'_j \boldsymbol{\mu}_j - \boldsymbol{\mu}'_l \boldsymbol{\mu}_l$, where $\mathbf{X} \sim N_p(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)$. Simple algebra completes the proof of part (a).

Similar calculation using (6), when $\boldsymbol{\Delta}_j = \boldsymbol{\Delta}_l = \boldsymbol{\Sigma}^{-1}$, shows that $Y^{j,l}(\mathbf{X}) = 2\mathbf{X}' (\boldsymbol{\mu}_l - \boldsymbol{\mu}_j) + \boldsymbol{\mu}'_j \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_j - \boldsymbol{\mu}'_l \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_l$, where $\mathbf{X} \sim N_p(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)$. This completes the proof of part

(b) using similar arguments as above. For part (c), let $\boldsymbol{\xi} \sim N_p(0, \mathbf{I})$. Since $\mathbf{X} \stackrel{d}{=} \boldsymbol{\Sigma}_l^{\frac{1}{2}} \boldsymbol{\xi} + \boldsymbol{\mu}_l$, using (6), we have

$$\begin{aligned} Y^{j,l}(\mathbf{X}) &= \mathbf{X}' (\boldsymbol{\Sigma}_j^{-1} - \boldsymbol{\Sigma}_l^{-1}) \mathbf{X} + 2\mathbf{X}' (\boldsymbol{\Sigma}_l^{-1} \boldsymbol{\mu}_l - \boldsymbol{\Sigma}_j^{-1} \boldsymbol{\mu}_j) \\ & \quad + \boldsymbol{\mu}'_j \boldsymbol{\Sigma}_j^{-1} \boldsymbol{\mu}_j - \boldsymbol{\mu}'_l \boldsymbol{\Sigma}_l^{-1} \boldsymbol{\mu}_l \\ & \stackrel{d}{=} (\boldsymbol{\Sigma}_l^{\frac{1}{2}} \boldsymbol{\xi} + \boldsymbol{\mu}_l)' (\boldsymbol{\Sigma}_j^{-1} - \boldsymbol{\Sigma}_l^{-1}) (\boldsymbol{\Sigma}_l^{\frac{1}{2}} \boldsymbol{\xi} + \boldsymbol{\mu}_l) \\ & \quad + 2(\boldsymbol{\Sigma}_l^{\frac{1}{2}} \boldsymbol{\xi} + \boldsymbol{\mu}_l)' (\boldsymbol{\Sigma}_l^{-1} \boldsymbol{\mu}_l - \boldsymbol{\Sigma}_j^{-1} \boldsymbol{\mu}_j) \\ & \quad + \boldsymbol{\mu}'_j \boldsymbol{\Sigma}_j^{-1} \boldsymbol{\mu}_j - \boldsymbol{\mu}'_l \boldsymbol{\Sigma}_l^{-1} \boldsymbol{\mu}_l \\ & = \boldsymbol{\xi}' (\boldsymbol{\Sigma}_{j|l} - \mathbf{I}) \boldsymbol{\xi} + 2\boldsymbol{\xi}' (\boldsymbol{\Sigma}_l^{\frac{1}{2}} \boldsymbol{\Sigma}_j^{-1}) (\boldsymbol{\mu}_l - \boldsymbol{\mu}_j) \\ & \quad + (\boldsymbol{\mu}_l - \boldsymbol{\mu}_j)' (\boldsymbol{\Sigma}_j^{-1}) (\boldsymbol{\mu}_l - \boldsymbol{\mu}_j) \end{aligned} \quad (7)$$

where $\boldsymbol{\Sigma}_{j|l} = \boldsymbol{\Sigma}_l^{\frac{1}{2}} \boldsymbol{\Sigma}_j^{-1} \boldsymbol{\Sigma}_l^{\frac{1}{2}}$. Let the spectral decomposition of $\boldsymbol{\Sigma}_{j|l}$ be given by $\boldsymbol{\Sigma}_{j|l} = \boldsymbol{\Gamma}_{j|l} \boldsymbol{\Lambda}_{j|l} \boldsymbol{\Gamma}'_{j|l}$, where $\boldsymbol{\Lambda}_{j|l}$ is a diagonal matrix containing the eigenvalues $\lambda_1, \lambda_2, \dots, \lambda_p$ of $\boldsymbol{\Sigma}_{j|l}$, and $\boldsymbol{\Gamma}_{j|l}$ is an orthogonal matrix containing the eigenvectors $\gamma_1, \gamma_2, \dots, \gamma_p$ of $\boldsymbol{\Sigma}_{j|l}$. Since $\mathbf{Z} \equiv \boldsymbol{\Gamma}_{j|l}' \boldsymbol{\xi} \sim N_p(0, \mathbf{I})$ as well, we get from (7) that

$$\begin{aligned} Y^{j,l}(\mathbf{X}) & \stackrel{d}{=} \boldsymbol{\xi}' (\boldsymbol{\Gamma}_{j|l} \boldsymbol{\Lambda}_{j|l} \boldsymbol{\Gamma}'_{j|l} - \boldsymbol{\Gamma}_{j|l} \boldsymbol{\Gamma}'_{j|l}) \boldsymbol{\xi} \\ & \quad + 2\boldsymbol{\xi}' (\boldsymbol{\Gamma}_{j|l} \boldsymbol{\Lambda}_{j|l} \boldsymbol{\Gamma}'_{j|l} \boldsymbol{\Sigma}_l^{-\frac{1}{2}}) (\boldsymbol{\mu}_l - \boldsymbol{\mu}_j) \\ & \quad + (\boldsymbol{\mu}_l - \boldsymbol{\mu}_j)' (\boldsymbol{\Sigma}_l^{-\frac{1}{2}} \boldsymbol{\Gamma}_{j|l} \boldsymbol{\Lambda}_{j|l} \boldsymbol{\Gamma}'_{j|l} \boldsymbol{\Sigma}_l^{-\frac{1}{2}}) (\boldsymbol{\mu}_l - \boldsymbol{\mu}_j) \\ & = (\boldsymbol{\Gamma}'_{j|l} \boldsymbol{\xi})' (\boldsymbol{\Lambda}_{j|l} - \mathbf{I}) (\boldsymbol{\Gamma}'_{j|l} \boldsymbol{\xi}) \\ & \quad + 2(\boldsymbol{\Gamma}'_{j|l} \boldsymbol{\xi})' (\boldsymbol{\Lambda}_{j|l} \boldsymbol{\Gamma}_{j|l} \boldsymbol{\Sigma}_l^{-\frac{1}{2}}) (\boldsymbol{\mu}_l - \boldsymbol{\mu}_j) \\ & \quad + (\boldsymbol{\mu}_l - \boldsymbol{\mu}_j)' (\boldsymbol{\Sigma}_l^{-\frac{1}{2}} \boldsymbol{\Gamma}_{j|l} \boldsymbol{\Lambda}_{j|l} \boldsymbol{\Gamma}'_{j|l} \boldsymbol{\Sigma}_l^{-\frac{1}{2}}) (\boldsymbol{\mu}_l - \boldsymbol{\mu}_j) \\ & \stackrel{d}{=} \sum_{i=1}^p (\lambda_i - 1) Z_i^2 + 2\lambda_i \delta_i Z_i + \lambda_i \delta_i^2, \end{aligned} \quad (8)$$

where $\delta_i, i = 1, \dots, p$ are as in the statement of the theorem. Depending on the values of λ_i , one can simplify the expression in (8). If $\lambda_i > 1$: $(\lambda_i - 1)Z_i^2 + 2\lambda_i \delta_i Z_i + \lambda_i \delta_i^2 = (\sqrt{\lambda_i - 1} Z_i + \lambda_i \delta_i / \sqrt{\lambda_i - 1})^2 - \lambda_i \delta_i^2 / (\lambda_i - 1)$, which is distributed as a $(\lambda_i - 1) \chi_{1, \lambda_i^2 \delta_i^2 / (\lambda_i - 1)}^2$ -random variable. If $\lambda_i < 1$: $(\lambda_i - 1)Z_i^2 + 2\lambda_i \delta_i Z_i + \lambda_i \delta_i^2 = -(\sqrt{1 - \lambda_i} Z_i + \lambda_i \delta_i / \sqrt{1 - \lambda_i})^2 - \lambda_i \delta_i^2 / (\lambda_i - 1)$, which is distributed as a $(1 - \lambda_i) \chi_{1, \lambda_i^2 \delta_i^2 / (\lambda_i - 1)}^2$ -random variable. In the case of $\lambda_i = 1$, $(\lambda_i - 1)Z_i^2 + 2\lambda_i \delta_i Z_i + \lambda_i \delta_i^2 = 2\lambda_i \delta_i Z_i + \lambda_i \delta_i^2$. The proof follows from incorporating these expressions and rearranging terms in (8). \square

Remark 2: Computing $\Pr[Y^{j,l}(\mathbf{X}) < 0]$ in case (a) and (b) of the theorem involves calculating Gaussian probabilities. For the third case (part (c) of the theorem) we see that this involves calculating the probability of a linear combination of independent non-central χ^2 and Gaussian variable, for which we use Algorithm AS 155 of [36]. Once $\Pr[Y^{j,l}(\mathbf{X}) < 0]$ and $\Pr[Y^{l,j}(\mathbf{X}) < 0]$ are computed, the index can be calculated from (5) using $p^{j,l}$ and $p^{l,j}$ computed from (4).

2.1.1 Properties of our Preliminary Index

In this section, we highlight some properties of our preliminary index with regard to achievable values and scaling that will be used for our second objective of simulating random configurations satisfying a desired value of our index.

Theorem 3: Fix $c > 0$, $\alpha \in (0, 1)$ and consider two clusters \mathcal{C}_l and \mathcal{C}_j . Each \mathcal{C}_i has n_i p -variate Gaussian observations with mean $c\boldsymbol{\mu}_i$ and covariance $\boldsymbol{\Sigma}_i$, $i \in \{j, l\}$. Then for specific choices of the distance metric, the following properties hold:

- (a) For the Euclidean distance ($\boldsymbol{\Delta}_j = \boldsymbol{\Delta}_l = \mathbf{I}$), define $\theta_i \equiv \Phi(\sqrt{n_i}(1 - 2\alpha))$, $i \in \{j, l\}$ and $\mathcal{I}_0^{j,l} \equiv 1 - \left(\frac{\theta_j + \theta_l}{2}\right)^{1/n_{j,l}}$. Then for large n_j, n_l ,

$$\lim_{c \rightarrow 0} \mathcal{I}^{j,l} \approx \mathcal{I}_0^{j,l} \quad \text{and} \quad \lim_{c \rightarrow \infty} \mathcal{I}^{j,l} = 1. \quad (9)$$

- (b) For the special case of Mahalanobis' distance with identical covariance structure (i.e., $\boldsymbol{\Delta}_j \equiv \boldsymbol{\Delta}_l = \boldsymbol{\Sigma}^{-1}$), define $\theta_j, \theta_l, \mathcal{I}_0^{j,l}$ as in part (a). Then for large n_j, n_l ,

$$\lim_{c \rightarrow 0} \mathcal{I}^{j,l} \approx \mathcal{I}_0^{j,l} \quad \text{and} \quad \lim_{c \rightarrow \infty} \mathcal{I}^{j,l} = 1. \quad (10)$$

