SN Computer Science _####################_
https://doi.org/10.1007/s42979-020-0067-z

ORIGINAL RESEARCH

A Probabilistic Distance Clustering Algorithm Using Gaussian and Student-*t* Multivariate Density Distributions

Cristina Tortora[1] · **Paul D. McNicholas**[2] · **Francesco Palumbo**[3]

Received: 4 November 2019 / Accepted: 9 January 2020

Abstract

A new dissimilarity measure for cluster analysis is presented and used in the context of probabilistic distance (PD) clustering. The basic assumption of PD-clustering is that for each unit, the product between the probability of the unit belonging to a cluster and the distance between the unit and the cluster is constant. This constant is a measure of the classifiability of the point, and the sum of the constant over units is called joint distance function (JDF). The parameters that minimize the JDF maximize the classifiability of the units. The new dissimilarity measure is based on the use of symmetric density functions and allows the method to find clusters characterized by different variances and correlation among variables. The multivariate Gaussian and the multivariate Student-*t* distributions have been used, outperforming classical PD clustering, and its variation PD clustering adjusted for cluster size, on simulated and real datasets.

Keywords Cluster analysis · PD-clustering · Multivariate distributions · Dissimilarity measures

Introduction

Cluster analysis refers to a wide range of numerical methods aiming to find distinct groups of homogeneous units. Clustering in two or three dimensions is a natural task that humans can often do visually; however, machine approaches are needed for all but such low dimensions. We focus on partitioning clustering methods; given a number of clusters K, partitioning methods assign units to the K clusters optimizing a given criterion. These methods are generally divided into not model-based and model-based, according to the distributional assumptions. Model-based clustering or finite mixture model clustering assumes that the population probability density function is a convex linear combination of a finite number of density functions; accordingly, they are very well suited to clustering problems. A variety of methods and algorithms have been proposed for finite mixture model parameter estimation. The most widely used strategy is to find the parameters that maximize the complete-data likelihood function using the expectation-maximization (EM) algorithm, which was proposed in 1977 [10] building on prior work (e.g., [5, 7, 23, 24]). The non-model-based methods generally optimize a criterion based on distance or dissimilarity measures. Different dissimilarity measures can be used based on the type of data, in this paper we focus on continuous data.

Formally, let us consider an $n \times J$ data matrix $\mathbf{X}$, with generic row vector $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{iJ})$. Partitioning algorithms aim to find a set of K clusters, $\mathscr{C}_k$, with $k = 1, \ldots, K$, such that the elements inside a cluster are homogeneous and $\mathscr{C}_1 \cup \mathscr{C}_2 \cup \cdots \cup \mathcal{C}_K = X$. If, for any pair $\{k, k'\} \in 1, \ldots, K$, $\mathcal{C}_k \cap \mathscr{C}_{k'} = \emptyset$, then the clustering technique is called hard or crisp, otherwise it is called fuzzy or soft. In the latter case, each unit can belong to more than one cluster with certain membership degree.

The most frequently used non-model-based methods for continuous data are k-means [20] and its fuzzy analogue c-means [4], which minimize the sum of the within groups sum of squares over all variables. In spite of their simplicity, the optimal solution can only be found applying an iterative intuitively reasonable procedure. More recently, [3] proposed probabilistic distance (PD) clustering, a distribution free fuzzy clustering technique (i.e., non-model-based),

✉ Cristina Tortora
cristina.tortora@sjsu.edu

[1] Department of Mathematics and Statistics, San José State University, San José, CA, USA

[2] Department of Mathematics and Statistics, McMaster University, Hamilton, ON, Canada

[3] Dipartimento di Scienze Politiche, University of Naples Federico II, Naples, Italy

Journal : Large 42979	Article No : 67	Pages : 22	MS Code : 67	Dispatch : 21-1-2020

where the membership degree is defined as heuristic probability. PD clustering optimization problems represents a special case of the Weber–Fermat's problem, when the number of the 'attraction points' is greater or equal to three, see [16] among others. In this framework, PD clustering assumes that the product of the probability of a point belonging to a cluster and the distance of the point from the center of the cluster is constant, and this constant is a measure of the classificability of the point. The method obtains the centers that maximize the classificability of all the points. A newer version of the algorithm that considers clusters of different size, PDQ-clustering, was proposed by [14] and an extension for high-dimensional data was proposed by [35, 36].

Generally, non-model-based clustering techniques are only based on the distances between the points and the centers; therefore, they do not take into account the shape and the size of the clusters. Accordingly, these techniques may fail when clusters are either non-spherical or spherical with different radii. To overcome this issue we propose a new dissimilarity measure based on symmetric density functions that have the advantage of considering the variability and the correlation among the variables. We use two different density functions, the multivariate Gaussian and the multivariate Student-t, but it could be extended to other symmetric densities. We then integrate this measure with PD-clustering and obtain new more flexible clustering techniques. Preliminary results can be found in [29].

After a background section on PD-clustering and PDQ-clustering, Sect. "Background", we introduce the new dissimilarity measure and the new techniques, Sect. "Flexible Extensions of PD-Clustering". We then compare them with some model-based and distance-based algorithms on simulated and real datasets, Sect. Empirical Evidence from Simulated and Real Data.

Background

In this section we briefly introduce PD-clustering [3], a distance-based soft clustering algorithm, and its extension, PD-clustering adjusted for cluster size [14].

Probabilistic Distance Clustering

Ben-Israel and Iyigun [3] proposed a non-hierarchical distance-based clustering method, called probabilistic distance (PD) clustering. They then extended the method to account for clusters of different size, i.e., PDQ [14]. Tortora et al. [35] proposed a factor version of the method to deal with high-dimensional data. Recently, [29] further extended the method to include more flexibility.

In PD-clustering, the number of clusters K is assumed to be a priori known, and a wide review on how to choose K can be found in [8]. Given some random centers, the probability of any point belonging to a cluster is assumed to be inversely proportional to the distance from the center of that cluster [13]. Suppose we have a data matrix $\mathbf{X}$ with N units and J variables, and consider K (non-empty) clusters. PD-clustering is based on two quantities: the distance of each data point $\boldsymbol{x}_i$ from each cluster centre $\mathbf{c}_k$, denoted $d(\boldsymbol{x}_i, \mathbf{c}_k)$, and the probability of each point belonging to a cluster, i.e., $p(\boldsymbol{x}_i, \mathbf{c}_k)$, for $k = 1, \dots, K$ and $i = 1, \dots, N$.

For convenience, define $p_{ik} := p(\boldsymbol{x}_i, \mathbf{c}_k)$ and $d_{ik} := d(\boldsymbol{x}_i, \mathbf{c}_k)$. PD-clustering is based on the principle that the product of the distances and the probabilities is a constant depending only on $\boldsymbol{x}_i$ [3]. Denoting this constant as $F(\boldsymbol{x}_i)$, we can write this principle as

$$p_{ik} d_{ik} = F(\boldsymbol{x}_i), \tag{1}$$

where $F(\boldsymbol{x}_i)$ depends only on $\boldsymbol{x}_i$, i.e., $F(\boldsymbol{x}_i)$ does not depend on the cluster k. As the distance from the cluster centre decreases, the probability of the point belonging to the cluster increases. The quantity $F(\boldsymbol{x}_i)$ is a measure of the closeness of $\boldsymbol{x}_i$ to the cluster centres, and it determines the classificability of the point $\boldsymbol{x}_i$ with respect to the centres $\mathbf{c}_k$, for $k = 1, \dots, K$. The smaller the $F(\boldsymbol{x}_i)$ value, the higher the probability of the point belonging to one cluster. If all of the distances between the point $\boldsymbol{x}_i$ and the centers of the clusters are equal to d_i, then $F(\boldsymbol{x}_i) = d_i/K$ and all of the probabilities of belonging to each cluster are equal, i.e., $p_{ik} = 1/K$. The sum of $F(\boldsymbol{x}_i)$ over i is called joint distance function (JDF). Starting from (1), it is possible to compute p_{ik}, i.e.,

$$p_{ik} = \frac{\prod_{m \neq k} d_{im}}{\sum_{m=1}^{K} \prod_{r \neq m} d_{ir}}, \tag{2}$$

for $k = 1, \dots K$, and $i = 1, \dots, N$. The whole clustering problem consists in the identification of the centers that minimize the JDF:

$$\text{JDF} = \sum_{i=1}^{n} \sum_{k=1}^{K} d_{ik} p_{ik}. \tag{3}$$

Extensive details on PD clustering are given in [3], who suggest using p^2 in (3) because it is a smoothed version of the problem. It follows that the optimized functions become

$$\text{JDF} = \sum_{i=1}^{n} \sum_{k=1}^{K} d_{ik} p_{ik}^2, \tag{4}$$

and the centers can be computed as

$$c_k = \frac{\sum_{i=1}^{N} u_{ik} \boldsymbol{x}_i}{\sum_{j=1}^{N} u_{jk}}, \tag{5}$$

with $u_{ik} = p_{ik}^2 / d_{ik}$.

Journal : Large 42979	Article No : 67	Pages : 22	MS Code : 67	Dispatch : 21-1-2020

It is worth noting that the function p_{ik} respects all necessary conditions to be a probability and yet no assumptions are made on the distribution of this function; further, p_{ik} can only be computed given $\boldsymbol{x}_i$ and for every $\mathbf{c}_k$ [28]. Following [13], we refer to p_{ik} as subjective probabilities, which are based on degree of belief (see [2]).

PD Clustering Adjusted for Cluster Size

The probabilities obtained using 1 do not consider the cluster size, and the algorithm tends to fail when clusters are unbalanced. Moreover, the resulting clusters have similar variance and covariance matrices. To overcome these issues [14] proposed PD-clustering adjusted for cluster size (PDQ). They assume that

$$\frac{p_{ik}^2 d_{ik}}{q_k} = F(\boldsymbol{x}_i), \tag{6}$$

where q_k is the cluster size, with the constraint that $\sum_{k=1}^K q_k = N$. The p_{ik} can then be computed via

$$p_{ik} = \frac{\prod_{m \neq k} d_{im}/q_m}{\sum_{m=1}^K \prod_{r \neq m} d_{ir}/q_r}. \tag{7}$$

The cluster size is considered a variable, the value of q_k that minimizes (6) is

$$q_k = N \frac{\left(\sum_{i=1}^N d_{ik} p_{ik}^2\right)^{1/2}}{\sum_{k=1}^K \left(\sum_{i=1}^N d_{ik} p_{ik}^2\right)^{1/2}}, \tag{8}$$

for $k = 1, \dots, K-1$, and

$$q_K = N - \sum_{k=1}^{K-1} q_k.$$

Flexible Extensions of PD-Clustering

Gaussian PD-Clustering

The PDQ algorithm can detect clusters of different size and with different within-cluster variability; however, it can still fail at detecting the clustering partition when variables are correlated or when there are outliers in the data. To overcome these issues we proposed a new dissimilarity measure based on a density function. Let $M_k = \max\{f(\boldsymbol{x}_i; \boldsymbol{\mu}_k, \boldsymbol{\theta}_k)\}$ and define the quantity

$$\delta_{ik} = \log\left(M_k f(\boldsymbol{x}_i; \boldsymbol{\mu}_k, \boldsymbol{\theta}_k)^{-1}\right), \tag{9}$$

which is a dissimilarity measure where $f(\boldsymbol{x}_i; \boldsymbol{\mu}_k, \boldsymbol{\theta}_k)$ is a symmetric unimodal density function with location parameter $\boldsymbol{\mu}_k$ and parameter vector $\boldsymbol{\theta}_k$. See appendix for the proof.

Recall that the density of a multivariate Gaussian distribution is

$$\phi(\boldsymbol{x}_i; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{\frac{J}{2}}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}-\boldsymbol{\mu})\right\}, \tag{10}$$

and define $\phi_{ik} := \phi(\boldsymbol{x}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, where $k = 1, \dots, K$. Using (10) in (9) and the result in (6), the JDF becomes

$$\begin{aligned} \mathrm{JDF} = & \sum_{i=1}^n \sum_{k=1}^K \frac{p_{ik}^2}{q_k} \log(M_k) + \sum_{i=1}^n \sum_{k=1}^K \frac{1}{2} \frac{p_{ik}^2}{q_k} \log\left((2\pi)^J |\boldsymbol{\Sigma}_k|\right) \\ & + \sum_{i=1}^n \sum_{k=1}^K \frac{1}{2} \frac{p_{ik}^2}{q_k} (\boldsymbol{x}_i - \boldsymbol{\mu}_k)' \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}_k). \end{aligned} \tag{11}$$

This technique is called Gaussian PD-clustering, GPDC, and it offers many advantages when compared to PD-clustering. The new dissimilarity measure already takes into account the impact of different within cluster variances and the correlation among variables.

The clustering problem now consists in the estimation of $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$, with $k = 1, \dots, K$, that minimize (11). A differentiation procedure leads to these estimates. An iterative algorithm is then used to compute the belonging probabilities and update the parameter estimates. More specifically, differentiating (11) with respect to $\boldsymbol{\mu}_k$ gives

$$\frac{\partial \mathrm{JDF}}{\partial \boldsymbol{\mu}_k} = -\frac{1}{2} \sum_{i=1}^n \frac{p_{ik}^2}{q_k} \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}_k). \tag{12}$$

Setting (12) equal to zero and solving for $\boldsymbol{\mu}_k$ gives

$$\boldsymbol{\mu}_k = \frac{\sum_{i=1}^n p_{ik}^2 \boldsymbol{x}_i}{\sum_{i=1}^n p_{ik}^2} \tag{13}$$

Now, differentiating (11) with respect to $\boldsymbol{\Sigma}_k$ gives

$$\begin{aligned} \frac{\partial \mathrm{JDF}}{\partial \boldsymbol{\Sigma}_k} &= \sum_{i=1}^n \frac{1}{2} \frac{p_{ik}^2}{q_k} \boldsymbol{\Sigma}_k^{-1} - \boldsymbol{\Sigma}_k^{-1} \sum_{i=1}^n \frac{1}{2} (\boldsymbol{x}_i - \boldsymbol{\mu}_k)(\boldsymbol{x}_i - \boldsymbol{\mu}_k)' \frac{p_{ik}^2}{q_k} \boldsymbol{\Sigma}_k^{-1} \\ &= \frac{1}{2} \boldsymbol{\Sigma}_k^{-1} \left[\sum_{i=1}^n \frac{p_{ik}^2}{q_k} - \sum_{i=1}^n (\boldsymbol{x}_i - \boldsymbol{\mu}_k)(\boldsymbol{x}_i - \boldsymbol{\mu}_k)' p_{ik}^2 \boldsymbol{\Sigma}_k^{-1} \right]. \end{aligned} \tag{14}$$

Setting (14) equal to zero and solving for $\boldsymbol{\Sigma}_k$ gives

$$\boldsymbol{\Sigma}_k = \frac{\sum_{i=1}^n (\boldsymbol{x}_i - \boldsymbol{\mu}_k)(\boldsymbol{x}_i - \boldsymbol{\mu}_k)' p_{ik}^2}{\sum_{i=1}^n p_{ik}^2}. \tag{15}$$

It follows that, at generic iteration $(t+1)$, the parameters that minimize the (11) are:

Journal : Large 42979 | Article No : 67 | Pages : 22 | MS Code : 67 | Dispatch : 21-1-2020

$$\boldsymbol{\mu}_k^{(t+1)} = \frac{\sum_{i=1}^n p_{ik}^2 \boldsymbol{x}_i}{\sum_{i=1}^n p_{ik}^2}, \tag{16}$$

$$\boldsymbol{\Sigma}_k^{(t+1)} = \frac{\sum_{i=1}^n (\boldsymbol{x}_i - \boldsymbol{\mu}_k^{(t+1)})(\boldsymbol{x}_i - \boldsymbol{\mu}_k^{(t+1)})' p_{ik}^2}{\sum_{i=1}^n p_{ik}^2}. \tag{17}$$

Our iterative procedure for Gaussian mixture model-based clustering parameter estimation can be summarized as follows:

Algorithm 1 GPDC

```
1:  procedure GPDC(X, K)
2:      for k = 1, ..., K do                          ▷ Initialization
3:          Random initialization of μ_k
4:          set Σ_k as identity matrix
5:      while μ_k^(t) ≠ μ_k^(t+1) do                  ▷ Core
6:          for k = 1, ..., K do
7:              update q_k according to (8)
8:              update p_ik according to (7)
9:              update μ_k according to (16)
10:             update Σ_k according to (17)
11:     return p_ik, μ_k, Σ_k
```

Generalization to a Multivariate Student-*t* Distribution

The same procedure can be generalized to any symmetric distribution. In this subsection we use the multivariate Student-*t* distribution, generating an algorithm identified as Student-*t* PD-Clustering (TPDC). TPDC can detect clusters characterized by heavy tails; furthermore, the Student-*t* distribution has been often used on datasets characterized by outliers [17]. Now, replace (10) with a multivariate Student-*t* distribution, i.e.,

$$f(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, v) = \frac{\Gamma\left(\frac{v+J}{2}\right)|\boldsymbol{\Sigma}|^{-\frac{1}{2}}}{(\pi v)^{\frac{1}{2}J}\,\Gamma\left(\frac{v}{2}\right)\left\{1 + \frac{\delta(\boldsymbol{x},\boldsymbol{\mu},\boldsymbol{\Sigma})}{v}\right\}^{\frac{1}{2}(v+J)}}, \tag{18}$$

where $\delta(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = (\boldsymbol{x} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu})$, and proceed as in Sect. 3.1 Then, the JDF becomes:

$$\begin{aligned}
\mathrm{JDF} = &\sum_{i=1}^n \sum_{k=1}^K \frac{p_{ik}^2}{q_k} \log(M_k) \\
&+ \sum_{i=1}^n \sum_{k=1}^K \frac{p_{ik}^2}{q_k} \left[-\log\left\{\Gamma\left(\frac{v_k+J}{2}\right)|\boldsymbol{\Sigma}_k|^{-\frac{1}{2}}\right\}\right] \\
&+ \sum_{i=1}^n \sum_{k=1}^K \frac{p_{ik}^2}{q_k} \log\left\{(\pi v_k)^{\frac{J}{2}}\,\Gamma\left(\frac{v_k}{2}\right)\left(1 + \frac{\delta\left(\boldsymbol{x}_i, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right)}{v_k}\right)^{\frac{v_k+J}{2}}\right\}.
\end{aligned} \tag{19}$$

The parameters that optimize (19) can be found by differentiating with respect to $\boldsymbol{\mu}_k$, $\boldsymbol{\Sigma}_k$, and v_k, respectively. Specifically, at a generic iteration $(t+1)$, the parameters that minimize (19) are:

$$\boldsymbol{\mu}_k^{(t+1)} = \frac{\sum_{i=1}^n w_{ik} \boldsymbol{x}_i}{\sum_{i=1}^n w_{ik}}, \tag{20}$$

with $w_{ik} = p_{ik}^2 / [v_k^{(t)} + \delta(\boldsymbol{x}_i, \boldsymbol{\mu}_k^{(t)}, \boldsymbol{\Sigma}_k^{(t)})]$,

$$\boldsymbol{\Sigma}_k^{(t+1)} = \frac{\sum_{i=1}^n p_{ik}^2 (\boldsymbol{x}_i - \boldsymbol{\mu}_k^{(t+1)})(\boldsymbol{x}_i - \boldsymbol{\mu}_k^{(t+1)})' s_{ik}}{\sum_{i=1}^n p_{ik}^2}, \tag{21}$$

with $s_{ik} = (v_k^{(t)} + J)/[v_k^{(t)} + \delta(\boldsymbol{x}_i, \boldsymbol{\mu}_k^{(t+1)}, \boldsymbol{\Sigma}_k^{(t)})]$, and the degrees of freedom update $v_k^{(t+1)}$ is the solution to the following equation:

$$\begin{aligned}
&\sum_{i=1}^n p_{ik}^2 \left[\psi\left(\frac{v_k}{2}\right) - \psi\left(\frac{v_k + J}{2}\right) + \frac{J}{2v_k}\right] \\
&+ \sum_{i=1}^n p_{ik}^2 \left[\frac{1}{2} \log\left(1 + \frac{\delta\left(\boldsymbol{x}_i, \boldsymbol{\mu}_k^{(t+1)}, \boldsymbol{\Sigma}_k^{(t+1)}\right)}{v_k^{(t)}}\right)\right] \\
&- \frac{1}{2} \frac{v_k + J}{v_k} \sum_{i=1}^n p_{ik}^2 \frac{\delta\left(\boldsymbol{x}_i, \boldsymbol{\mu}_k^{(t+1)}, \boldsymbol{\Sigma}_k^{(t+1)}\right)}{v_k^{(t)} + \delta\left(\boldsymbol{x}_i, \boldsymbol{\mu}_k^{(t+1)}, \boldsymbol{\Sigma}_k^{(t+1)}\right)} = 0,
\end{aligned} \tag{22}$$

where

$$\psi(v) = \left(\frac{1}{\Gamma(v)}\right) \frac{\delta \Gamma(v)}{\delta v}.$$

Our iterative algorithm can be summarized as follows:

Algorithm 2 TPDC

```
1:  procedure TPDC(X, K)
2:      for k = 1, ..., K do                          ▷ Initialization
3:          Random initialization of μ_k
4:          set Σ_k as identity matrix
5:          v_k = 20
6:      while μ_k^(t) ≠ μ_k^(t+1) do                  ▷ Core
7:          for k = 1, ..., K do
8:              update q_k according to (8)
9:              update p_ik according to (7)
10:             update μ_k according to (20)
11:             update Σ_k according to (21)
12:             update v_k solving (22)
13:     return p_ik, μ_k, Σ_k, v_k
```

Algorithm Details

All the proposed techniques require a random initialization. Random starts can lead to unstable solutions, to avoid this

Journal : Large 42979	Article No : 67	Pages : 22	MS Code : 67	Dispatch : 21-1-2020

problem the algorithms use multiple starts. Moreover, the functions include the option to use PD-clustering or partition around medoids (PAM; [15]) to start. As for many other clustering techniques, the optimized function, the JDF in (4), is not convex—not even quasi-convex—and may have other stationary points. For a fixed value of $\boldsymbol{\Sigma}_k$, the JDF is a monotonically decreasing function, this guarantees that the function converges to a minimum, not necessarily a global minimum. The proposed techniques, GPDC and TPDC, introduce the estimate of $\boldsymbol{\Sigma}_k$, giving much more flexibility, albeit the JDF is no longer monotonically decreasing. Using (9) in (4), we obtain

$$\mathrm{JDF} = \sum_{i=1}^{n}\sum_{k=1}^{K} p_{ik}^2(\log M_k - \log \phi(\boldsymbol{x}_i;\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k))$$

with $M_k \geq \phi(\boldsymbol{x}_i;\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$. Therefore, for every $k = 1, \dots, K$, the function is upper-bounded for non-degenerate density functions. The convergence of the algorithm cannot depend on the JDF but is based on $\boldsymbol{\mu}_k$. The time complexity of the algorithm is comparable to the EM algorithm, both algorithms require the inversion and the determinant of a $J \times J$ matrix, therefore, the time complexity is of $O(n^3JK)$, where n is the number of observations, J the number of variables, and K the number of clusters.