- (c) For the general Mahalanobis distance (i.e., $\boldsymbol{\Delta}_j = \boldsymbol{\Sigma}_j^{-1}$, $\boldsymbol{\Delta}_l = \boldsymbol{\Sigma}_l^{-1}$ with $\boldsymbol{\Sigma}_l \neq \boldsymbol{\Sigma}_j$), let $\lambda_1, \lambda_2, \dots, \lambda_p$ be the eigenvalues of $\boldsymbol{\Sigma}_{j|l} \equiv \boldsymbol{\Sigma}_l^{\frac{1}{2}} \boldsymbol{\Sigma}_j^{-1} \boldsymbol{\Sigma}_l^{\frac{1}{2}}$ and Z_1, \dots, Z_l be independent $N(0, 1)$ random variables. Define for $i \in \{j, l\}$, $\theta_i \equiv \Phi\left(\frac{\sqrt{n_i}(\kappa - \alpha)}{\sqrt{\kappa(1 - \kappa)}}\right)$, where $\kappa = \Pr\left[\sum_{i=1}^p (\lambda_i - 1)Z_i^2 < 0\right]$. Also define $\mathcal{I}_0^{j,l} \equiv 1 - \left(\frac{\theta_j + \theta_l}{2}\right)^{1/n_{j,l}}$. Then for large n_j, n_l . Then

$$\lim_{c \rightarrow 0} \mathcal{I}^{j,l} \approx \mathcal{I}_0^{j,l} \quad \text{and} \quad \lim_{c \rightarrow \infty} \mathcal{I}^{j,l} = 1. \quad (11)$$

Proof: First note that for large n_l and the normal approximation to binomial probabilities,

$$\begin{aligned} p^{j,l} &= \sum_{i=n_l\alpha}^{n_l} \binom{n_l}{i} \Pr[Y^{j,l}(\mathbf{X}) < 0]^i \Pr[Y^{j,l}(\mathbf{X}) > 0]^{n_l-i} \\ &\approx \Phi\left(\frac{\sqrt{n_l}(\Pr[Y^{j,l}(\mathbf{X}) < 0] - \alpha)}{\sqrt{\Pr[Y^{j,l}(\mathbf{X}) < 0] \Pr[Y^{j,l}(\mathbf{X}) > 0]}}\right). \end{aligned} \quad (12)$$

We get from Theorem 1(a), $Y^{j,l}(\mathbf{X}) \sim N(c^2(\boldsymbol{\mu}_l - \boldsymbol{\mu}_j)'(\boldsymbol{\mu}_l - \boldsymbol{\mu}_j), 4c^2(\boldsymbol{\mu}_l - \boldsymbol{\mu}_j)' \boldsymbol{\Sigma}_l (\boldsymbol{\mu}_l - \boldsymbol{\mu}_j))$ and hence,

$$\Pr[Y^{j,l}(\mathbf{X}) < 0] = \Phi\left(\frac{-c(\boldsymbol{\mu}_l - \boldsymbol{\mu}_j)'(\boldsymbol{\mu}_l - \boldsymbol{\mu}_j)}{2\sqrt{((\boldsymbol{\mu}_l - \boldsymbol{\mu}_j)' \boldsymbol{\Sigma}_l (\boldsymbol{\mu}_l - \boldsymbol{\mu}_j))}}\right).$$

Taking limits of both sides, we get using continuity of $\Phi(\cdot)$,

$$\lim_{c \rightarrow 0} \Pr[Y^{j,l}(\mathbf{X}) < 0] = 0.5, \quad \text{and} \quad \lim_{c \rightarrow \infty} \Pr[Y^{j,l}(\mathbf{X}) < 0] = 0.$$

This, together with (12) and the definition of $\mathcal{I}^{j,l}$ in (5), completes the proof of part (a).

The proof for Part (b) is very similar to that of part (a) and is omitted. For part (c), note that from Theorem 1(c),

$$Y^{j,l}(\mathbf{X}) \stackrel{d}{=} \sum_{i=1}^p (\lambda_i - 1)Z_i^2 + 2\lambda_i\delta_i c Z_i + \lambda_i\delta_i^2 c^2, \quad (13)$$

where δ_i , $i = 1, \dots, p$ are as defined there (and using the fact that the means are $c\boldsymbol{\mu}_j$ and $c\boldsymbol{\mu}_l$ here). Then by continuity arguments, it follows that

$$\lim_{c \rightarrow 0} \Pr[Y^{j,l}(\mathbf{X}) < 0] = \Pr\left[\sum_{l=1}^p (\lambda_l - 1)Z_l^2 < 0\right] = \kappa.$$

Note that the right-side of (13) is a quadratic in c with leading coefficient $\sum_{i=1}^p \lambda_i \delta_i^2 > 0$. Hence, for large values of c (in particular, when c is larger than the largest root of the quadratic equation), the right side in (13) is positive. Taking limits on both sides of (13) and using continuity arguments we get $\lim_{c \rightarrow \infty} \Pr[Y^{j,l}(\mathbf{X}) < 0] = 0$. This completes the proof of Theorem 3. \square

The following corollary is immediate from the proof above.

Corollary 4: Fix $\alpha \in (0, 1)$. Let for any $c > 0$, $\mathcal{I}^{j,l}(c) = \mathcal{I}^{j,l}$ be our preliminary index of separation between two groups \mathcal{C}_l and \mathcal{C}_j , each having n_i number of p -variate Gaussian observations with mean $c\boldsymbol{\mu}_i$ and covariance $\boldsymbol{\Sigma}_i$, $i \in \{j, l\}$. For different choices of the distance metric, let $\mathcal{I}_0^{j,l}$ be as defined in the three parts of Theorem 3. Then $\mathcal{I}^{j,l}(c)$ is a continuous function of c with its range containing $(\mathcal{I}_0^{j,l}, 1)$.

From Theorem 1 (see Remark 2 and Corollary 4), one can compute the value of the preliminary index for given values of the cluster means and dispersions. In real datasets, we can apply some clustering method to obtain the cluster memberships and estimate the cluster means and the variance structures and compute an estimated version of the index. See Section 4 for one such application.

2.2 Generating Data with given values of the Index

From Corollary 4, it is clear that any value in $(\mathcal{I}_0^{j,l}, 1)$ is a possible value of $\mathcal{I}^{j,l}$, for a given set of model parameters $(c\boldsymbol{\mu}, \boldsymbol{\Sigma}, n_1, n_2)$ by suitably choosing the scaling parameter $c > 0$. This leads to the data-generation algorithm described in the appendix. The main idea is that for a given target value of the index, start with some initial sets of parameters $p, (\boldsymbol{\mu}, \boldsymbol{\Sigma}, n_1, n_2)$, and compute our index for this initial configuration. Then find $c > 0$ iteratively so that the data with parameters $(c\boldsymbol{\mu}, \boldsymbol{\Sigma}, n_1, n_2)$ attains the target value. The algorithm described in the appendix gives a more general version of the index discussed in Section 3.

3 ILLUSTRATIONS AND ADJUSTMENTS

In this section, we provide some illustrative examples obtained by simulating realizations from groups under different scenarios using the preliminary index and the algorithm in the appendix. We study the relationship of realizations obtained at these different values of the index with that of clustering difficulty, which we measure in terms of the Adjusted Rand index (\mathcal{R}) of [37] obtained by clustering the observations in each dataset and comparing the result with the true classification. Note that by design, \mathcal{R} takes an average value of zero if all observations are assigned to a group completely at random. A perfect grouping of the observations matching the true, on the other hand, yields \mathcal{R} its highest value of unity.

3.1 Homogeneous Spherical Clusters

Here, it is assumed that all groups have the same spherical dispersion structure, i.e., $\boldsymbol{\Sigma}_k = \sigma^2 \mathbf{I}$ for all $k \in \{1, 2, \dots, K\}$. In this section, the idealized \mathcal{R} is calculated on each simulated dataset by comparing the true classification with that obtained

using the k -means algorithm of [35] started with the true (known) group means (in order to eliminate initialization issues on the obtained clustering). The obtained clustering may thus be regarded as the best possible grouping and its degree of ability (as measured by \mathcal{R}) to recover the true classes may be considered to be an indication of the clustering complexity of the simulated dataset.

3.1.1 The Two-Groups Case

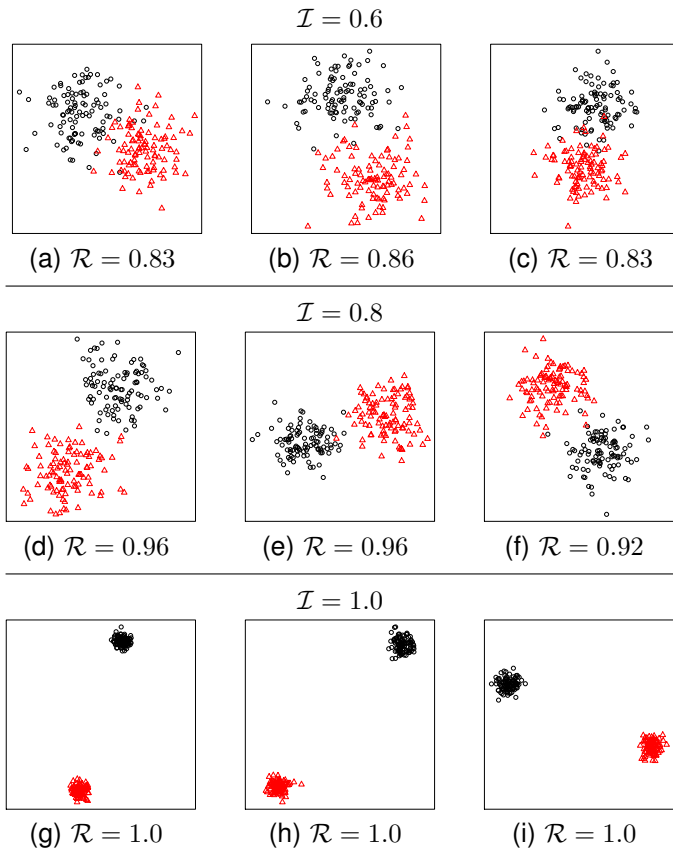


Fig. 1. Simulated datasets at three different values of \mathcal{I} . Colors and symbols represent the true cluster class. Above each set of plots we give the value of \mathcal{I} and below each individual plot we give the obtained \mathcal{R} .

Figures 1a-i display simulated datasets for different values of our preliminary index $\mathcal{I} \equiv \mathcal{I}^{1,2}$ with $\alpha = 0.75$. See [38] for plots of additional realizations and at other values of \mathcal{I} . In each case, 100 observations were generated from each of two groups separated according to \mathcal{I} . In the figures, color and character are used to distinguish the true grouping. Figure 1 demonstrates that as the value of \mathcal{I} increases, the clusters become more separated. This is confirmed by the values of \mathcal{R} . The classifications of the datasets in Figures 1a-c have the lowest \mathcal{R} between 0.83 and 0.86. Each subsequent row down has a range of \mathcal{R} higher than those in the previous row. Thus, there is some qualitative support for our measure of separation (\mathcal{I}) as an indicator of difficulty, however a more comprehensive analysis is needed for using \mathcal{I} as a quantitative measure. Therefore, we conducted a simulation study to investigate the validity of our index as a quantitative

surrogate for clustering complexity. Specifically, we simulated 25 datasets, each having observations from two groups at each of 15 evenly-spaced values of \mathcal{I} (using $\alpha = 0.75$) in $(0, 1)$, for nine different combinations of numbers of observations per cluster and dimensions. For each dataset obtained at each value of \mathcal{I} , we calculated the \mathcal{R} as outlined above.