Empirical Evidence from Simulated and Real Data

The proposed algorithm has been evaluated on real and simulated datasets. The simulated datasets have been used to illustrate the ability of the algorithms to recover the parameters of the distributions and to compare the new techniques with some existing methods. In the following sessions we used the software R [26], the functions for both GPDC and TPDC are included in the R package `FPDclustering` [37].

Simulation Study

The same design was used twice, the first time each cluster was generated from a multivariate Gaussian distribution with three variables and $K = 3$ clusters. The second time, using a multivariate Student-t distribution with five degrees of freedom, same number of variables and clusters. We set the parameter using a four factor full factorial design. There are two factors per each level, where the levels are

- Overlapping and not overlapping clusters
- Different number of elements per clusters
- Unitary variance and variance bigger than 1
- Uncorrelated and correlated variables

Table 1 Model parameters used to generate the simulated datasets

Not overlapping	Overlapping
$\boldsymbol{\mu}_1 = (0, 0, 0)'$	$\boldsymbol{\mu}_1 = (0, 0, 0)'$
$\boldsymbol{\mu}_2 = (-7, 7, 0)'$	$\boldsymbol{\mu}_2 = (-4, 4, 0)'$
$\boldsymbol{\mu}_3 = (-7, 0, 7)'$	$\boldsymbol{\mu}_3 = (-4, 0, 4)'$
Option 1	Option 2
$n_1 = 200$	$n_1 = 200$
$n_2 = 300$	$n_2 = 100$
$n_3 = 100$	$n_3 = 300$
Unitary variance	bigger than 1
$\mathrm{diag}(\boldsymbol{\Sigma}_1) = 1$	$\mathrm{diag}(\boldsymbol{\Sigma}_1) = 1$
$\mathrm{diag}(\boldsymbol{\Sigma}_2) = 1$	$\mathrm{diag}(\boldsymbol{\Sigma}_2) = 16$
$\mathrm{diag}(\boldsymbol{\Sigma}_3) = 1$	$\mathrm{diag}(\boldsymbol{\Sigma}_3) = 2.25$
Not correlated	Correlated
$\boldsymbol{\Sigma}_1 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$	$\boldsymbol{\Sigma}_1 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$
$\boldsymbol{\Sigma}_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$	$\boldsymbol{\Sigma}_2 = \begin{pmatrix} 1 & -0.5 & -0.5 \\ -0.5 & 1 & 0.5 \\ -0.5 & 0.5 & 1 \end{pmatrix}$
$\boldsymbol{\Sigma}_3 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$	$\boldsymbol{\Sigma}_3 = \begin{pmatrix} 1 & 0.7 & 0.7 \\ 0.7 & 1 & 0.7 \\ 0.7 & 0.7 & 1 \end{pmatrix}$

Table 1 shows the parameters used in the simulation study.

The datasets have been generated using the R package `mvtnorm` [11]. Tables 5, 6, 7, 8, 9, 10, 11, 12 in Appendix B.2 show the true and the average estimated values of the parameters obtained from 50 runs of the GPDC and TPDC algorithms. For sake of space, comments are limited to groups of scenarios. The factors that affect the estimates the most are the change in variances and the amount of overlap. Specifically, when data are simulated using multivariate Gaussian distributions, in cases 5–8 and 13–16, the variances are not homogeneous and the GPDC tends to underestimate the bigger variances and overestimate the smaller ones. The TPDC is less affected by this issue, i.e., it underestimates some of the variances but the degrees of freedom recover; however, in the two extreme scenarios, 8 and 16, it cannot recover the cluster structures. Similar outcomes occur when data are simulated using a multivariate Student-t distribution; moreover, as expected on those datasets, the GPDC tends to overestimate the variances and TPDC tends to underestimate the variances and compensate with the degrees of freedom.

On the same datasets we used the functions `gpcm`, option `VVV`, of the R package `mixture` [6] for the Gaussian mixture models (GMM) and the function `teigen`, option `UUUU`, of the homonymous R package [1] for the mixtures of multivariate Student-t distributions (TMM). The k-means

Journal : Large 42979	Article No : 67	Pages : 22	MS Code : 67	Dispatch : 21-1-2020

Table 2 Average ARI and standard deviation on 50 datasets per scenario

	Correlated	Unitary Variance	Over-lapping	n_k	GPDC		TPDC		PDQ		k-means		GMM		TMM	
					ARI	SD	ARI	SD	ARI	SD	ARI	SD	ARI	SD	ARI	SD
Multivariate Gaussian distribution																
1	No	Yes	No	Op 1	1.00	0.00	1.00	0.00	1.00	0.00	0.91	0.20	0.99	0.06	1.00	0.00
2	No	Yes	no	Op 2	1.00	0.00	1.00	0.00	1.00	0.00	0.92	0.18	0.96	0.12	1.00	0.00
3	No	Yes	Yes	Op 1	0.97	0.01	0.98	0.01	0.98	0.01	0.99	0.01	0.99	0.01	0.99	0.01
4	No	Yes	Yes	Op 2	0.97	0.01	0.98	0.01	0.98	0.01	0.98	0.01	0.99	0.01	0.99	0.01
5	No	No	No	Op 1	0.80	0.03	0.84	0.03	0.74	0.04	0.61	0.08	0.93	0.07	0.93	0.02
6	No	No	No	Op 2	0.94	0.02	0.93	0.02	0.74	0.05	0.89	0.09	0.97	0.01	0.97	0.01
7	No	No	Yes	Op 1	0.59	0.04	0.60	0.05	0.30	0.04	0.35	0.06	0.76	0.08	0.76	0.04
8	No	No	Yes	Op 2	0.56	0.12	0.50	0.09	0.60	0.06	0.75	0.08	0.87	0.02	0.87	0.02
9	Yes	Yes	No	Op 1	1.00	0.00	1.00	0.00	1.00	0.00	0.86	0.22	0.95	0.14	1.00	0.00
10	Yes	Yes	No	Op 2	1.00	0.00	1.00	0.00	1.00	0.00	0.91	0.19	0.91	0.18	1.00	0.00
11	Yes	Yes	Yes	Op 1	0.97	0.02	0.98	0.02	0.98	0.01	0.93	0.15	0.96	0.09	0.99	0.01
12	Yes	yes	Yes	Op 2	0.98	0.02	0.99	0.01	0.98	0.01	0.98	0.07	0.98	0.08	1.00	0.00
13	Yes	No	No	Op 1	0.76	0.08	0.90	0.03	0.67	0.05	0.52	0.08	0.95	0.07	0.96	0.01
14	Yes	No	No	Op 2	0.97	0.02	0.96	0.02	0.77	0.06	0.91	0.02	0.99	0.01	0.99	0.01
15	Yes	No	Yes	Op 1	0.44	0.08	0.45	0.08	0.33	0.05	0.32	0.04	0.82	0.09	0.83	0.03
16	Yes	No	Yes	Op 2	0.54	0.07	0.45	0.08	0.54	0.08	0.80	0.10	0.94	0.01	0.94	0.01
Multivariate Student-t distribution, 5 df																
1	no	Yes	No	Op 1	0.98	0.01	0.99	0.01	0.99	0.01	0.98	0.07	0.97	0.06	0.99	0.01
2	No	Yes	No	Op 2	0.99	0.01	0.99	0.01	0.99	0.01	0.93	0.17	0.95	0.11	0.99	0.01
3	No	Yes	Yes	Op 1	0.88	0.02	0.89	0.02	0.91	0.02	0.90	0.02	0.90	0.02	0.91	0.02
4	No	Yes	Yes	Op 2	0.88	0.03	0.89	0.02	0.90	0.02	0.89	0.07	0.89	0.05	0.90	0.02
5	No	No	No	Op 1	0.92	0.02	0.94	0.02	0.94	0.02	0.89	0.07	0.93	0.07	0.95	0.01
6	No	No	No	Op 2	0.96	0.02	0.96	0.01	0.93	0.02	0.96	0.01	0.91	0.06	0.96	0.01
7	No	No	Yes	Op 1	0.69	0.04	0.71	0.04	0.70	0.04	0.63	0.04	0.56	0.11	0.76	0.03
8	No	No	Yes	Op 2	0.76	0.03	0.76	0.03	0.62	0.05	0.81	0.03	0.66	0.06	0.78	0.05
9	Yes	Yes	No	Op 1	0.99	0.01	0.99	0.01	0.99	0.01	0.90	0.19	0.92	0.17	0.99	0.01
10	Yes	Yes	No	Op 2	0.99	0.01	0.99	0.01	0.99	0.01	0.87	0.22	0.93	0.14	0.99	0.01
11	Yes	Yes	Yes	Op 1	0.87	0.03	0.88	0.03	0.90	0.02	0.83	0.16	0.90	0.07	0.93	0.02
12	Yes	Yes	Yes	Op 2	0.87	0.05	0.89	0.04	0.93	0.02	0.92	0.11	0.94	0.06	0.96	0.01
13	Yes	No	No	Op 1	0.93	0.02	0.95	0.02	0.92	0.02	0.79	0.17	0.93	0.10	0.96	0.01
14	Yes	No	No	Op 2	0.97	0.01	0.98	0.01	0.95	0.02	0.97	0.01	0.96	0.03	0.98	0.01
15	Yes	No	Yes	Op 1	0.56	0.10	0.59	0.08	0.65	0.04	0.55	0.08	0.72	0.13	0.82	0.03
16	Yes	No	Yes	Op 2	0.49	0.29	0.51	0.32	0.56	0.31	0.67	0.37	0.64	0.35	0.71	0.38

algorithm is part of the `stats` package [27], and the `PDQ-clust` function for PDQ clustering is part of the `FPDclustering` package [37].

To compare the clustering performance of the methods we used the adjusted Rand index (ARI) [12]. It compares predicted classifications with true classes. The ARI corrects the Rand index [30] for chance, its expected value under random classification is 0, and it takes a value of 1 when there is perfect class agreement. Steinley [31] gives guidelines for interpreting ARI values. Table 2 shows the average ARI and the standard deviation on 50 runs for each algorithm.