Figures 2a-c display the results for $p = 10, 100$ and 1000 , respectively, with \mathcal{I} on the x -axis and the corresponding \mathcal{R} on the y -axis. Color is used to indicate the number of observations per group in each dataset at setting. In each figure, the shaded region for each color denotes the spread between the first and third quartiles of \mathcal{R} based on 25 datasets. From the figures, we note that \mathcal{I} tracks \mathcal{R} very well, again providing qualitative support for our index as a surrogate for clustering complexity. However, the relationship is not linear. There is also more variability in \mathcal{R} when the number of observations is low. Further, the bands representing the spread between the first and third quartiles of \mathcal{R} do not often overlap with a change in dimension. Thus, there is some inconsistency in the relationship between our preliminary index \mathcal{I} and \mathcal{R} across dimensions. Indeed, this inconsistency is most apparent when the number of observations in the dataset is less than the number of dimensions. This inconsistency is not terribly surprising, and in line with the so-called ‘‘curse of dimensionality’’ and the need for larger sample sizes with increasing p [39] to obtain the same kinds of clustering performance as with lower dimensions. In order to maintain interpretability across dimensions, we therefore investigate adjustments to our preliminary index to account for the effect of n and p . We pursue this course in the remainder of this section.

3.1.1.1 An Initial Adjustment to \mathcal{I} for group size and dimension: To understand further the relationship between \mathcal{I} and \mathcal{R} we simulated 25 datasets each with observations from two homogeneous spherical groups for all combinations of (n_1, n_2, p) , where $n_1 \leq n_2$ (assumed without loss of generality) were the observations in the first and second groups, and $p \in \{2, 4, 5, 10, 20, 50, 100, 200, 500, 1000\}$, and $(n_1, n_2) \in \{(20, 20), (50, 50), (75, 75), (100, 100), (200, 200), (500, 500), (1000, 1000), (30, 100), (20, 50), (60, 75), (90, 100), (150, 250), (50, 600), (100, 1000)\}$. We simulated datasets according to ten values of \mathcal{I} evenly spaced between 0 and 1, and for $\alpha = 0.75$. For each of these 105000 datasets thus obtained, we again obtained \mathcal{R} using k -means. We explored several relationships between \mathcal{I} and \mathcal{R} very extensively. Of these explorations, the following multiplicative relationship between \mathcal{I} and \mathcal{R} was found to perform the best:

$$\log \left(\frac{\mathcal{R} + 1}{1 - \mathcal{R}} \right) \approx \exp(\delta_\alpha) \left(\log \left(\frac{1}{1 - \mathcal{I}} \right) \right)^{\theta_\alpha}, \quad (14)$$

where $\delta_\alpha = \sum_{1 \leq i \leq j \leq k \leq 4} \zeta_{\omega_i, \omega_j, \omega_k} \log(\omega_i) \log(\omega_j) \log(\omega_k)$, $\theta_\alpha = \sum_{1 \leq i \leq j \leq 4} \beta_{\omega_i, \omega_j} \log(\omega_i) \log(\omega_j)$, and $(\omega_1, \omega_2, \omega_3, \omega_4) = (n_1, n_2, p, e)$. Then for $\alpha = 0.75$ we fit the linear model

$$\log \left[\log \left(\frac{\mathcal{R} + 1}{1 - \mathcal{R}} \right) \right] \approx \theta_\alpha \log \left[\log \left(\frac{1}{1 - \mathcal{I}} \right) \right] + \delta_\alpha. \quad (15)$$

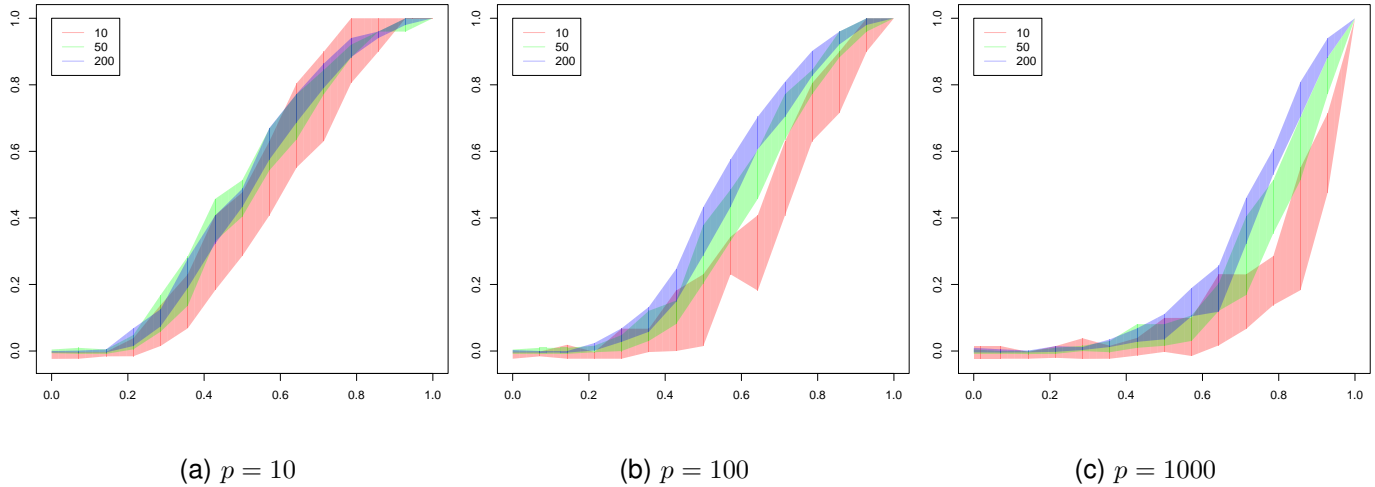


Fig. 2. Plots for $K = 2$ clusters comparing \mathcal{R} (y -axis) against \mathcal{I} (x -axis) for $\alpha = 0.75$. The three colors designate number of observations per cluster such that $n_1 = n_2$. The lower and upper bound of the bands represent the first and third quartile of \mathcal{R} calculated on 25 datasets for several different \mathcal{I} .

The estimates for the coefficients in (15) were obtained using least-squares to obtain the best possible fit of \mathcal{I} to \mathcal{R} and are displayed in Table 1 under the columns labeled *hom*. See [38] for parameter estimates obtained at $\alpha = (0.25, 0.50)$.

This relationship has the property that a large value of \mathcal{I} corresponds to \mathcal{R} that is close to 1 while values of \mathcal{I} that are close to zero correspond to \mathcal{R} s that are also close to zero. The sum of the estimates for the four three-way interaction terms for the n s is close to zero for each α . This suggests that when $n_1 = n_2$ the cubed term for the n s is not that different from zero. Therefore, based on the parameter estimates in the left panel of Table 1, we define $\mathcal{R}_{\mathcal{I},\alpha}$ as follows:

$$\mathcal{R}_{\mathcal{I},\alpha} = \frac{\exp\left(\exp(\delta_\alpha) \left(\log\left(\frac{1}{1-\mathcal{I}}\right)\right)^{\theta_\alpha}\right) - 1}{\exp\left(\exp(\delta_\alpha) \left(\log\left(\frac{1}{1-\mathcal{I}}\right)\right)^{\theta_\alpha}\right) + 1}. \quad (16)$$

We call $\mathcal{R}_{\mathcal{I},\alpha}$ as our initial adjusted index. We now investigate its performance in indexing clustering complexity in a similar framework to Figures 1 and 2. Figure 3 illustrates two-cluster datasets simulated in a similar manner as in Figure 1 but using our initial adjusted index $\mathcal{R}_{\mathcal{I},0.75}$. Similar is the case with Figure 4 which mimics the setup of Figure 2 with the only difference being that datasets are generated here using $\mathcal{R}_{\mathcal{I},0.75}$ (instead of \mathcal{I}). In both cases, we note that the agreement between the numerical values of \mathcal{R} and $\mathcal{R}_{\mathcal{I},0.75}$ is substantially improved. In particular, Figures 3a-i show that the range of actual values of \mathcal{R} contains the value of $\mathcal{R}_{\mathcal{I},0.75}$. Also, Figures 4a-c demonstrate that $\mathcal{R}_{\mathcal{I},0.75}$ tracks \mathcal{R} very well. Similar is the case with $\alpha = 0.25$ and 0.50 (see [38]), providing support for the use of $\mathcal{R}_{\mathcal{I},\alpha}$ as a surrogate of clustering complexity. This support is consistent for different numbers of observations and dimension. Note however, that Figure 4 continues to indicate greater variability in the obtained \mathcal{R} when there are fewer observations in a group. This is expected because smaller sample sizes lead to results with higher variability.

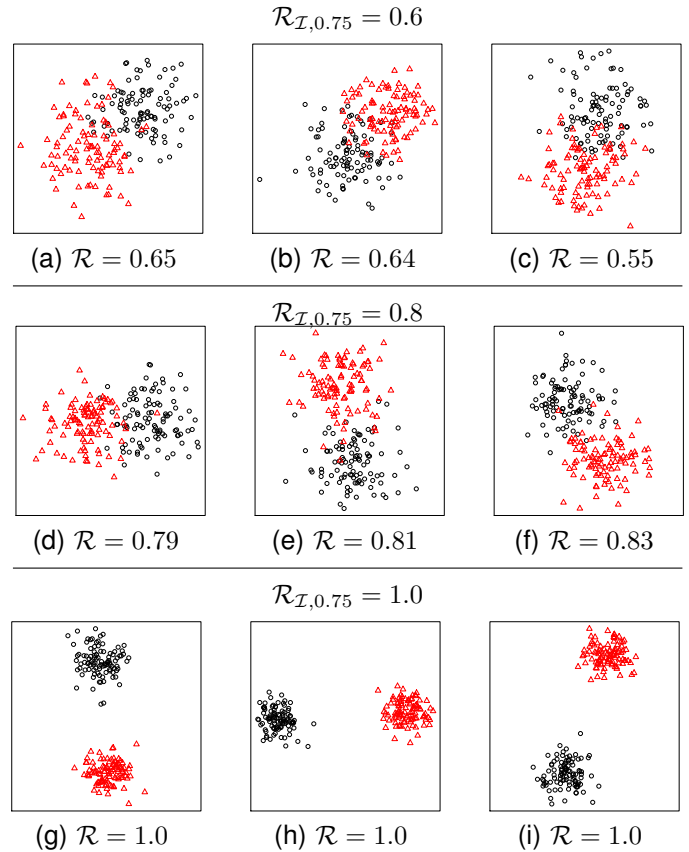


Fig. 3. Simulated datasets at different $\mathcal{R}_{\mathcal{I},0.75}$ (top of each row). Color and symbol represent true class. Obtained \mathcal{R} values are provided below each plot.

3.1.2 The Case with Many Groups ($K \geq 2$)

So far we have analyzed and developed our index for two-class datasets. In this section, we extend it to the general multi-group case. Clearly, separation between different pairs of clusters impacts clustering difficulty. We investigated several

TABLE 1

Table of estimated parameter values to adjust index for n and p when $\alpha = 0.75$, for clusters with homogeneous spherical (*hom*) and the general heterogeneous (*het*) dispersion structures. Fonts in the table represent bounds on the p -values (bold and underline for a p -value < 0.001 , bold and italic for a p -value < 0.01 , bold for a p -value < 0.05 , italic for a p -value < 0.1 and regular font otherwise).

	<i>hom</i>	<i>het</i>		<i>hom</i>	<i>het</i>		<i>hom</i>	<i>het</i>		<i>hom</i>	<i>het</i>		<i>hom</i>	<i>het</i>
ζ	<u>1.191</u>	-0.18	$\zeta_{n_1,p}$	0.02	<u>-0.19</u>	ζ_{n_1,n_1,n_1}	<u>0.202</u>	<u>-0.51</u>	β	<u>0.543</u>	<u>0.578</u>	$\beta_{n_1,p}$	<u>0.031</u>	<u>0.078</u>
ζ_{n_1}	<u>1.705</u>	0.634	$\zeta_{n_2,p}$	<u>0.047</u>	<u>0.198</u>	ζ_{n_1,n_1,n_2}	<u>-0.53</u>	<u>0.721</u>	β_{n_1}	<u>-0.085</u>	<u>-0.303</u>	$\beta_{n_2,p}$	<u>-0.009</u>	<u>-0.039</u>
ζ_{n_2}	<u>-2.231</u>	-0.344	ζ_{n_1,n_1}	0.276	<u>2.684</u>	ζ_{n_1,n_2,n_2}	<u>0.54</u>	0.003	β_{n_2}	<u>0.567</u>	<u>0.596</u>	β_{n_1,n_1}	<u>0.065</u>	<u>0.406</u>
ζ_p	<u>-0.188</u>	<u>-0.12</u>	ζ_{n_2,n_2}	<u>1.359</u>	<u>2.766</u>	ζ_{n_2,n_2,n_2}	<u>-0.222</u>	<u>-0.21</u>	β_p	<u>-0.024</u>	<u>-0.029</u>	β_{n_1,n_2}	<u>-0.145</u>	<u>-0.543</u>
$\zeta_{p,p}$	<u>-0.046</u>	<u>-0.061</u>	ζ_{n_1,n_2}	<u>-1.512</u>	<u>-5.51</u>							β_{n_2,n_2}	<u>0.033</u>	<u>0.121</u>

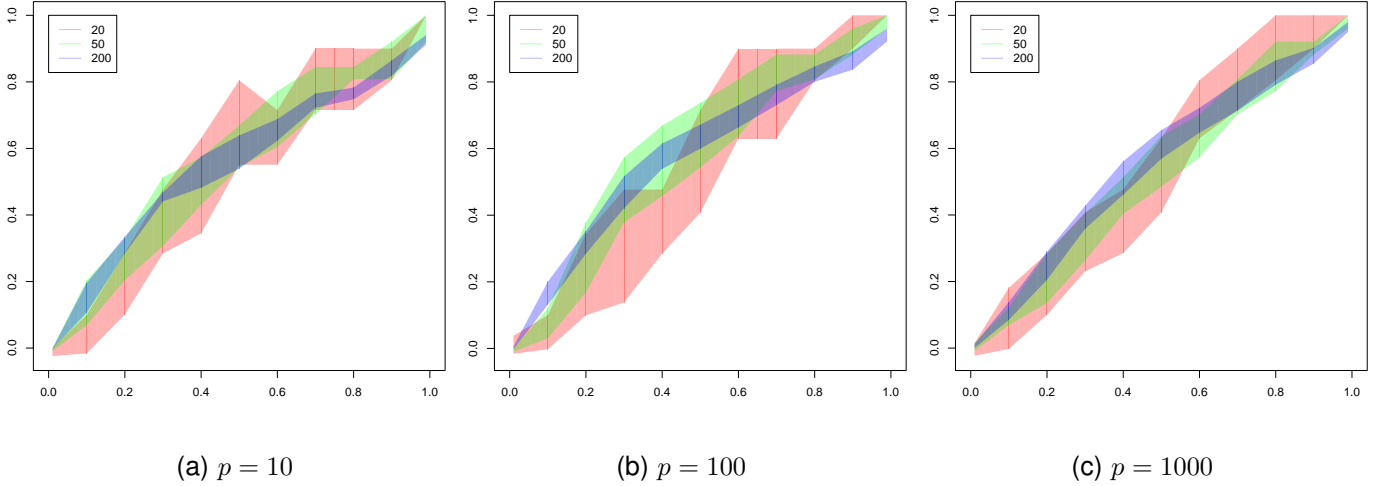


Fig. 4. Plots for $K = 2$ clusters comparing \mathcal{R} against $\mathcal{R}_{\mathcal{I},0.75}$. Colors and bands are as in Figure 2.

possibilities for summarizing clustering difficulty based on all $\binom{K}{2}$ pairwise indices, however, they all possessed several drawbacks. For instance, the average pairwise separation is typically high because many of the $\mathcal{R}_{\mathcal{I},\alpha}$'s are close to 1, while the minimum is overly influenced by the presence of (only) two close groups. Therefore, we investigated an adaptation of the summarized multiple Jaccard similarity index proposed by [40]. The Jaccard coefficient of similarity [41] measures similarity or overlap between two species or populations. This was extended by [40] for summarizing many pairwise indices for the case of multiple populations. Note that both the Jaccard index and its summarized multiple version address similarity or overlap between populations while our pairwise index measures separability. This needs to be adapted for our case. Therefore, we define $\mathcal{R}_{\mathcal{I},\alpha}^{ii} = 0$ for $i = 1, \dots, K$. Further, for each pair $\mathcal{R}_{\mathcal{I},\alpha}^{ij}$: $1 \leq i, j \leq K$ of clusters, let $\mathcal{R}_{\mathcal{I},\alpha}^{ij} = \mathcal{R}_{\mathcal{I},\alpha}^{ji} \equiv \mathcal{R}_{\mathcal{I},\alpha}$, i.e., the adjusted index defined using clusters \mathcal{C}_i and \mathcal{C}_j . Also, let $\Upsilon = ((\mathcal{R}_{\mathcal{I},\alpha}^{ij}))_{1 \leq i, j \leq K}$ be the matrix of corresponding $\mathcal{R}_{\mathcal{I},\alpha}^{ij}$'s. Then we define the following summarized index for K clusters:

$$\mathcal{R}_{\mathcal{I},\alpha} = 1 - \frac{(\mathbf{J}_K - \Upsilon)_{(1)}^\lambda - 1}{K - 1}, \quad (17)$$

where \mathbf{J}_K is a $K \times K$ matrix with all entries equal to unity, and $(\mathbf{J}_K - \Upsilon)_{(1)}^\lambda$ is the largest eigenvalue of the matrix $(\mathbf{J}_K - \Upsilon)$. [40] motivates his summary using principal components analysis (PCA) in the context of a correlation

matrix where the first principal component is that orthogonal projection of the dataset that captures the greatest variability in the K coordinates. Like his summary, our summary index (17) has some very appealing properties. When the matrix of pairwise separation indices is $\mathbf{J}_K - \mathbf{I}_K$, i.e., $\mathcal{R}_{\mathcal{I},\alpha}^{jl} = 1 \forall j \neq l$, then the first (largest) eigenvalue captures only $1/K$ proportion of the total eigenvalues. In this case, $\mathcal{R}_{\mathcal{I},\alpha} = 1$. On the other hand when every element in the matrix is zero, there is perfect overlap between all groups, and only the largest eigenvalue is non-zero. In this case $\mathcal{R}_{\mathcal{I},\alpha} = 0$. Finally, when $K = 2$, $\mathcal{R}_{\mathcal{I},\alpha}$ is consistent with the (sole) pairwise index $\mathcal{R}_{\mathcal{I},\alpha}^{ij}$.

3.1.2.1 Final Adjustments to Separation Index:

While (17) provides a summarized pairwise measure that is appealing in some special cases, we still need to investigate its performance as a surrogate measure of clustering complexity. We therefore performed a detailed simulation study to study the relationship between the summarized $\mathcal{R}_{\mathcal{I},\alpha}$ and \mathcal{R} . Specifically, we generated 25 K -cluster datasets for $K = 3, 5$ and 10 each with 100 dimensions, at each of 10 different values of $\mathcal{R}_{\mathcal{I},0.75}$ between 0 and 1, for equal numbers of observations in each group, (i.e., $n_k \equiv n_0$) where $n_0 \in \{20, 100, 1000\}$. For each dataset, we used the k -means algorithm, and computed \mathcal{R} of the subsequent clustering. Figures 5a-c plot the interquartile ranges of \mathcal{R} for each combination of K and n_0 . Here we have $\mathcal{R}_{\mathcal{I},\alpha}$ on the x -axis and \mathcal{R} on the y -axis. Figures 5a-c demonstrate that $\mathcal{R}_{\mathcal{I},\alpha}$ tracks \mathcal{R} . In addition, the relationship between $\mathcal{R}_{\mathcal{I},\alpha}$ and \mathcal{R} is consistent for different

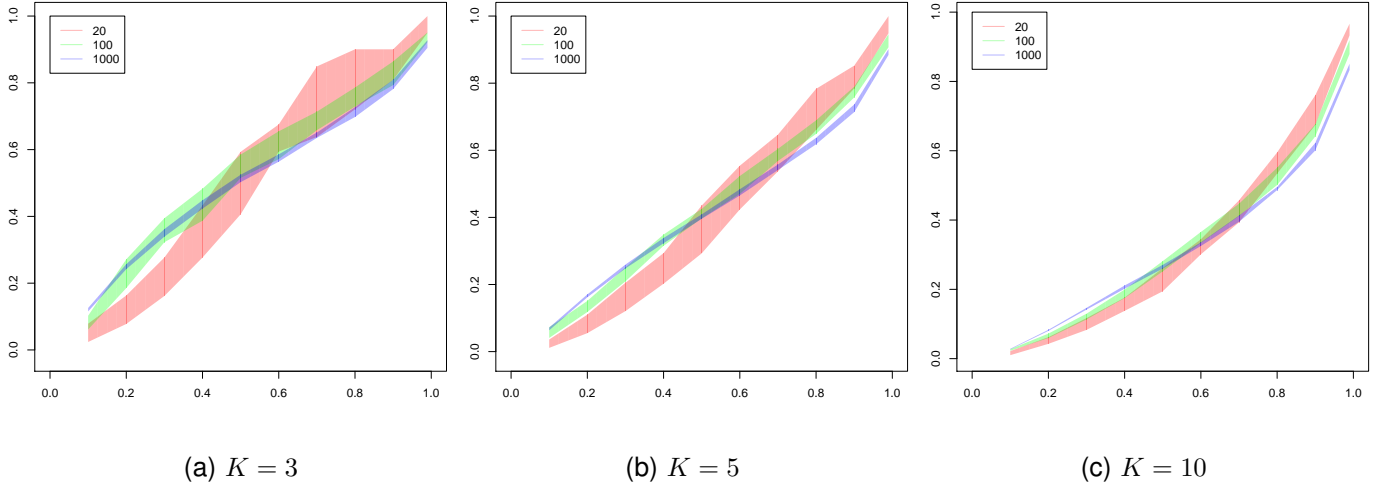


Fig. 5. Plots for \mathcal{R} against $\mathcal{R}_{\mathcal{I},0.75}$. The three colors designate numbers of observations per cluster, set to be equal and $\in \{20, 100, 1000\}$. Other aspects of the plots are as in Figure 2.