As pointed out in the previous sections, GPDC and TPDC are framed in a non-parametric view; however, to evaluate the performance we compare them with the GMM and TMM. The performance is not expected to be better than those techniques, although in most scenarios GPDC and TPDC perform as well as finite mixture models. As expected, k-means results are impacted by correlations and not homogeneous variances. PDQ cannot recover the correct clustering partition in case of overlapping and not homogeneous variance. It is not affected by changes in group size or correlation. The proposed techniques GPDC and TPDC outperform k-means and PDQ in most scenarios, they show weakness in the two most extreme situations, i.e., scenarios 8 and 16. Specifically, when clusters have different variances and the biggest variance is associated with the smallest

Journal : Large 42979 | Article No : 67 | Pages : 22 | MS Code : 67 | Dispatch : 21-1-2020

Table 3 Number of units, variables, and clusters for the three real datasets

	n	*J*	*K*	
Seed	210	4	3	UCI machine learning repository
HSCT	9702	3	4	Terry Fox Lab
AIS	202	4	2	EMMIXuskew R package

Table 4 Adjusted Rand index for the real datasets

	GPDC	TPDC	PDQ	k-means	GMM	TMM
Seed	0.41	0.33	0.17	0.16	0.44	0.53
HSCT	0.99	0.98	0.98	0.72	0.88	0.99
AIS	0.88	0.90	0.16	0.06	0.72	0.81

cluster, they fail detecting the clustering partitions. Figures 5, 6, 7, 8, 9, 10, 11, 12 in Appendix B.2 show examples of simulated datasets for each scenario.

Real Data Analysis

We performed a real data analysis on three datasets that differ in size and number of clusters (details in Table 3). We performed variable selection prior to cluster analysis (details in Appendix B.1). The seed dataset[1] contains information about kernels belonging to three different varieties of wheat—Kama, Rosa and Canadian—with 70 observations per variety (see Fig. 1). We used the variables: compactness, length of kernel, width of kernel, and asymmetry coefficient. The hematopoietic steam cell transplant (HSCT) data were collected in the Terry Fox Lab at the British Columbia Cancer Agency. It contains information about 9780 cells, each stained with four fluorescent dyes. Experts identified four clusters; moreover, 78 cells were deemed “dead”, leaving a total of 9702 observation, we selected the three most informative variables. Figure 2 shows the partitions defined by the experts. The Australian Institute of sport dataset[2]. contains data on 102 male and 100 female athletes for the Australian institute of sports. We selected the variables: height in cm, hematocrit, plasma ferritin concentration, and percent body fat, see Fig. 3.

Table 4 shows the ARI on the three datasets. On the seed dataset, GPDC and TPDC perform better than PDQ and k-means. The improvement from PDQ is noticeable, PDQ gives an ARI of 0.17, while GPDC gives an ARI of 0.41. On this dataset, TMM gives the best performance. On the HSCT dataset, GPDC, TPDC, PDQ, and TMM have a very high ARI. On the AIS dataset, GPDC and TPDC give the best performance.

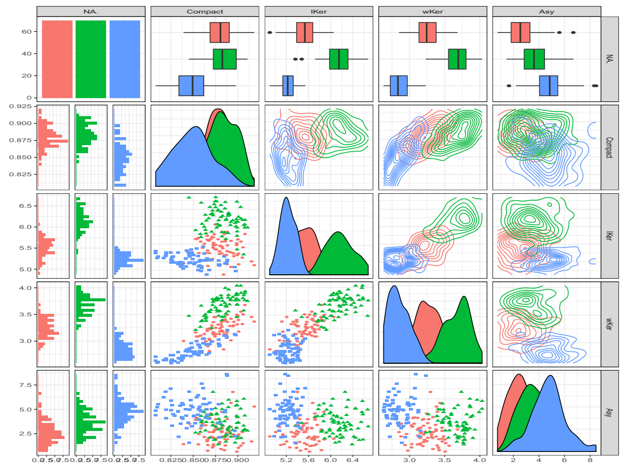

Fig. 1 Seed dataset, each color and symbol representing a different variety of wheat

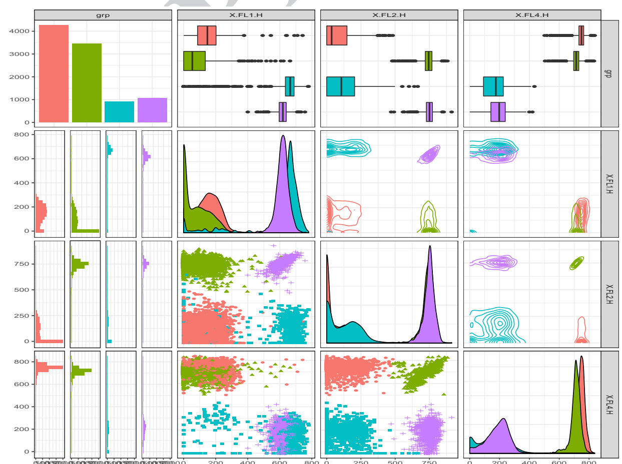

Fig. 2 HSCT dataset, each color and symbol representing a partition defined by the experts

Conclusion

A new distance measure based on density functions is introduced and used in the context of probabilistic distance clustering adjusted for cluster size (PDQ). PDQ assumes that, for a generic unit, the product between the probability of belonging to a cluster and its distance from the cluster is constant. The minimization of the sum of these constants over the units leads to clusters that maximize the classifiability of the data. We introduce two algorithms based on PDQ that use distance measures based on the multivariate

[1] http://archive.ics.uci.edu/ml/.

[2] GLMsData R package.

Journal : Large 42979 | Article No : 67 | Pages : 22 | MS Code : 67 | Dispatch : 21-1-2020

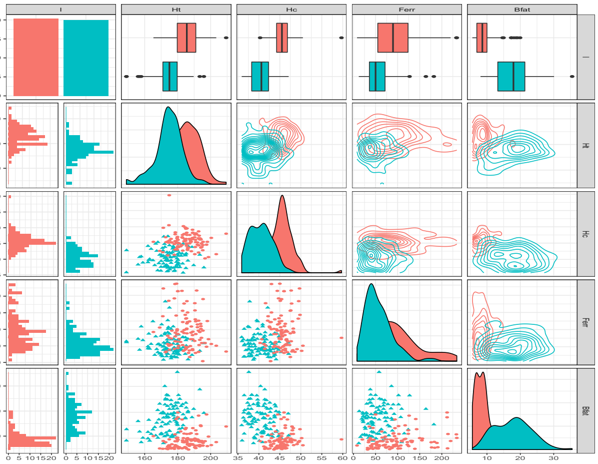

Fig. 3 AIS dataset, each color and symbol representing male and female athletes

Gaussian distribution and on the multivariate Student-*t* distribution. Using simulated and real datasets we show how the new algorithms over-perform PDQ and the well known k-means algorithm.

The algorithm could be extended using different distributions. Further to this point, we mentioned outliers as a possible motivation for the PDQ approach with the multivariate Student-*t* distribution (Sect. 3). However, if the objective is dealing with outliers, it will be better to consider the PDQ approach with the multivariate contaminated normal distribution [25] and this will be a topic of future work. Other approaches for handling cluster concentration will also be considered (e.g., [9]) as will methods that accommodate asymmetric, or skewed, clusters (e.g., [18, 19, 21, 22, 32, 34]).

Funding this study was funded by a Discovery Grant from the Natural Sciences and Engineering Research Council of Canada and the Canada Research Chairs program (McNicholas). During the development of the present work, Prof. Francesco Palumbo had a short term visiting at the San Jose State University (CA) financially supported by the International short mobility program with foreign universities and research centers of the Università degli Studi di Napoli Federico II (DR 2243).

Compliance with Ethical Standards

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

Dissimilarity Measure

A general measure $d(\boldsymbol{x}, \mathbf{y})$ is a dissimilarity measure if the following conditions are verified [33, p.404]:

1. $\mathrm{d}(\boldsymbol{x}, \mathbf{y}) \geq 0$
2. $\mathrm{d}(\boldsymbol{x}, \mathbf{y}) = 0 \Leftrightarrow \boldsymbol{x} = \mathbf{y}$
3. $\mathrm{d}(\boldsymbol{x}, \mathbf{y}) = d(\boldsymbol{x}, \mathbf{y})$.

Journal : **Large 42979**	Article No : **67**	Pages : **22**	MS Code : **67**	Dispatch : **21-1-2020**

Let $f(\boldsymbol{x}_i;\boldsymbol{\mu}_k,\boldsymbol{\theta}_k)$ be the generic symmetric unimodal multivariate density function of the random variable $\mathbf{X}$ with parameter $\boldsymbol{\theta}_k$ and location parameter $\boldsymbol{\mu}_k$ then

$$\mathrm{d}(\boldsymbol{x}_i,\boldsymbol{\mu}_k) = \log\left(\frac{M_k}{f(\boldsymbol{x}_i;\boldsymbol{\mu}_k,\boldsymbol{\theta}_k)}\right), \tag{23}$$

satisfies all the three properties and it is a dissimilarity measure for $k = 1,\dots,K$.

1. $\mathrm{d}(\boldsymbol{x}_i,\boldsymbol{\mu}_k) > 0,\ \forall \boldsymbol{x}_i$.
Proof

$$0 < \frac{f(\boldsymbol{x}_i;\boldsymbol{\mu}_k,\boldsymbol{\theta}_k)}{M_k} \le 1 \Rightarrow \frac{M_k}{f(\boldsymbol{x}_i;\boldsymbol{\mu}_k,\boldsymbol{\theta}_k)} \ge 1 \Rightarrow \\ \Rightarrow \log\left(\frac{M_k}{f(\boldsymbol{x}_i;\boldsymbol{\mu}_k,\boldsymbol{\theta}_k)}\right) \ge 0.$$

2. $\mathrm{d}(\boldsymbol{x}_i,\boldsymbol{\mu}_k) = 0 \Leftrightarrow \boldsymbol{x}_i = \boldsymbol{\mu}_k$.
2a. $\boldsymbol{x}_i = \boldsymbol{\mu}_k \Rightarrow d(\boldsymbol{x}_i,\boldsymbol{\mu}_k) = 0\ \forall \boldsymbol{x}_i$. Proof

$$\boldsymbol{x}_i = \boldsymbol{\mu}_k \Rightarrow f(\boldsymbol{x}_i;\boldsymbol{\mu}_k,\boldsymbol{\theta}_k) = f(\boldsymbol{\mu}_k;\boldsymbol{\mu}_k,\boldsymbol{\theta}_k) = M_k \Rightarrow \\ \Rightarrow \frac{M_k}{M_k} = 1 \Rightarrow \log(1) = 0,$$

2b. $\mathrm{d}(\boldsymbol{x}_i,\boldsymbol{\mu}_k) = 0 \Rightarrow \boldsymbol{x}_i = \boldsymbol{\mu}_k,\ \forall \boldsymbol{x}_i$.
Proof

$$\log\left(\frac{M_k}{f(\boldsymbol{x}_i;\boldsymbol{\mu}_k,\boldsymbol{\theta}_k)}\right) = 0 \Rightarrow \frac{M_k}{f(\boldsymbol{x}_i;\boldsymbol{\mu}_k,\boldsymbol{\theta}_k)} = 1 \Rightarrow \\ \Rightarrow f(\boldsymbol{x}_i;\boldsymbol{\mu}_k,\boldsymbol{\theta}_k) = M_k \\ = f(\boldsymbol{\mu}_k;\boldsymbol{\mu}_k,\boldsymbol{\theta}_k) \Rightarrow \boldsymbol{x}_i = \boldsymbol{\mu}_k.$$

3. $\mathrm{d}(\boldsymbol{x}_i,\boldsymbol{\mu}_k) = d(\boldsymbol{\mu}_k,\boldsymbol{x}_i),\ \forall \boldsymbol{x}_i$ Proof Given $\boldsymbol{\theta}_k$

$$f(\boldsymbol{x}_i;\boldsymbol{\mu}_k,\boldsymbol{\theta}_k) = f(\boldsymbol{\mu}_k;\boldsymbol{x}_i,\boldsymbol{\theta}_k), \Rightarrow \log\left(\frac{M_k}{f(\boldsymbol{x}_i;\boldsymbol{\mu}_k,\boldsymbol{\theta}_k)}\right) = \log\left(\frac{M_k}{f(\boldsymbol{\mu}_k;\boldsymbol{x}_i,\boldsymbol{\theta}_k)}\right)$$

□

Addition Details for Data Analyses

Variable Selection

On each dataset we selected one variable per group using AQ1 hierarchical clustering (Fig. 4).