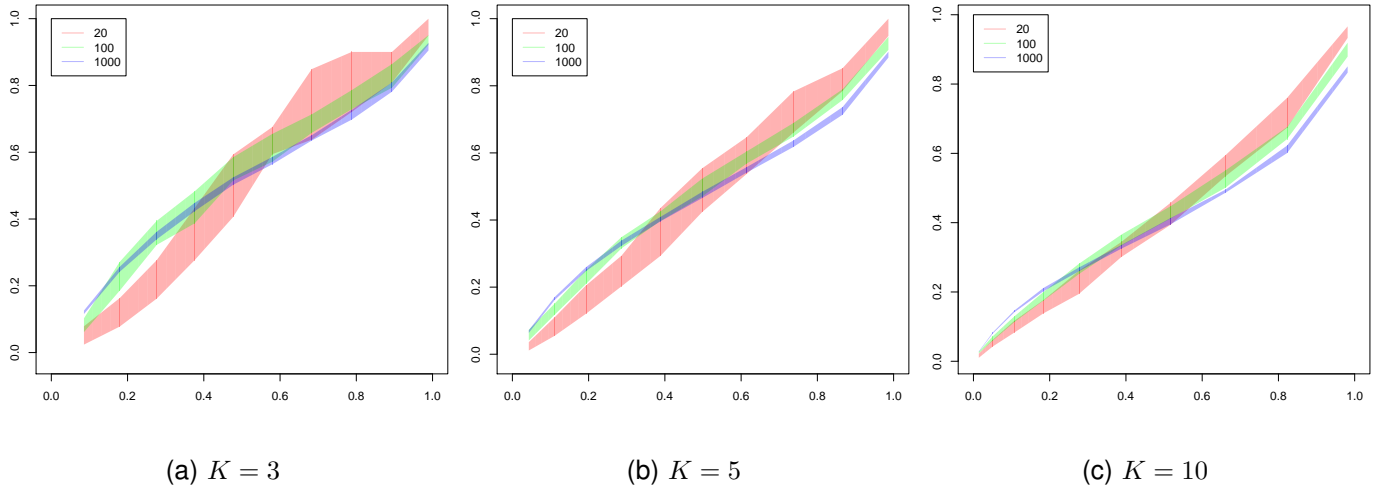


Fig. 6. Plots of \mathcal{R} against \mathcal{I}_* for when $\alpha = 0.75$. Other aspects of the plot are as in Figure 5.

numbers of observations. However, it is also clear that the exact relationship between $\mathcal{R}_{\mathcal{I},\alpha}$ and \mathcal{R} depends on K so some more adjustments are called for. To study this issue further, we simulated 25 datasets for each combination of $p \in \{2, 4, 5, 10, 20, 50, 100, 200, 500, 1000\}$, $n_i = n_j \equiv n_0$ for all i, j , where $n_0 \in \{20, 50, 75, 100, 200, 500, 1000\}$, at ten evenly-spaced values of $\mathcal{R}_{\mathcal{I},\alpha}$ in $(0,1)$, $\alpha \in \{0.25, 0.5, 0.75\}$ and $K \in \{3, 5, 7, 10\}$. For each dataset we calculated \mathcal{R} . Using the \mathcal{R} from each of these datasets, and for each combination of (p, α, K) we fit the multiplicative model:

$$\mathcal{R} \approx \mathcal{R}_{\mathcal{I},\alpha}^{\beta_{k,p}}. \quad (18)$$

Using the parameter estimates for $\beta_{k,p}$ for each combination of dimension and number of clusters, we then fit the following linear model separately for each tuning parameter:

$$\beta_{k,p} = \eta + \eta_1 k + \eta_2 k^2 + \eta_3 p + \eta_4 p^2 + \eta_5 kp. \quad (19)$$

Parameter estimates for this model for the case of $\alpha = 0.75$

TABLE 2

Table of estimated parameter values to adjust index for K and p when $\alpha = 0.75$, for clusters with homogeneous spherical (*hom*) and the general heterogeneous (*het*) dispersion structures. For the estimated parameters, two of the p -values are < 0.01 and the rest are < 0.001 . For any two constants a and b , $aE-b$ means $a * 10^{-b}$.

	η	η_1	η_2	η_3	η_4	η_5
<i>hom</i>	0.51	0.21	-0.01	2.18E-4	-1.9E-7	2.28E-5
<i>het</i>	0.65	0.013	-0.0051	0.002	-1.42E-5	3.98E-4

are presented in Table 2 (see [38] for the parameter estimates when $\alpha \in \{0.25, 0.5\}$). Thus the final version of our index, after all adjustments, for the case of homogeneous spherical clusters is

$$\mathcal{I}_* = \mathcal{R}_{\mathcal{I},\alpha}^{\beta_{k,p}}. \quad (20)$$

Figures 6a-c are constructed similarly to Figures 5a-c except that the datasets are now generated using the algorithm in the appendix and using \mathcal{I}_* (instead of $\mathcal{R}_{\mathcal{I},0.75}$). Note that the effect of K has been largely addressed and the relationship between \mathcal{I}_* and \mathcal{R} is fairly similar across dimensions.

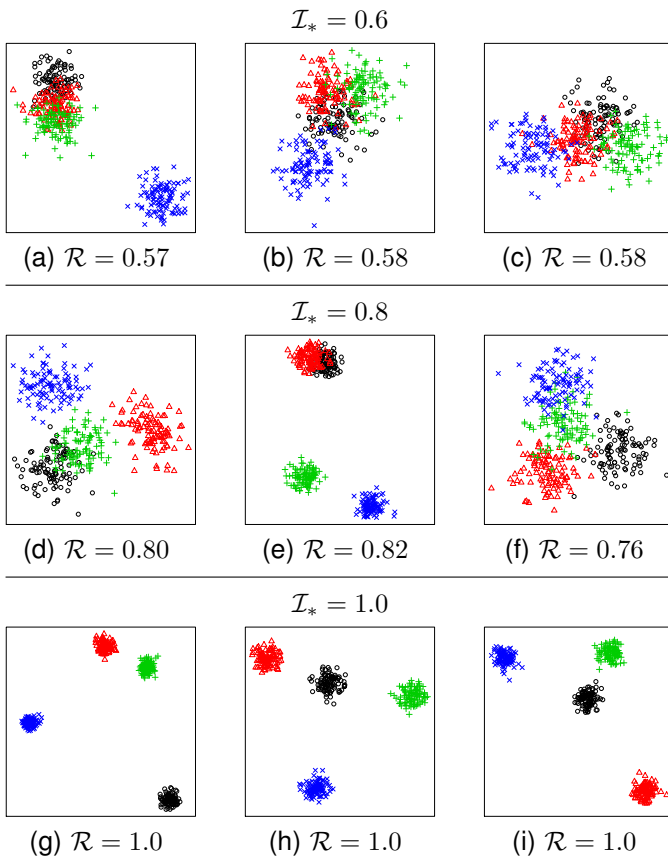


Fig. 7. Four-component simulated datasets at different \mathcal{I}_* -values, with color and plotting character used to represent the true class. Obtained \mathcal{R} -values are displayed below each plot.

3.1.2.2 Illustrations: We now present, in Figures 7a-i, some simulated four-class datasets to demonstrate possible configurations obtained using the algorithm in the appendix for three different values of \mathcal{I}_* . Additional realizations and at different values of \mathcal{I}_* are in [38]. In each case we used $\alpha = 0.75$. The different colors and characters in the figures represent the true classes. Note that the clusters are well-separated as \mathcal{I}_* increases. Clustering complexity also decreases as confirmed by the computed values of \mathcal{R} . The datasets in the first row have the lowest \mathcal{I}_* , with each subsequent row down having a range of \mathcal{R} higher than those in the previous row. In each row we see that \mathcal{I}_* is within the range of the actual \mathcal{R} index values. This provides further evidence that the transformed version of our index in the form of \mathcal{I}_* is similar to \mathcal{R} .

We conclude this section with a few comments. Note that \mathcal{I}_* is strictly between 0 and 1. Further, let us denote $\mathcal{I}_*(c)$ to be the value of \mathcal{I}_* that is defined using (15), (18) and (20), but with $\mathcal{I}^{i,j}$ replaced by $\mathcal{I}(c)^{i,j}$ as in Corollary 4 for the cluster \mathcal{C}_i with n_i p -variate Gaussian observations with mean $c\mu_i$ and covariance Σ_i , $i \in \{1, \dots, K\}$. The following corollary

implies that we can generate datasets with any value of the final index between $\mathcal{I}_*(0)$ and unity using the algorithm in the appendix.

Corollary 5: Fix α . Let $c > 0$, then for positive θ_α and $\beta_{k,p}$, $\mathcal{I}_*(c)$ is a continuous function where its range contains $(\mathcal{I}_*(0), 1)$. In addition, $\mathcal{I}_*(c)$ is an increasing function of c .

Proof: The result follows directly from Theorem 3 and Corollary 4. \square

Note that we found the adjustments in (15) and (20) empirically through simulations. In all of the cases we considered, θ_α and $\beta_{k,p}$ are positive and thus Corollary 5 holds. Ideally, we should consider conducting further simulations if we desire adjustments for n_i, n_j, K or p that are very different from the cases we considered in our simulations. In this case we need to find other parameter estimates as in Tables 1 and 2.

3.2 Clusters with General Ellipsoidal Dispersions

In this section, we assume that Σ_k is any general nonnegative definite covariance matrix, for each $k \in \{1, \dots, K\}$. In this case, the k -means algorithm is no longer applicable so we used hierarchical clustering with Euclidean distance and Ward's criterion [42] to obtain the tree, which was subsequently cut to yield K clusters. The \mathcal{R} -value of this grouping relative to the true was taken to indicate the difficulty of clustering. The focus in this section is therefore to broadly relate our preliminary index and its adjustments to the \mathcal{R} thus obtained.

3.2.1 The Two-Groups Case

Figures 8a-i display simulated datasets for different values of our preliminary index \mathcal{I} using the algorithm in the appendix with $\alpha = 0.75$. In each case, 100 observations were generated from each of two groups separated according to \mathcal{I} . In these figures, color and character distinguish the true grouping. For each simulated dataset, we also obtained \mathcal{R} as described above. Figures 8a-c have the lowest \mathcal{R} between 0.56 and 0.67, but each subsequent row down has \mathcal{R} values, on the average, higher than in previous rows. In general, therefore, Figures 8a-i demonstrate as the value of \mathcal{I} increases, the clusters become more separated. This coincides with an increase in the values of \mathcal{R} . Thus lower values of \mathcal{I} correspond to clustering problems of higher difficulty while higher values of \mathcal{I} are associated with higher values of \mathcal{R} and hence clustering complexity.

Similar to Section 3.1, we investigated further the relationship between \mathcal{I} and \mathcal{R} in the case of nonhomogeneous groups. Specifically, we simulated 25 datasets each of observations from two nonhomogeneous populations each with arbitrary ellipsoidal dispersion structures for all possible combinations of $(n_1, n_2, p, \mathcal{I})$, where $n_1 \leq n_2$ (assumed w.l.o.g.), and n_1, n_2 are the numbers of observations in the two groups. In our experiments, $p \in \{2, 4, 5, 10, 20, 50, 100, 200, 500, 1000\}$ and $(n_1, n_2) \in \{(20, 20), (50, 50), (75, 75), (100, 100), (200, 200), (350, 350), (500, 100), (30, 100), (20, 50), (60, 75), (90, 100), (150, 250), (50, 600), (100, 500)\}$. We simulated 25 datasets according to each of ten values of \mathcal{I} spaced evenly between 0 and 1, and for $\alpha = 0.75$. Each

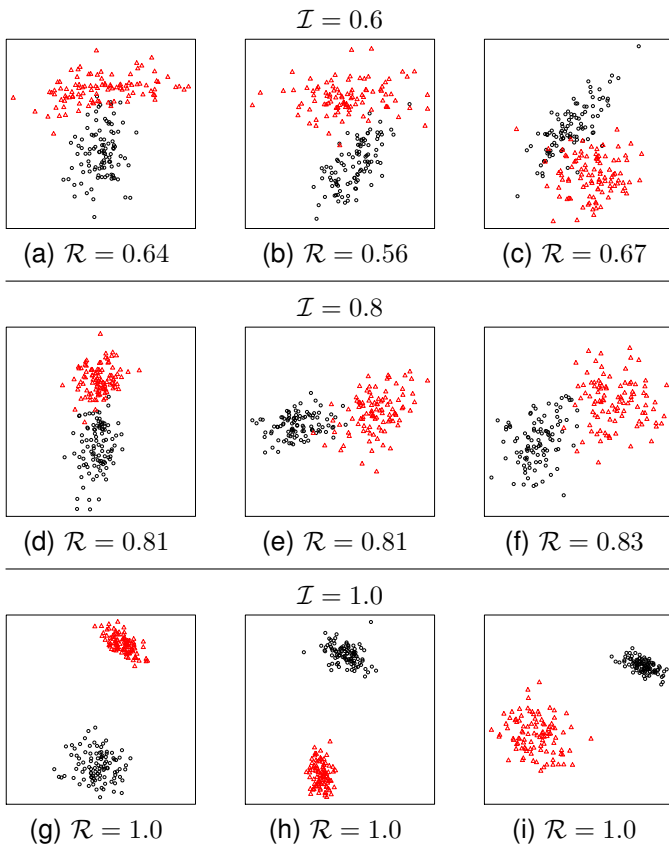


Fig. 8. Simulated datasets at different \mathcal{I} in the case with general ellipsoidal-dispersed groups. Color and plotting character represent true groups. Under each plot we also report the \mathcal{R} comparing the true grouping with that using hierarchical clustering with Ward’s linkage and $K = 2$.

dataset was partitioned into two groups using hierarchical clustering with [42]’s linkage and the resulting partition evaluated with the true grouping of the simulated dataset in terms of \mathcal{R} . We noticed nonlinearity in the general relationship between \mathcal{I} and \mathcal{R} so as in Section 3.1.1, we

explored a relationship between \mathcal{I} and \mathcal{R} along with the parameters p, n . Similar to (14), (15), (18) and (19) we found appropriate adjustments and defined $\mathcal{R}_{\mathcal{I},\alpha}$ for this case. (See the columns leveled *het* of Table 1 for the estimated values when $\alpha = 0.75$ and [38] for estimates obtained when α equals 0.25 and 0.50.)

3.2.2 The Case with Many Groups ($K \geq 2$)

Summarizing the index for $K > 2$ groups brings the same issues outlined in Section 3.1.2. We propose adapting [40]’s summarized multiple Jaccard similarity index in the same manner as before but noting that $\mathcal{R}_{\mathcal{I},\alpha}$ is calculated within the setting of nonhomogeneous ellipsoidal clusters. As in Section 3.1.2, we conducted an extensive simulation experiment to study the relationship between $\mathcal{R}_{\mathcal{I},\alpha}$ and \mathcal{R} . We simulated 25 datasets for each combination of $p, n_i, \alpha, \mathcal{R}_{\mathcal{I},\alpha}$ and K where $p \in \{2, 4, 5, 10, 20, 50, 100, 200\}$, $\alpha = 0.75$, $n_i = n_j = n_0$ for all i, j with $n_0 \in \{20, 50, 75, 100, 200, 500\}$, $K \in \{3, 5, 7, 10\}$ and $\mathcal{R}_{\mathcal{I},\alpha}$ evenly-spaced over ten values in $(0, 1)$. We partitioned each simulated dataset using hierarchical clustering with [42]’s linkage and calculated \mathcal{R} of the resulting partitioning relative to the true. We used the results to adjust our index as in (18) and (19) and obtained the final summary (\mathcal{I}_*) of similar form to (20), and coefficient estimates provided in Table 2.

Figure 9 presents results of simulation experiments in the same spirit as Figures 5 and 6 except that the datasets are now generated using the algorithm in the appendix with \mathcal{I}_* as described here. Note that as in Figures 6 the effect of K has been largely addressed and the relationship between \mathcal{I}_* and \mathcal{R} is largely consistent across dimensions. Thus the curse of dimensionality no longer affects our index and as such it is interpretable across dimensions.

3.3 Illustrative Examples

We now provide some illustrations of the range of multi-clustered datasets that can be obtained using the algorithm in the appendix and for different values of the summarized index

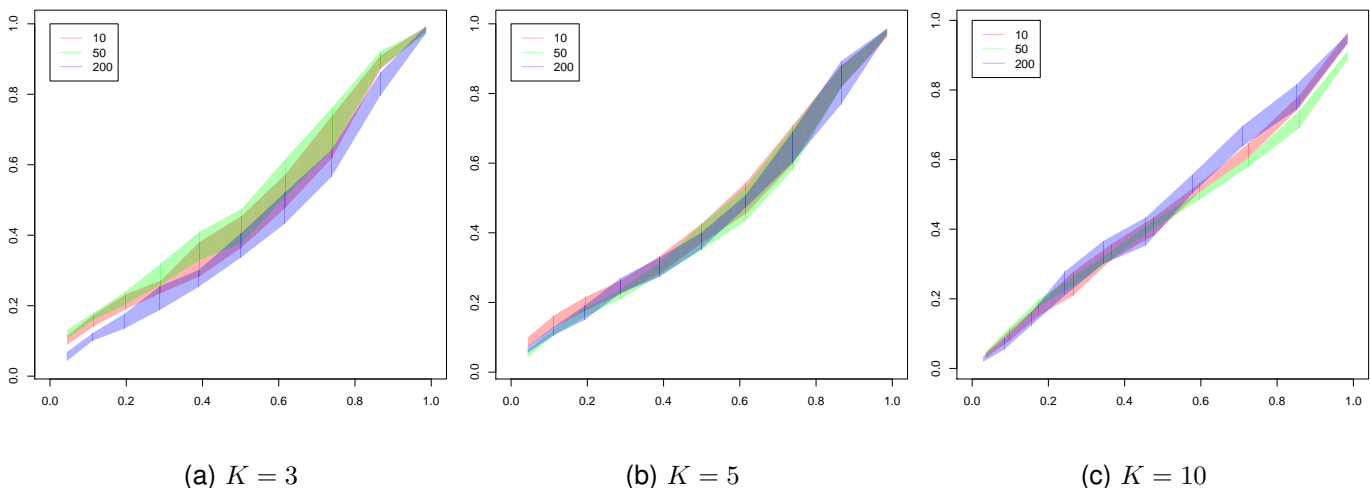


Fig. 9. Plot of \mathcal{R} against \mathcal{I}_* for the case with general ellipsoidal clusters. The three colors represent the dimensions of the simulated datasets ($p \in \{10, 50, 200\}$). Other aspects of the plot are as in Figure 2.

\mathcal{I}_* . We first display realizations obtained in two dimensions, and then move on to higher dimensions.

Figures 10a-i mimic the setup of Figure 7 for nonhomogeneous ellipsoidal clusters. Here the observations are grouped using hierarchical clustering with Ward's criterion [42] and cutting the tree hierarchy at $K = 4$. The grouping of the datasets in Figures 10a-c have the lowest \mathcal{R} between 0.54 and 0.64, but each subsequent row down has \mathcal{R} values, on the average, higher than in previous rows. In general, therefore, Figures 10a-i demonstrate as the value of \mathcal{I}_* increases, the clusters become more separated. This coincides with an increase in the values of \mathcal{R} . Thus lower values of \mathcal{I}_* correspond to clustering problems of higher difficulty while higher values of \mathcal{I}_* are associated with higher values of \mathcal{R} and hence clustering complexity. Figures 10a-i demonstrate the various possible configurations obtained using the algorithm in the appendix for three different values of \mathcal{I}_* . Note that in some cases only two of the four groups are well-separated while in other cases none of the groups are well-separated.

We used *Radviz* or radial visualization plots [43] to display multi-dimensional multi-class datasets in Figure 11. We display three realizations each of five-dimensional five-clustered datasets, obtained using our algorithm with \mathcal{I}_* values of 0.6, 0.8 and 1. Additional simulated realizations and other values of \mathcal{I}_* are presented in [38]. Below each plot we provide \mathcal{R} obtained upon comparing the partitioning obtained using

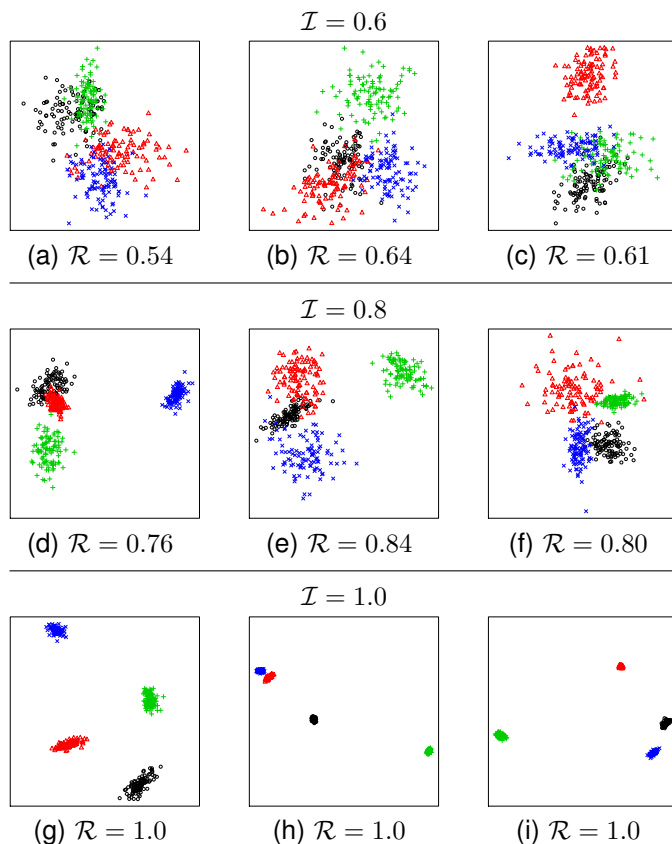


Fig. 10. Simulated datasets at different \mathcal{I}_* in heterogeneous case for $K = 4$, drawn in the same spirit as Figure 8.

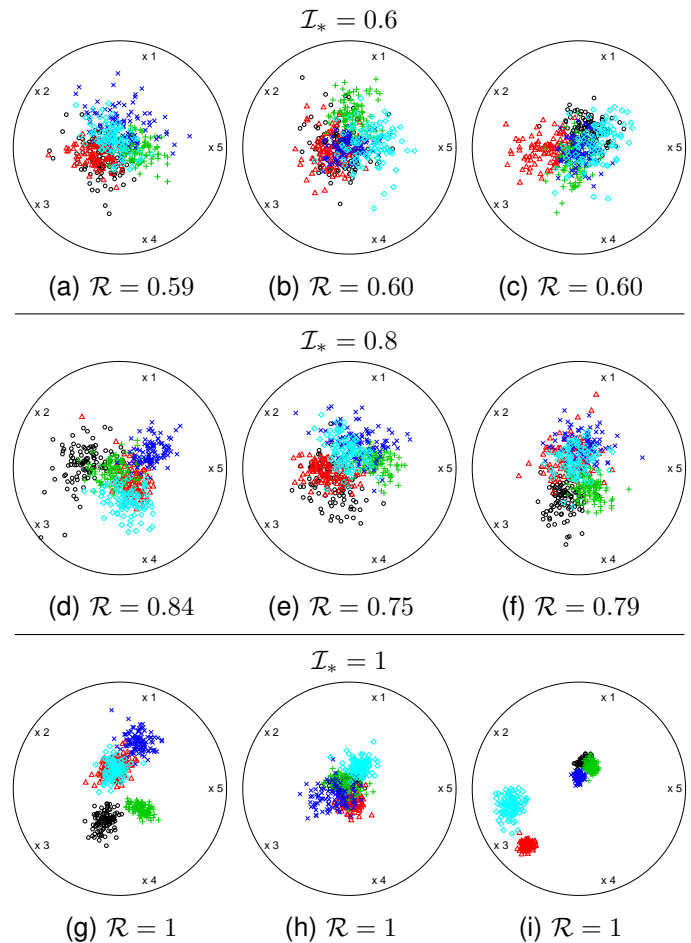


Fig. 11. Radial visualization plots of 5 dimension and 5 components at different values for \mathcal{I}_*

hierarchical clustering with [42]'s linkage. From Figure 11 we see that clusters with higher values of \mathcal{I}_* are more separated in general than datasets with lower values of \mathcal{I}_* . The clusters with $\mathcal{I}_* = 0.6$ appear to overlap quite a bit; however, the datasets corresponding to $\mathcal{I}_* = 1$ seem more separated. Additionally, \mathcal{I}_* is close to \mathcal{R} in all cases.