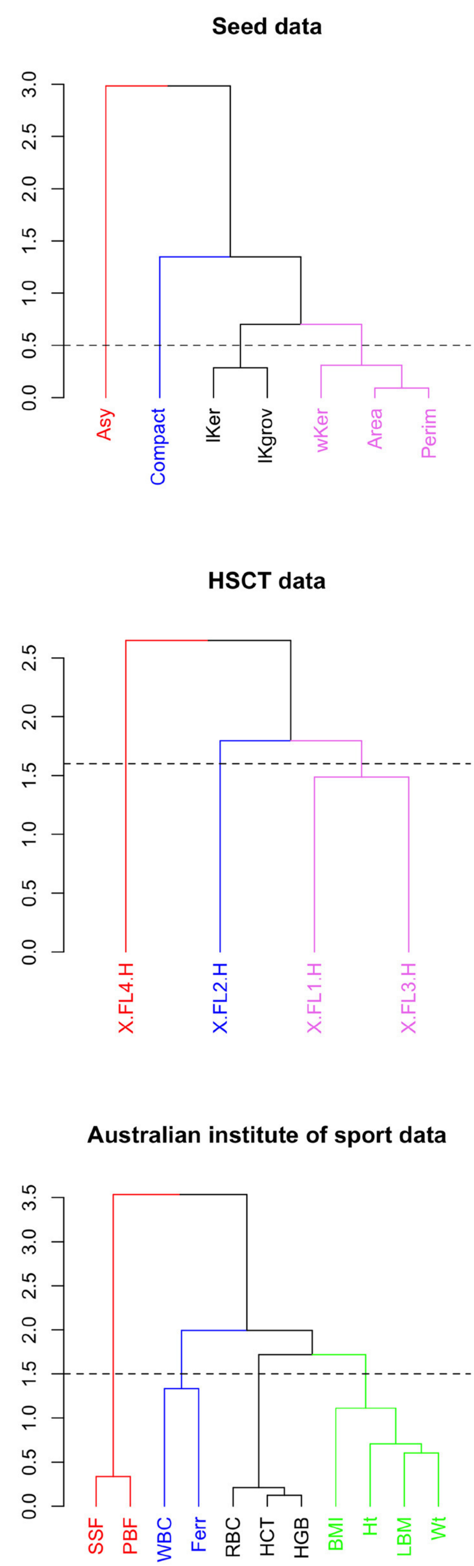

Fig. 4 Variable selection

Journal : **Large 42979**	Article No : **67**	Pages : **22**	MS Code : **67**	Dispatch : **21-1-2020**

Simulated Datasets

Table 5 Simulated datasets using multivariate Gaussian distributions Scenarios 1-4

	True			GPDC			TPDC		
$\boldsymbol{\mu}_1$	0.00	0.00	0.00	−0.02	0.03	−0.01	0.01	−0.00	−0.01
$\boldsymbol{\mu}_2$	−7.00	7.00	0.00	−7.00	7.00	0.01	−7.01	7.01	0.01
$\boldsymbol{\mu}_3$	−7.00	0.00	7.00	−6.98	0.05	6.97	−6.99	0.02	7.02
$\boldsymbol{\Sigma}_1$	1.00	0.00	0.00	1.29	−0.29	−0.02	0.97	−0.06	0.01
	0.00	1.00	0.00	−0.29	1.23	0.00	−0.06	0.95	0.00
	0.00	0.00	1.00	−0.02	0.00	0.93	0.01	0.00	0.84
$\boldsymbol{\Sigma}_2$	1.00	0.00	0.00	1.04	−0.11	0.00	0.90	−0.02	−0.00
	0.00	1.00	0.00	−0.11	1.09	−0.04	−0.02	0.93	−0.01
	0.00	0.00	1.00	0.00	−0.04	0.90	−0.00	−0.01	0.84
$\boldsymbol{\Sigma}_3$	1.00	0.00	0.00	1.04	−0.02	−0.11	0.87	−0.01	−0.01
	0.00	1.00	0.00	−0.02	1.18	−0.28	−0.01	0.89	−0.06
	0.00	0.00	1.00	−0.11	−0.28	1.35	−0.01	−0.06	0.95
$\boldsymbol{\mu}_1$	0.00	0.00	0.00	−0.02	−0.00	0.03	0.02	−0.00	−0.01
$\boldsymbol{\mu}_2$	−7.00	7.00	0.00	−6.99	6.94	0.07	−7.00	6.99	0.04
$\boldsymbol{\mu}_3$	−7.00	0.00	7.00	−6.99	0.01	7.00	−7.00	0.01	7.01
$\boldsymbol{\Sigma}_1$	1.00	0.00	0.00	1.28	−0.03	−0.27	0.97	−0.00	−0.04
	0.00	1.00	0.00	−0.03	0.91	0.00	−0.00	0.82	0.00
	0.00	0.00	1.00	−0.27	0.00	1.25	−0.04	0.00	0.97
$\boldsymbol{\Sigma}_2$	1.00	0.00	0.00	1.03	−0.12	−0.03	0.85	−0.02	−0.02
	0.00	1.00	0.00	−0.12	1.36	−0.30	−0.02	0.95	−0.07
	0.00	0.00	1.00	−0.03	−0.30	1.23	−0.02	−0.07	0.92
$\boldsymbol{\Sigma}_3$	1.00	0.00	0.00	1.03	0.00	−0.12	0.90	0.00	−0.02
	0.00	1.00	0.00	0.00	0.91	−0.03	0.00	0.84	−0.01
	0.00	0.00	1.00	−0.12	−0.03	1.07	−0.02	−0.01	0.90
$\boldsymbol{\mu}_1$	0.00	0.00	0.00	−0.08	0.12	−0.02	0.01	0.03	−0.03
$\boldsymbol{\mu}_2$	−4.00	4.00	0.00	−3.99	4.04	−0.01	−4.02	4.05	−0.01
$\boldsymbol{\mu}_3$	−4.00	0.00	4.00	−3.92	0.22	3.79	−3.95	0.10	3.97
$\boldsymbol{\Sigma}_1$	1.00	0.00	0.00	1.51	−0.52	−0.05	1.00	−0.20	−0.01
	0.00	1.00	0.00	−0.52	1.38	−0.00	−0.20	0.94	0.00
	0.00	0.00	1.00	−0.05	−0.00	0.87	−0.01	0.00	0.68
$\boldsymbol{\Sigma}_2$	1.00	0.00	0.00	1.01	−0.19	−0.00	0.80	−0.07	−0.00
	0.00	1.00	0.00	−0.19	1.13	−0.05	−0.07	0.87	−0.02
	0.00	0.00	1.00	−0.00	−0.05	0.81	−0.00	−0.02	0.69
$\boldsymbol{\Sigma}_3$	1.00	0.00	0.00	1.16	−0.05	−0.26	0.71	−0.02	−0.09
	0.00	1.00	0.00	−0.05	1.50	−0.70	−0.02	0.85	−0.27
	0.00	0.00	1.00	−0.26	−0.70	1.90	−0.09	−0.27	0.99
$\boldsymbol{\mu}_1$	0.00	0.00	0.00	−0.08	−0.01	0.11	0.01	−0.01	0.02
$\boldsymbol{\mu}_2$	−4.00	4.00	0.00	−3.93	3.77	0.24	−3.97	3.94	0.13
$\boldsymbol{\mu}_3$	−4.00	0.00	4.00	−3.98	−0.01	4.03	−4.01	−0.01	4.05
$\boldsymbol{\Sigma}_1$	1.00	0.00	0.00	1.51	−0.06	−0.51	0.99	−0.02	−0.19
	0.00	1.00	0.00	−0.06	0.86	−0.01	−0.02	0.67	−0.01
	0.00	0.00	1.00	−0.51	−0.01	1.40	−0.19	−0.01	0.94
$\boldsymbol{\Sigma}_2$	1.00	0.00	0.00	1.14	−0.28	−0.05	0.70	−0.11	−0.02
	0.00	1.00	0.00	−0.28	1.91	−0.72	−0.11	1.01	−0.29
	0.00	0.00	1.00	−0.05	−0.72	1.54	−0.02	−0.29	0.88
$\boldsymbol{\Sigma}_3$	1.00	0.00	0.00	1.02	−0.00	−0.20	0.78	−0.00	−0.07
	0.00	1.00	0.00	−0.00	0.82	−0.05	−0.00	0.67	−0.02
	0.00	0.00	1.00	−0.20	−0.05	1.12	−0.07	−0.02	0.82

Journal : Large 42979	Article No : 67	Pages : 22	MS Code : 67	Dispatch : 21-1-2020

Table 6 Simulated datasets using multivariate Gaussian distributions Scenarios 5-8