We close this section here by commenting on the role of α . Our preliminary index \mathcal{I}_α is dependent on the choice of α . However, our final adjustments ensuring the final index \mathcal{I}_* to be approximately linear in \mathcal{R} means that the value of α is not that crucial in our calculations.

4 APPLICATION TO GROUPED DATA

In this section we illustrate some applications of our index to some datasets which contains class information for each observation. The datasets studied here are the textures [10], wine [11], Iris [12], crabs [13], image [14], *E. coli* [15] and [16]'s synthetic two-dimensional datasets.

4.1 Estimate \mathcal{I}_* for classification datasets

Our first application is in using \mathcal{I}_* as a measure of how well-separated the groups in each dataset are, on the whole, and how they relate to the difficulty in clustering or classification.

TABLE 3

The estimated index \mathcal{I}_* ($\hat{\mathcal{I}}_*$), the index derived by [27] (QJ) and exact- c -separation (c) are presented with adjusted Rand indices ($\mathcal{R}_Q, \mathcal{R}_E$) obtained using quadratic discriminant analysis (QDA) and EM-clustering respectively on the datasets (\mathcal{S}) (i) textures, (ii) Ruspini, (iii) wine, (iv) image, (v) crabs, (vi) Iris and (vii) *E. coli*.

\mathcal{S}	n	p	K	$\hat{\mathcal{I}}_*$	\mathcal{R}_Q	\mathcal{R}_E	QJ	c
(i)	5500	37	11	1.0	1	1	0.87	0.09
(ii)	75	2	4	1.0	1	1	0.63	3.05
(iii)	178	13	3	0.94	0.98	0.98	0.49	0.19
(iv)	2,310	11	7	0.94	0.82	0.68	0.07	0.07
(v)	200	5	4	0.91	0.90	0.83	0.23	0.11
(vi)	150	4	3	0.90	0.94	0.90	0.57	0.97
(vii)	327	5	5	0.90	0.82	0.78	0.35	0.52

Table 3 provides summaries of our estimated values of \mathcal{I}_* ($\hat{\mathcal{I}}_*$) for each dataset. We use the true classification and the actual observations to obtain estimates of the mean and covariance of each cluster. Using the estimated mean, covariance and number of observations from each cluster we find estimates of \mathcal{I}_* for $\alpha = 0.75$ and using Mahalanobis distance. We also calculated \mathcal{R} based on the clustering using quadratic discriminant analysis (QDA) and EM-clustering done using the R package `MClust`. The corresponding calculated \mathcal{R} 's were called \mathcal{R}_Q and \mathcal{R}_E respectively. Note that our EM algorithms were initialized using the estimated mean and covariance matrix and the actual cluster size proportions. Therefore we consider the final clustering as the best we could possibly do over all other choices for initialization values. Table 3 compares the estimated values of $\hat{\mathcal{I}}_*$ and \mathcal{R} evaluating the classification on the corresponding dataset done using QDA and model-based clustering. The datasets are ordered in Table 3 from the largest to the smallest $\hat{\mathcal{I}}_*$ value. For each dataset, we also calculated [27]'s index as well as exact- c -separation. With the exception of the *Iris* and *image* datasets, higher values of $\hat{\mathcal{I}}_*$ correspond to higher values of both \mathcal{R}_Q and \mathcal{R}_E . The relationship between $\hat{\mathcal{R}}$ and c -separation is not as clear. The textures dataset for example has $\hat{\mathcal{R}} = 1$ but a very small c -separation = 0.09. There also does not appear to be much relationship between the index of [27] and \mathcal{R} . Both c -separation and [27] do however pick up *image* as the most difficult dataset to group.

4.2 Indexing Distinctiveness of Classes

In this section, we discuss the use of our index to summarize the distinctiveness of each class with respect to the other. To illustrate this point, we estimate the pairwise indices $\mathcal{R}_{\mathcal{I},\alpha}$ of Section 3.2.1 for each pair of classes in each dataset. These pairwise indices are presented in Figure 12. For each dataset, we display the value quantitatively and qualitatively by means of a color map. Darker values indicate well-separated groups with index values closer to 1 while lighter regions represent pairs of groups that are harder to separate. The map of these pairwise indices provides us with an idea of the groups that are easier or harder to separate. In the *Iris* dataset example, $\hat{\mathcal{R}}_{\mathcal{I},0.75} = 1$ for species 1 (*I. Setosa*) and 2 (*I. Versicolor*) and for species groups 1 (*I. Setosa*) and 3 (*I. Virginica*).

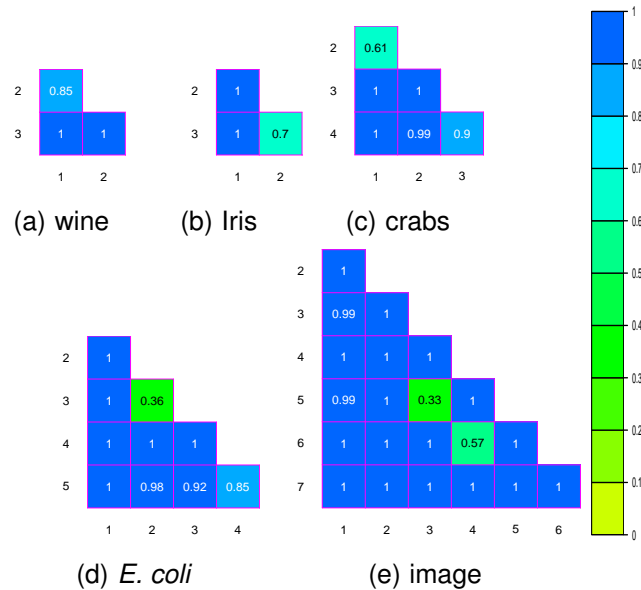


Fig. 12. Plots displaying the the pairwise indices $\hat{\mathcal{R}}_{\mathcal{I},0.75}$ for five commonly used classification datasets.

This indicates very-well separated groups. On the other hand, $\hat{\mathcal{R}}_{\mathcal{I},0.75} = 0.7$ for the species groups 2 (*I. Versicolor* and *I. Virginica*). This suggests that the difficulty in classifying the *Iris* dataset is largely due to the similarities between the species *I. Virginica* and *I. Versicolor*. The wine dataset is similar to the *Iris* dataset in that most of the difficulty in classification or clustering appears due to just two groups, as evidenced by $\hat{\mathcal{R}}_{\mathcal{I},0.75} = 0.847$ for between Groups 1 and 2. All other pairs of groups in the wine dataset have a value of $\hat{\mathcal{R}}_{\mathcal{I},0.75}$ greater than 0.99. The pairwise indices of the crabs dataset also produces some interesting insights into the difficulty of classification or clustering. Note that this dataset has 50 male and 50 female crabs each of orange and blue colored crabs. An interesting finding is that $\hat{\mathcal{R}}_{\mathcal{I},0.75}$ is above 0.99 for all pairs of categories except for the two pairs of crab groups having the same color. For the blue male and female crabs, we have $\hat{\mathcal{R}}_{\mathcal{I},0.75} = 0.61$ while $\hat{\mathcal{R}}_{\mathcal{I},0.75} = 0.90$ for the orange male and female crabs. This suggests that colors separate the crab groups better than gender and thus the difficulty in clustering this dataset is largely to the small differences between genders. Both the Ruspini dataset and the textures dataset have $\hat{\mathcal{R}}_{\mathcal{I},0.75} = 1$ for all pairs of groups and are not displayed for brevity. Figure 12d displays the pairwise values of $\hat{\mathcal{R}}_{\mathcal{I},0.75}$ for the 5 groups in the *E. coli* dataset. Once again, groups 2 and 3 are very difficult to separate, since $\hat{\mathcal{R}}_{\mathcal{I},0.75} = 0.365$ between these two groups. Figure 12e displays the pairwise values for $\hat{\mathcal{R}}_{\mathcal{I},0.75}$ for the seven groups in the image dataset. Once again, there is very little separation between groups 3 and 5 corresponding to images of foliage and a window as well as groups 4 and 6 corresponding to images of cement and a path. All other pairs of groups appear very well-separated.

Our pairwise index thus provides an idea of how separated each individual grouping is relative to the others. Then, we can use our index to characterize the relationships between

groups of observations in a dataset. The above can also be used to characterize derived groupings from clustering algorithms using Euclidean or Mahalanobis distances. The overall index \mathcal{I}_* provides us with a sense of the quality of the derived grouping, while the pairwise indices provide us with the relative distinctiveness of any two groups, and how easy it is to tell them apart.

5 CONCLUSION

In this paper, we derive a separation index \mathcal{I} for Euclidean or Mahalanobis' distance-based classification or hard clustering algorithms. Our index is geared towards capturing the difficulty in clustering and classification of datasets formulated in terms of order statistics for the difference in the distance of an observation from the center of a cluster to that from the center of its own cluster. Furthermore, we found an adjusted version of our index \mathcal{I}_* which is consistent over different combinations of number of observations, dimensions and number of clusters. Our \mathcal{I}_* is found to track very well the Adjusted Rand index [37] and thus is a good surrogate for clustering complexity. We have also explored some theoretical properties of our index. Our index is general enough to handle any number of overlapping and non-overlapping clusters, dimensions, sample sizes and ellipsoidal dispersion structures. We have used the index to characterize separation between different pairs of groups in datasets where class information is available. The appendix also provides an algorithm for generating datasets of a given sample size according to a pre-specified index. Although this index was developed within the context of Gaussian clusters, we can define an approximate version of the index for the non-Gaussian but ellipsoidal cases using the true means, true covariance matrices, but under the assumption of normality. Of course, relevance of this index in this case will depend on how dissimilar the groups are from Gaussian. For the even more general case, the preliminary index is much more difficult to compute under more general distributional assumptions. We suggest therefore, using multivariate transformations, such as the multivariate Box-Cox transformation [44] to obtain approximate versions of the index and for corresponding algorithms to simulate clustered data.

There are several ways our index could be utilized. As mentioned in the paper, our index could be used to characterize differences between different groups for datasets for which class information is available. Along with the algorithm in the appendix, our index could be used to generate datasets for a fixed value of \mathcal{I}_* . Such datasets could be used to evaluate the performance of different clustering and classification methodologies for a wide range of situations having different clustering difficulties. This would provide for a comprehensive understanding of the strengths and weaknesses of these methods. For instance, we have evaluated the performance of a wide range of initialization strategies proposed in the literature for the k -means algorithms vis-a-vis clustering difficulty [45]. Our index could also be used to refine partitions that have already been obtained via some algorithm. This is similar to the work of [8] or [9], both of whom start with an initial

Gaussian model-based clustering of the dataset. [8] combine the obtained components in a hierarchical fashion based on an entropy criterion. In addition, [9] also suggested combining components using a hierarchical approach based on concepts of unimodality and misclassification. It would be interesting to investigate performance of similar algorithms using \mathcal{I}_* . Thus, we note that there are a wide range of other scenarios where our index may be useful.

APPENDIX CLUSTER GENERATION ALGORITHM

We use \mathcal{I}_K^* to represent the generic version of the index for a dataset in p dimensions having K clusters with the number of observations n_1, n_2, \dots, n_K . The objective is to produce a dataset for which \mathcal{I}_K^* has a desired pre-specified value ι . The main steps are as follows:

- 1) **Parameter initialization:** If desired, the user can provide starting means $\mu_1, \mu_2, \dots, \mu_K$ and covariance matrices $\Sigma_1, \Sigma_2, \dots, \Sigma_K$. Otherwise, the initial parameters are generated as follows: Generate initial $\mu_1, \mu_2, \dots, \mu_K$ randomly from $U[0, a]^p$, and use these as the means for the K groups. To generate K covariance matrices Σ_k in the homogeneous spherical covariance case, we randomly generate $\sigma^2 > 0$ from one-dimensional uniform distribution and choose our dispersion matrices as diagonal matrices with the diagonal element equal to σ^2 . If the desired covariance matrices are general spherical we take a sample of K from a chosen distribution (such as a uniform distribution) and the dispersion matrices are diagonal with these chosen diagonal elements. Otherwise, we generate K covariate matrices from the Wishart distribution with degrees of freedom equal to p^2 and scale matrix equal to the identity matrix of dimension p . This choice of degrees of freedom allows for great flexibility in the shape and orientation of the Covariance matrices. However, the generated covariance matrix may be close to singular. To address this issue we use the method proposed by [33] and put a restriction on the maximum eigenvalue relative to the minimum eigenvalue. Let $\lambda_{(1)} > \lambda_{(2)} > \dots > \lambda_{(p)}$ where $\lambda_{(i)}$ is the i^{th} largest eigenvalue of a proposal covariance matrix Σ_k . Thus given a realization Σ_k , which we can decompose Σ_k such that $\Sigma_k = \Gamma_k \Lambda_k \Gamma_k'$ with Λ_k and Γ_k as the the diagonal matrix of eigenvalues and matrix of eigenvectors respectively. Assume Λ_k has diagonals in decreasing order. Define *maximum eccentricity* $e_{max} = \sqrt{1 - \lambda_{(p)}/\lambda_{(1)}}$. Then we specify the condition that e_{max} must be less than a predetermined value for all K . We suggest using the restriction that $e_{max} \leq .95$. For those Σ_k 's (say, Σ_m) for which $e_k > e_{max}$, we let $\lambda_{m,(j)}^* = \lambda_{m,(j)}$ when $\lambda_{m,(j)} \leq \lambda_{m,(p)}/(1 - e_{max}^2)$ otherwise set $\lambda_{m,(j)}^*$ to random draw from uniform($\lambda_{m,(p)}$, $\lambda_{m,(p)}/(1 - e_{max}^2)$). Then reconstruct $\Sigma_k^* = \Gamma_k \Lambda_k^* \Gamma_k'$ where Λ_k^* is the diagonal matrix of new eigenvalues $\lambda_{m,(1)}^* \geq \lambda_{m,(2)}^* \geq \dots \geq \lambda_{m,(p)}^*$.

- 2) **Calculating current value of the index:** For the current configuration of the parameters, compute the index for a given choice of $\alpha = 0.25, 0.5$ or 0.75 : For 2 clusters, the index is defined in (5) is calculated as described in Remark 2. For general versions of the index, each pairwise index is calculated as above and are combined (for $K \geq 3$) and adjusted using transformations specified in Sections 3.1 and 3.2.
- 3) **Iteration step and termination:** For a small ϵ , the pre-specified tolerance level, stop if $|\mathcal{I}_K^*(c) - \iota| < \epsilon$, and the current configuration will generate data having the index values (close to) ι . If not, we use bisection method to find $c > 0$ and go to step 2 using the parameters $(n_k, c\mu_k, \Sigma_k^*) : k = 1, 2, \dots, K$.

REFERENCES

- [1] J. A. Hartigan, "Statistical theory in clustering," *Journal of Classification*, vol. 2, pp. 63–76, 1985.
- [2] D. B. Ramey, "Nonparametric clustering techniques," in *Encyclopedia of Statistical Science*. New York: Wiley, 1985, vol. 6, pp. 318–319.
- [3] L. Kaufman and P. J. Rousseeuw, *Finding Groups in Data*. New York: John Wiley and Sons, Inc., 1990.
- [4] B. S. Everitt, S. Landau, and M. Leesem, *Cluster Analysis (4th ed.)*. New York: Hodder Arnold, 2001.
- [5] G. J. McLachlan and K. E. Basford, *Mixture Models: Inference and Applications to Clustering*. New York: Marcel Dekker, 1988.
- [6] C. Fraley and A. E. Raftery, "Model-based clustering, discriminant analysis, and density estimation," *Journal of the American Statistical Association*, vol. 97, pp. 611–631, 2002.
- [7] J. R. Kettner, "The practice of cluster analysis," *Journal of Classification*, vol. 23, pp. 3–30, 2006.
- [8] J. P. Baudry, A. E. Raftery, G. Celeux, K. Lo, and R. Gottardo, "Combining mixture components for clustering," *Journal of Computational and Graphical Statistics*, vol. 19, no. 2, pp. 332–353, 2010.
- [9] C. Hennig, "Methods for merging Gaussian mixture components," *Advances in Data Analysis and Classification*, vol. 4, no. 1, pp. 3–34, 2010.
- [10] P. Brodatz, *A Photographic Album for Artists and Designers*. New York: Dover, 1966.
- [11] M. F. et al, "PARVUS - an extendible package for data exploration, classification and correlation," *Institute of Pharmaceutical and Food Analysis and Technologies, Via Brigata Salerno*, 1991.
- [12] E. Anderson, "The irises of the gaspe peninsula," *Bulletin of the American Iris Society*, vol. 59, pp. 2–5, 1935.
- [13] R. J. Campbell, N. A. and Mahon, "A multivariate study of variation in two species of rock crab of genus *leptograsmus*," *Australian Journal of Zoology*, vol. 22, pp. 417–25, 1974.
- [14] D. Newman, S. Hettich, C. L. Blake, and C. J. Merz, "UCI repository of machine learning databases," 1998. [Online]. Available: <http://www.ics.uci.edu/~mllearn/MLRepository.html>
- [15] K. Nakai and M. Kinehasa, "Expert system for predicting protein localization sites in gram-negative bacteria," *PROTEINS: Structure, Function, and Genetics*, vol. 11, pp. 95–110, 1991.
- [16] E. H. Ruspini, "Numerical methods for fuzzy clustering," *Information Science*, vol. 2, pp. 319–350, 1970.
- [17] G. W. Milligan, "An algorithm for generating artificial test clusters," *Psychometrika*, vol. 50, pp. 123–127, 1985.
- [18] G. W. Milligan, S. C. Soon, and L. M. Sokol, "The effect of cluster size, dimensionality, and the number of clusters on recovery of true cluster structure," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 5, pp. 40–47, 1983.
- [19] G. W. Milligan and M. C. Cooper, "A study of the comparability of external criteria for hierarchical cluster analysis," *Multivariate Behavioral Research*, vol. 21, pp. 441–458, 1986.
- [20] —, "Study of standardization of variables in cluster analysis," *Journal of Classification*, vol. 5, pp. 181–204, 1988.
- [21] P. V. Balakrishnan, M. C. Cooper, V. S. Jacob, and P. A. Lewis, "A study of the classification capabilities of neural networks using unsupervised learning: A comparison with k-means clustering," *Psychometrika*, vol. 59, pp. 509–525, 1994.
- [22] M. J. Brusco and J. D. CREDIT, "A variable-selection heuristic for K-means clustering," *Psychometrika*, vol. 66, pp. 249–270, 2001.
- [23] R. Maitra and I. P. Ramler, "Clustering in the presence of scatter," *Biometrics*, vol. 65, pp. 341–352, 2009.
- [24] R. M. McIntyre and R. K. Blashfield, "A nearest-centroid technique for evaluating the minimum-variance clustering procedure," *Multivariate Behavioral Research*, vol. 15, pp. 225–238, 1980.
- [25] D. Steinley and R. Henson, "OCLUS: An analytic method for generating clusters with known overlap," *Journal of Classification*, vol. 22, pp. 221–250, 2005.
- [26] W. Qiu and H. Joe, "Separation index and partial membership for clustering," *Computational Statistics and Data Analysis*, vol. 50, pp. 585–603, 2006.
- [27] —, "Generation of random clusters with specified degree of separation," *Journal of Classification*, vol. 23, pp. 315–334, 2006.
- [28] N. Likas, A. Vlassis, and J. J. Verbeek, "The global k-means clustering algorithm," *Pattern Recognition*, vol. 36, pp. 451–461, 2003.
- [29] J. Verbeek, N. Vlassis, and B. Krose, "Efficient greedy learning of gaussian mixture models," *Neural Computation*, vol. 15, pp. 469–485, 2003.
- [30] J. Verbeek, N. Vlassis, and J. Nunnink, "A variational EM algorithm for large-scale mixture modeling," *Annual Conference of the Advanced School for Computing and Imaging*, pp. 1–7, 2003.
- [31] S. Dasgupta, "Learning mixtures of gaussians," in *Proc. IEEE Symposium on Foundations of Computer Science*, New York, 1999, pp. 633–644.
- [32] R. Maitra, "Initializing partition-optimization algorithms," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 6, pp. 144–157, 2009.
- [33] R. Maitra and V. Melnykov, "Simulating data to study performance of finite mixture modeling and clustering algorithms," *Journal of Computational and Graphical Statistics*, vol. 19, no. 2, pp. 354–376, 2010.
- [34] J. A. Hartigan, *Clustering Algorithms*. New York: Wiley, 1975.
- [35] J. A. Hartigan and M. A. Wong, "A k-means clustering algorithm," *Applied Statistics*, vol. 28, pp. 100–108, 1979.
- [36] R. Davies, "The distribution of a linear combination of χ^2 random variables," *Applied Statistics*, vol. 29, pp. 323–333, 1980.
- [37] L. Hubert and P. Arabie, "Comparing partitions," *Journal of Classification*, vol. 2, pp. 193–218, 1985.
- [38] A. P. Ghosh, R. Maitra, and A. D. Peterson, "A separability index for distance-based clustering and classification algorithms," Iowa State University, Department of Statistics, Ames, IA, Tech. Rep. 06, 2010, 2010.
- [39] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*, 2nd ed., ser. Springer Series in Statistics. Springer, September 2009. [Online]. Available: <http://www-stat.stanford.edu/tibs/ElemStatLearn/main.html>
- [40] R. Maitra, "A re-defined and generalized percent-overlap-of-activation measure for studies of fMRI reproducibility and its use in identifying outlier activation maps," *Neuroimage*, vol. 50, no. 1, pp. 124–135, 2010.
- [41] P. Jaccard, "Étude comparative de la distribution florale dans une portion des alpes et des jura," *Bulletin del la Société Vaudoise des Sciences Naturelles*, vol. 37, p. 547?579, 1901.
- [42] J. H. Ward, "Hierarchical grouping to optimize an objective function," *Journal of American Statistical Association*, vol. 58, pp. 236–244, 1963.
- [43] P. Hoffman, G. G., K. Marx, I. Grosse, and E. Stanley, "DNA visual and analytic data mining," in *IEEE Visualization '97 Proceedings*, Phoenix, AZ, 1997, pp. 437–441. [Online]. Available: <http://www.cs.uml.edu/phoffman/viz>
- [44] G. E. P. Box and D. R. Cox, "An analysis of transformations," *Journal of the Royal Statistical Society*, vol. 26, no. 2, pp. 211–252, 1964.
- [45] A. D. Peterson, A. P. Ghosh, and R. Maitra, "A systematic evaluation of different methods for initializing the k-means clustering algorithm," Iowa State University, Department of Statistics, Ames, IA, Tech. Rep. 07, 2010, 2010.