	True			GPDC			TPDC		
$\boldsymbol{\mu}_1$	0.00	0.00	0.00	−0.20	0.23	−0.03	−0.02	0.04	−0.02
$\boldsymbol{\mu}_2$	−7.00	7.00	0.00	−7.33	8.08	−0.46	−7.28	7.85	−0.40
$\boldsymbol{\mu}_3$	−7.00	0.00	7.00	−7.09	0.86	6.33	−7.03	0.40	6.78
$\boldsymbol{\Sigma}_1$	1.00	0.00	0.00	2.68	−0.93	0.09	0.90	−0.09	0.02
	0.00	1.00	0.00	−0.93	2.60	−0.06	−0.09	0.89	−0.00
	0.00	0.00	1.00	0.09	−0.06	1.69	0.02	−0.00	0.74
$\boldsymbol{\Sigma}_2$	16.00	0.00	0.00	12.84	0.95	0.09	11.36	0.46	−0.00
	0.00	16.00	0.00	0.95	13.43	1.24	0.46	12.01	0.80
	0.00	0.00	16.00	0.09	1.24	11.82	−0.00	0.80	10.36
$\boldsymbol{\Sigma}_3$	2.25	0.00	0.00	4.40	−0.23	0.39	2.14	−0.09	0.13
	0.00	2.25	0.00	−0.23	6.37	−2.79	−0.09	2.79	−0.85
	0.00	0.00	2.25	0.39	−2.79	7.28	0.13	−0.85	2.94
$\boldsymbol{\mu}_1$	0.00	0.00	0.00	−0.12	0.06	0.01	−0.00	0.01	−0.02
$\boldsymbol{\mu}_2$	−7.00	7.00	0.00	−7.10	7.23	0.80	−7.07	7.11	0.87
$\boldsymbol{\mu}_3$	−7.00	0.00	7.00	−7.02	0.06	7.00	−7.01	0.01	7.06
$\boldsymbol{\Sigma}_1$	1.00	0.00	0.00	2.09	−0.29	−0.27	0.91	−0.04	−0.02
	0.00	1.00	0.00	−0.29	1.31	−0.11	−0.04	0.73	−0.02
	0.00	0.00	1.00	−0.27	−0.11	1.76	−0.02	−0.02	0.87
$\boldsymbol{\Sigma}_2$	16.00	0.00	0.00	11.06	−0.17	−0.08	9.14	−0.44	−0.06
	0.00	16.00	0.00	−0.17	17.36	−4.90	−0.44	14.91	−4.73
	0.00	0.00	16.00	−0.08	−4.90	14.27	−0.06	−4.73	11.95
$\boldsymbol{\Sigma}_3$	2.25	0.00	0.00	2.42	−0.03	−0.02	1.86	−0.01	−0.02
	0.00	2.25	0.00	−0.03	2.29	−0.44	−0.01	1.70	−0.09
	0.00	0.00	2.25	−0.02	−0.44	3.02	−0.02	−0.09	2.01
$\boldsymbol{\mu}_1$	0.00	0.00	0.00	−0.07	0.17	−0.03	0.02	0.04	−0.03
$\boldsymbol{\mu}_2$	−4.00	4.00	0.00	−4.65	5.39	−0.79	−4.55	5.17	−0.65
$\boldsymbol{\mu}_3$	−4.00	0.00	4.00	−4.12	0.66	3.84	−4.07	0.39	4.00
$\boldsymbol{\Sigma}_1$	1.00	0.00	0.00	2.67	−0.07	0.01	1.04	0.01	0.00
	0.00	1.00	0.00	−0.07	2.53	−0.04	0.01	1.02	−0.01
	0.00	0.00	1.00	0.01	−0.04	2.18	0.00	−0.01	0.93
$\boldsymbol{\Sigma}_1$	16.00	0.00	0.00	12.16	1.15	−0.07	9.85	0.67	−0.06
	0.00	16.00	0.00	1.15	13.16	1.16	0.67	11.00	0.70
	0.00	0.00	16.00	−0.07	1.16	11.51	−0.06	0.70	9.30
$\boldsymbol{\Sigma}_3$	2.25	0.00	0.00	5.01	0.07	0.18	2.59	−0.01	0.04
	0.00	2.25	0.00	0.07	5.12	−0.24	−0.01	2.60	−0.05
	0.00	0.00	2.25	0.18	−0.24	6.08	0.04	−0.05	2.92
$\boldsymbol{\mu}_1$	0.00	0.00	0.00	−0.31	0.16	−0.01	−0.18	0.10	0.03
$\boldsymbol{\mu}_2$	−4.00	4.00	0.00	−3.84	2.77	2.35	−1.30	0.33	1.09
$\boldsymbol{\mu}_3$	−4.00	0.00	4.00	−4.18	−0.23	3.82	−4.21	−0.25	3.86
$\boldsymbol{\Sigma}_1$	1.00	0.00	0.00	3.00	−0.40	−0.15	1.75	−0.20	−0.20
	0.00	1.00	0.00	−0.40	2.29	−0.32	−0.20	1.40	−0.11
	0.00	0.00	1.00	−0.15	−0.32	2.80	−0.20	−0.11	1.61
$\boldsymbol{\Sigma}_2$	16.00	0.00	0.00	5.58	−0.02	0.27	1.99	−0.17	−0.12
	0.00	16.00	0.00	−0.02	9.10	−3.30	−0.17	1.88	−0.38
	0.00	0.00	16.00	0.27	−3.30	6.61	−0.12	−0.38	2.05
$\boldsymbol{\Sigma}_3$	2.25	0.00	0.00	2.72	−0.01	0.18	2.01	−0.05	0.13
	0.00	2.25	0.00	−0.01	2.83	−0.61	−0.05	2.04	−0.47
	0.00	0.00	2.25	0.18	−0.61	3.37	0.13	−0.47	2.34

Journal : Large 42979 | Article No : 67 | Pages : 22 | MS Code : 67 | Dispatch : 21-1-2020

Table 7 Simulated datasets using multivariate Gaussian distributions Scenarios 9-12

	True			GPDC			TPDC		
$\boldsymbol{\mu}_1$	0.00	0.00	0.00	−0.04	0.04	0.02	0.01	0.00	0.02
$\boldsymbol{\mu}_2$	−7.00	7.00	0.00	−7.01	7.00	0.03	−7.02	7.01	0.02
$\boldsymbol{\mu}_3$	−7.00	0.00	7.00	−6.98	0.03	7.02	−6.97	0.03	7.03
$\boldsymbol{\Sigma}_1$	1.00	0.00	0.00	1.36	−0.34	−0.01	1.00	−0.07	0.00
	0.00	1.00	0.00	−0.34	1.30	−0.01	−0.07	0.98	−0.00
	0.00	0.00	1.00	−0.01	−0.01	0.96	0.00	−0.00	0.86
$\boldsymbol{\Sigma}_2$	1.00	−0.50	−0.50	1.07	−0.60	−0.47	0.92	−0.47	−0.43
	−0.50	1.00	0.50	−0.60	1.11	0.46	−0.47	0.94	0.43
	−0.50	0.50	1.00	−0.47	0.46	0.91	−0.43	0.43	0.85
$\boldsymbol{\Sigma}_3$	1.00	0.70	0.70	0.93	0.63	0.64	0.86	0.59	0.61
	0.70	1.00	0.70	0.63	0.93	0.62	0.59	0.85	0.60
	0.70	0.70	1.00	0.64	0.62	0.98	0.61	0.60	0.89
$\boldsymbol{\mu}_1$	0.00	0.00	0.00	−0.04	−0.00	0.05	0.00	−0.00	0.01
$\boldsymbol{\mu}_2$	−7.00	7.00	0.00	−7.00	6.97	0.05	−7.02	7.00	0.04
$\boldsymbol{\mu}_3$	−7.00	0.00	7.00	−6.99	0.02	7.01	−6.99	0.02	7.01
$\boldsymbol{\Sigma}_1$	1.00	0.00	0.00	1.40	−0.01	−0.35	1.01	−0.00	−0.06
	0.00	1.00	0.00	−0.01	0.97	0.01	−0.00	0.88	0.00
	0.00	0.00	1.00	−0.35	0.01	1.36	−0.06	0.00	1.01
$\boldsymbol{\Sigma}_2$	1.00	−0.50	−0.50	1.06	−0.59	−0.50	0.90	−0.46	−0.46
	−0.50	1.00	0.50	−0.59	1.19	0.39	−0.46	0.93	0.43
	−0.50	0.50	1.00	−0.50	0.39	1.06	−0.46	0.43	0.91
$\boldsymbol{\Sigma}_3$	1.00	0.70	0.70	0.94	0.63	0.64	0.88	0.60	0.61
	0.70	1.00	0.70	0.63	0.88	0.63	0.60	0.84	0.60
	0.70	0.70	1.00	0.64	0.63	0.94	0.61	0.60	0.88
$\boldsymbol{\mu}_1$	0.00	0.00	0.00	−0.19	0.22	0.04	−0.09	0.11	0.05
$\boldsymbol{\mu}_2$	−4.00	4.00	0.00	−4.05	4.04	0.08	−4.06	4.06	0.07
$\boldsymbol{\mu}_3$	−4.00	0.00	4.00	−3.95	0.05	4.04	−3.94	0.05	4.08
$\boldsymbol{\Sigma}_1$	1.00	0.00	0.00	1.78	−0.76	−0.00	1.32	−0.41	−0.01
	0.00	1.00	0.00	−0.76	1.62	−0.08	−0.41	1.20	−0.05
	0.00	0.00	1.00	−0.00	−0.08	0.97	−0.01	−0.05	0.83
$\boldsymbol{\Sigma}_2$	1.00	−0.50	−0.50	1.05	−0.64	−0.44	0.84	−0.47	−0.38
	−0.50	1.00	0.50	−0.64	1.17	0.43	−0.47	0.91	0.37
	−0.50	0.50	1.00	−0.44	0.43	0.83	−0.38	0.37	0.72
$\boldsymbol{\Sigma}_3$	1.00	0.70	0.70	0.86	0.57	0.57	0.57	0.40	0.42
	0.70	1.00	0.70	0.57	0.90	0.55	0.40	0.59	0.42
	0.70	0.70	1.00	0.57	0.55	0.99	0.42	0.42	0.65
$\boldsymbol{\mu}_1$	0.00	0.00	0.00	−0.17	−0.00	0.24	−0.08	0.01	0.12
$\boldsymbol{\mu}_2$	−4.00	4.00	0.00	−3.94	3.85	0.12	−4.03	4.02	0.08
$\boldsymbol{\mu}_3$	−4.00	0.00	4.00	−3.99	0.03	4.05	−3.99	0.03	4.04
$\boldsymbol{\Sigma}_1$	1.00	0.00	0.00	1.89	0.04	−0.79	1.34	−0.02	−0.40
	0.00	1.00	0.00	0.04	0.97	0.02	−0.02	0.85	0.00
	0.00	0.00	1.00	−0.79	0.02	1.70	−0.40	0.00	1.25
$\boldsymbol{\Sigma}_2$	1.00	−0.50	−0.50	1.19	−0.79	−0.46	0.68	−0.40	−0.32
	−0.50	1.00	0.50	−0.79	1.66	0.22	−0.40	0.81	0.28
	−0.50	0.50	1.00	−0.46	0.22	1.23	−0.32	0.28	0.70
$\boldsymbol{\Sigma}_3$	1.00	0.70	0.70	0.88	0.59	0.62	0.78	0.53	0.56
	0.70	1.00	0.70	0.59	0.77	0.59	0.53	0.70	0.53
	0.70	0.70	1.00	0.62	0.59	0.91	0.56	0.53	0.81

Journal : Large 42979	Article No : 67	Pages : 22	MS Code : 67	Dispatch : 21-1-2020

Table 8 Simulated datasets using multivariate Gaussian distributions Scenarios 13–16

	True			GPDC			TPDC		
$\boldsymbol{\mu}_1$	0.00	0.00	0.00	−0.44	0.48	−0.24	−0.09	0.11	−0.06
$\boldsymbol{\mu}_2$	−7.00	7.00	0.00	−7.94	8.25	0.64	−7.73	7.75	0.58
$\boldsymbol{\mu}_3$	−7.00	0.00	7.00	−6.98	0.63	6.52	−6.92	0.09	7.09
$\boldsymbol{\Sigma}_1$	1.00	0.00	0.00	4.11	−2.42	0.32	1.24	−0.36	0.10
	0.00	1.00	0.00	−2.42	3.92	−0.39	−0.36	1.20	−0.11
	0.00	0.00	1.00	0.32	−0.39	2.86	0.10	−0.11	1.06
$\boldsymbol{\Sigma}_2$	16.00	−8.00	−8.00	13.01	−6.32	−6.60	12.04	−5.73	−5.83
	−8.00	16.00	8.00	−6.32	13.59	6.70	−5.73	12.79	5.60
	−8.00	8.00	16.00	−6.60	6.70	11.52	−5.83	5.60	11.29
$\boldsymbol{\Sigma}_3$	2.25	1.57	1.57	3.11	0.73	0.83	1.29	0.88	0.93
	1.57	2.25	1.57	0.73	5.46	−0.75	0.88	1.37	0.90
	1.57	1.57	2.25	0.83	−0.75	6.25	0.93	0.90	1.46
$\boldsymbol{\mu}_1$	0.00	0.00	0.00	−0.16	0.10	0.00	−0.01	0.02	−0.01
$\boldsymbol{\mu}_2$	−7.00	7.00	0.00	−7.72	7.58	0.98	−7.61	7.51	0.92
$\boldsymbol{\mu}_3$	−7.00	0.00	7.00	−6.96	0.05	7.05	−6.96	0.05	7.06
$\boldsymbol{\Sigma}_1$	1.00	0.00	0.00	2.40	−0.40	−0.55	0.94	−0.05	−0.04
	0.00	1.00	0.00	−0.40	1.46	−0.08	−0.05	0.76	−0.03
	0.00	0.00	1.00	−0.55	−0.08	2.40	−0.04	−0.03	0.96
$\boldsymbol{\Sigma}_2$	16.00	−8.00	−8.00	12.21	−5.92	−5.73	10.56	−5.29	−4.72
	−8.00	16.00	8.00	−5.92	15.67	3.66	−5.29	13.38	2.96
	−8.00	8.00	16.00	−5.73	3.66	13.43	−4.72	2.96	11.09
$\boldsymbol{\Sigma}_3$	2.25	1.57	1.57	2.00	1.32	1.37	1.80	1.21	1.27
	1.57	2.25	1.57	1.32	1.87	1.29	1.21	1.66	1.20
	1.57	1.57	2.25	1.37	1.29	2.11	1.27	1.20	1.84
$\boldsymbol{\mu}_1$	0.00	0.00	0.00	−0.25	0.41	−0.21	−0.21	0.17	0.02
$\boldsymbol{\mu}_2$	−4.00	4.00	0.00	−4.99	5.11	0.35	−5.06	4.94	0.45
$\boldsymbol{\mu}_3$	−4.00	0.00	4.00	−4.12	1.09	3.35	−3.52	0.39	3.62
$\boldsymbol{\Sigma}_1$	1.00	0.00	0.00	3.91	−1.63	−1.06	2.06	−0.51	−0.61
	0.00	1.00	0.00	−1.63	3.78	0.90	−0.51	1.79	0.34
	0.00	0.00	1.00	−1.06	0.90	3.57	−0.61	0.34	2.05
$\boldsymbol{\Sigma}_2$	16.00	−8.00	−8.00	11.09	−5.47	−5.53	10.45	−5.07	−5.30
	−8.00	16.00	8.00	−5.47	12.68	6.01	−5.07	11.75	5.11
	−8.00	8.00	16.00	−5.53	6.01	10.45	−5.30	5.11	10.05
$\boldsymbol{\Sigma}_3$	2.25	1.57	1.57	5.54	−1.45	−0.85	1.76	0.22	0.27
	1.57	2.25	1.57	−1.45	6.76	1.81	0.22	1.96	0.84
	1.57	1.57	2.25	−0.85	1.81	6.94	0.27	0.84	2.03
$\boldsymbol{\mu}_1$	0.00	0.00	0.00	−0.37	0.29	−0.14	−0.32	0.32	−0.02
$\boldsymbol{\mu}_2$	−4.00	4.00	0.00	−2.03	1.59	1.47	−0.32	0.32	−0.02
$\boldsymbol{\mu}_3$	−4.00	0.00	4.00	−4.66	−0.55	3.40	−4.41	−0.34	3.66
$\boldsymbol{\Sigma}_1$	1.00	0.00	0.00	3.80	−1.26	−1.06	2.70	−1.22	−0.64
	0.00	1.00	0.00	−1.26	2.89	0.49	−1.22	2.39	0.36
	0.00	0.00	1.00	−1.06	0.49	3.82	−0.64	0.36	2.35
$\boldsymbol{\Sigma}_2$	16.00	−8.00	−8.00	4.97	−2.68	−0.61	2.70	−1.22	−0.64
	−8.00	16.00	8.00	−2.68	6.30	−0.13	−1.22	2.39	0.36
	−8.00	8.00	16.00	−0.61	−0.13	5.50	−0.64	0.36	2.35
$\boldsymbol{\Sigma}_3$	2.25	1.57	1.57	1.92	0.60	0.88	1.20	0.45	0.62
	1.57	2.25	1.57	0.60	2.02	0.78	0.45	1.18	0.50
	1.57	1.57	2.25	0.88	0.78	2.02	0.62	0.50	1.27

Journal : Large 42979 | Article No : 67 | Pages : 22 | MS Code : 67 | Dispatch : 21-1-2020

Table 9 Simulated datasets using multivariate Student-t distributions Scenarios 1–4

	True			GPDC			TPDC		
$\boldsymbol{\mu}_1$	0.00	0.00	0.00	−0.05	0.06	−0.01	0.01	−0.00	−0.01
$\boldsymbol{\mu}_2$	−7.00	7.00	0.00	−6.98	7.02	−0.00	−7.00	7.02	0.00
$\boldsymbol{\mu}_3$	−7.00	0.00	7.00	−6.96	0.11	6.90	−6.99	0.01	7.02
$\boldsymbol{\Sigma}_1$	1.00	0.00	0.00	2.02	−0.60	−0.08	0.96	−0.09	−0.03
	0.00	1.00	0.00	−0.60	1.94	−0.01	−0.09	0.94	−0.01
	0.00	0.00	1.00	−0.08	−0.01	1.27	−0.03	−0.01	0.76
$\boldsymbol{\Sigma}_2$	1.00	0.00	0.00	1.47	−0.22	−0.01	1.01	−0.04	−0.01
	0.00	1.00	0.00	−0.22	1.62	−0.04	−0.04	1.06	−0.00
	0.00	0.00	1.00	−0.01	−0.04	1.24	−0.01	−0.00	0.90
$\boldsymbol{\Sigma}_3$	1.00	0.00	0.00	1.65	−0.06	−0.32	0.81	−0.03	−0.04
	0.00	1.00	0.00	−0.06	2.13	−0.72	−0.03	0.88	−0.08
	0.00	0.00	1.00	−0.32	−0.72	2.59	−0.04	−0.08	0.97
$\boldsymbol{\mu}_1$	0.00	0.00	0.00	−0.06	−0.01	0.07	0.01	−0.01	−0.00
$\boldsymbol{\mu}_2$	−7.00	7.00	0.00	−6.96	6.87	0.11	−7.02	7.00	0.01
$\boldsymbol{\mu}_3$	−7.00	0.00	7.00	−7.00	−0.02	7.02	−7.02	−0.01	7.03
$\boldsymbol{\Sigma}_1$	1.00	0.00	0.00	2.05	−0.06	−0.63	0.97	−0.02	−0.11
	0.00	1.00	0.00	−0.06	1.28	−0.01	−0.02	0.78	−0.00
	0.00	0.00	1.00	−0.63	−0.01	1.92	−0.11	−0.00	0.95
$\boldsymbol{\Sigma}_2$	1.00	0.00	0.00	1.67	−0.34	−0.01	0.80	−0.05	0.00
	0.00	1.00	0.00	−0.34	2.51	−0.73	−0.05	0.95	−0.10
	0.00	0.00	1.00	−0.01	−0.73	2.07	0.00	−0.10	0.90
$\boldsymbol{\Sigma}_3$	1.00	0.00	0.00	1.52	−0.04	−0.25	1.03	−0.03	−0.06
	0.00	1.00	0.00	−0.04	1.24	−0.05	−0.03	0.92	−0.02
	0.00	0.00	1.00	−0.25	−0.05	1.60	−0.06	−0.02	1.05
$\boldsymbol{\mu}_1$	0.00	0.00	0.00	−0.09	0.13	−0.03	−0.01	0.05	−0.02
$\boldsymbol{\mu}_2$	−4.00	4.00	0.00	−3.99	4.06	−0.03	−4.02	4.07	−0.02
$\boldsymbol{\mu}_3$	−4.00	0.00	4.00	−3.88	0.29	3.71	−3.94	0.14	3.89
$\boldsymbol{\Sigma}_1$	1.00	0.00	0.00	1.98	−0.61	−0.08	0.95	−0.21	−0.03
	0.00	1.00	0.00	−0.61	1.86	−0.01	−0.21	0.91	−0.01
	0.00	0.00	1.00	−0.08	−0.01	1.19	−0.03	−0.01	0.64
$\boldsymbol{\Sigma}_2$	1.00	0.00	0.00	1.33	−0.20	−0.01	0.75	−0.05	−0.01
	0.00	1.00	0.00	−0.20	1.54	−0.03	−0.05	0.81	0.00
	0.00	0.00	1.00	−0.01	−0.03	1.11	−0.01	0.00	0.64
$\boldsymbol{\Sigma}_3$	1.00	0.00	0.00	1.70	−0.09	−0.40	0.86	−0.05	−0.16
	0.00	1.00	0.00	−0.09	2.22	−0.88	−0.05	1.09	−0.37
	0.00	0.00	1.00	−0.40	−0.88	2.85	−0.16	−0.37	1.32
$\boldsymbol{\mu}_1$	0.00	0.00	0.00	−0.10	−0.03	0.13	−0.01	−0.02	0.04
$\boldsymbol{\mu}_2$	−4.00	4.00	0.00	−3.90	3.67	0.28	−3.97	3.87	0.14
$\boldsymbol{\mu}_3$	−4.00	0.00	4.00	−4.00	−0.05	4.06	−4.03	−0.03	4.08
$\boldsymbol{\Sigma}_1$	1.00	0.00	0.00	2.01	−0.05	−0.64	0.96	−0.02	−0.23
	0.00	1.00	0.00	−0.05	1.20	−0.02	−0.02	0.64	−0.01
	0.00	0.00	1.00	−0.64	−0.02	1.82	−0.23	−0.01	0.91
$\boldsymbol{\Sigma}_2$	1.00	0.00	0.00	1.72	−0.40	−0.06	0.83	−0.15	−0.02
	0.00	1.00	0.00	−0.40	2.75	−0.90	−0.15	1.25	−0.39
	0.00	0.00	1.00	−0.06	−0.90	2.18	−0.02	−0.39	1.07
$\boldsymbol{\Sigma}_3$	1.00	0.00	0.00	1.38	−0.03	−0.23	0.78	−0.02	−0.08
	0.00	1.00	0.00	−0.03	1.12	−0.03	−0.02	0.66	−0.01
	0.00	0.00	1.00	−0.23	−0.03	1.53	−0.08	−0.01	0.82

Journal : Large 42979	Article No : 67	Pages : 22	MS Code : 67	Dispatch : 21-1-2020

Table 10 Simulated datasets using multivariate Student-t distributions Scenarios 5–8

	True			GPDC			TPDC		
μ_1	0.00	0.00	0.00	− 0.16	0.17	− 0.02	− 0.01	0.02	− 0.01
μ_2	− 7.00	7.00	0.00	− 7.04	7.27	− 0.07	− 7.04	7.18	− 0.05
μ_3	− 7.00	0.00	7.00	− 6.99	0.47	6.59	− 7.00	0.14	6.93
Σ_1	1.00	0.00	0.00	2.80	− 1.12	− 0.04	0.97	− 0.15	− 0.02
	0.00	1.00	0.00	− 1.12	2.72	− 0.05	− 0.15	0.96	− 0.01
	0.00	0.00	1.00	− 0.04	− 0.05	1.60	− 0.02	− 0.01	0.74
Σ_2	16.00	0.00	0.00	4.83	− 0.20	0.02	3.58	− 0.14	− 0.02
	0.00	16.00	0.00	− 0.20	5.39	0.09	− 0.14	3.84	0.06
	0.00	0.00	16.00	0.02	0.09	4.25	− 0.02	0.06	3.14
Σ_3	2.25	0.00	0.00	2.90	− 0.14	− 0.21	1.32	− 0.07	− 0.05
	0.00	2.25	0.00	− 0.14	4.69	− 2.42	− 0.07	1.77	− 0.56
	0.00	0.00	2.25	− 0.21	− 2.42	5.80	− 0.05	− 0.56	1.97
μ_1	0.00	0.00	0.00	− 0.10	0.00	0.08	0.00	− 0.01	− 0.00
μ_2	− 7.00	7.00	0.00	− 6.96	6.76	0.36	− 7.01	6.93	0.18
μ_3	− 7.00	0.00	7.00	− 7.02	− 0.02	7.04	− 7.03	− 0.02	7.06
Σ_1	1.00	0.00	0.00	2.32	− 0.18	− 0.65	0.91	− 0.03	− 0.10
	0.00	1.00	0.00	− 0.18	1.42	−0.03	− 0.03	0.71	− 0.01
	0.00	0.00	1.00	− 0.65	− 0.03	2.06	− 0.10	− 0.01	0.87
Σ_2	16.00	0.00	0.00	5.13	− 0.65	− 0.00	2.94	− 0.34	−0.02
	0.00	16.00	0.00	− 0.65	8.04	− 2.46	− 0.34	4.20	− 1.14
	0.00	0.00	16.00	− 0.00	− 2.46	6.76	− 0.02	− 1.14	3.78
Σ_3	2.25	0.00	0.00	2.13	− 0.05	− 0.22	1.48	− 0.04	− 0.08
	0.00	2.25	0.00	− 0.05	1.83	− 0.20	− 0.04	1.32	− 0.04
	0.00	0.00	2.25	− 0.22	− 0.20	2.40	− 0.08	− 0.04	1.56
μ_1	0.00	0.00	0.00	− 0.16	0.21	− 0.05	− 0.05	0.09	− 0.03
μ_2	− 4.00	4.00	0.00	− 4.12	4.46	− 0.19	− 4.12	4.38	− 0.15
μ_3	− 4.00	0.00	4.00	− 3.99	0.62	3.53	− 3.98	0.40	3.72
Σ_1	1.00	0.00	0.00	2.57	− 0.62	− 0.02	1.11	− 0.20	− 0.02
	0.00	1.00	0.00	− 0.62	2.48	− 0.07	− 0.20	1.08	− 0.02
	0.00	0.00	1.00	−0.02	− 0.07	1.80	− 0.02	−0.02	0.82
Σ_2	16.00	0.00	0.00	4.42	− 0.02	0.03	2.61	− 0.05	− 0.01
	0.00	16.00	0.00	− 0.02	5.15	0.19	− 0.05	2.91	0.10
	0.00	0.00	16.00	0.03	0.19	4.06	− 0.01	0.10	2.34
Σ_3	2.25	0.00	0.00	3.18	− 0.14	− 0.17	1.56	− 0.10	− 0.10
	0.00	2.25	0.00	− 0.14	3.89	− 1.26	− 0.10	1.96	− 0.63
	0.00	0.00	2.25	− 0.17	− 1.26	4.89	− 0.10	− 0.63	2.30
μ_1	0.00	0.00	0.00	− 0.16	0.01	0.10	− 0.05	− 0.00	0.03
μ_2	− 4.00	4.00	0.00	− 3.95	3.53	0.75	− 3.95	3.45	0.75
μ_3	−4.00	0.00	4.00	− 4.06	− 0.05	4.11	−4.06	− 0.05	4.14
Σ_1	1.00	0.00	0.00	2.40	− 0.17	− 0.49	1.08	− 0.05	− 0.21
	0.00	1.00	0.00	− 0.17	1.57	− 0.12	− 0.05	0.76	− 0.04
	0.00	0.00	1.00	− 0.49	− 0.12	2.15	− 0.21	− 0.04	1.02
Σ_2	16.00	0.00	0.00	4.25	− 0.42	− 0.04	2.83	− 0.38	−0.09
	0.00	16.00	0.00	− 0.42	6.72	−2.02	− 0.38	4.79	− 1.79
	0.00	0.00	16.00	− 0.04	− 2.02	5.15	− 0.09	− 1.79	3.74
Σ_3	2.25	0.00	0.00	2.04	− 0.05	− 0.13	1.10	− 0.03	− 0.07
	0.00	2.25	0.00	− 0.05	1.81	− 0.16	− 0.03	0.97	− 0.02
	0.00	0.00	2.25	− 0.13	−0.16	2.41	− 0.07	− 0.02	1.21

| Journal : Large 42979 | Article No : 67 | Pages : 22 | MS Code : 67 | Dispatch : 21-1-2020 |

Table 11 Simulated datasets using multivariate Student-t distributions Scenarios 9–12

	True			GPDC			TPDC		
μ_1	0.00	0.00	0.00	− 0.11	0.11	0.04	− 0.01	0.01	0.03
μ_2	− 7.00	7.00	0.00	− 7.01	7.02	0.04	− 7.02	7.03	0.03
μ_3	− 7.00	0.00	7.00	− 6.97	0.03	7.03	− 6.98	0.02	7.03
Σ_1	1.00	0.00	0.00	2.37	− 0.83	− 0.08	1.16	− 0.14	− 0.04
	0.00	1.00	0.00	− 0.83	2.26	− 0.06	− 0.14	1.12	− 0.02
	0.00	0.00	1.00	− 0.08	− 0.06	1.43	− 0.04	− 0.02	0.91
Σ_2	1.00	− 0.50	− 0.50	1.55	− 0.92	−0.69	1.07	− 0.57	− 0.51
	− 0.50	1.00	0.50	− 0.92	1.69	0.68	− 0.57	1.12	0.51
	− 0.50	0.50	1.00	− 0.69	0.68	1.31	− 0.51	0.51	0.97
Σ_3	1.00	0.70	0.70	1.32	0.89	0.89	0.77	0.53	0.56
	0.70	1.00	0.70	0.89	1.44	0.89	0.53	0.80	0.57
	0.70	0.70	1.00	0.89	0.89	1.54	0.56	0.57	0.87
μ_1	0.00	0.00	0.00	− 0.10	− 0.01	0.14	− 0.01	− 0.01	0.02
μ_2	− 7.00	7.00	0.00	− 6.98	6.92	0.06	− 7.03	7.01	0.02
μ_3	− 7.00	0.00	7.00	− 7.01	− 0.00	7.01	− 7.01	− 0.00	7.01
Σ_1	1.00	0.00	0.00	2.46	− 0.00	− 0.87	1.11	− 0.02	− 0.14
	0.00	1.00	0.00	− 0.00	1.47	0.03	− 0.02	0.91	−0.00
	0.00	0.00	1.00	− 0.87	0.03	2.26	− 0.14	− 0.00	1.07
Σ_2	1.00	− 0.50	− 0.50	1.71	− 1.04	− 0.66	0.83	− 0.45	− 0.40
	− 0.50	1.00	0.50	− 1.04	2.24	0.43	− 0.45	0.92	0.39
	− 0.50	0.50	1.00	− 0.66	0.43	1.72	− 0.40	0.39	0.86
Σ_3	1.00	0.70	0.70	1.32	0.86	0.89	1.04	0.69	0.72
	0.70	1.00	0.70	0.86	1.18	0.86	0.69	0.97	0.70
	0.70	0.70	1.00	0.89	0.86	1.34	0.72	0.70	1.06
μ_1	0.00	0.00	0.00	−0.26	0.29	0.02	− 0.13	0.14	0.03
μ_2	− 4.00	4.00	0.00	− 4.08	4.09	0.09	− 4.07	4.08	0.07
μ_3	−4.00	0.00	4.00	− 3.93	0.07	4.01	−3.96	0.02	4.06
Σ_1	1.00	0.00	0.00	2.49	− 1.03	− 0.08	1.27	− 0.44	− 0.04
	0.00	1.00	0.00	− 1.03	2.30	− 0.07	− 0.44	1.19	− 0.06
	0.00	0.00	1.00	− 0.08	− 0.07	1.47	− 0.04	− 0.06	0.84
Σ_2	1.00	− 0.50	− 0.50	1.36	−0.80	− 0.62	0.89	− 0.49	− 0.41
	− 0.50	1.00	0.50	− 0.80	1.55	0.61	− 0.49	0.98	0.42
	− 0.50	0.50	1.00	− 0.62	0.61	1.14	− 0.41	0.42	0.77
Σ_3	1.00	0.70	0.70	1.25	0.77	0.73	0.60	0.42	0.44
	0.70	1.00	0.70	0.77	1.43	0.76	0.42	0.65	0.46
	0.70	0.70	1.00	0.73	0.76	1.62	0.44	0.46	0.73
μ_1	0.00	0.00	0.00	− 0.28	− 0.04	0.36	− 0.11	−0.01	0.18
μ_2	− 4.00	4.00	0.00	− 3.89	3.71	0.18	− 4.01	3.97	0.04
μ_3	− 4.00	0.00	4.00	− 4.01	− 0.00	4.06	− 4.01	0.01	4.05
Σ_1	1.00	0.00	0.00	2.81	0.18	− 1.16	1.36	0.03	− 0.50
	0.00	1.00	0.00	0.18	1.43	0.11	0.03	0.82	0.03
	0.00	0.00	1.00	−1.16	0.11	2.44	− 0.50	0.03	1.25
Σ_2	1.00	− 0.50	− 0.50	1.84	− 1.10	−0.55	0.81	− 0.49	− 0.35
	−0.50	1.00	0.50	− 1.10	2.72	0.13	− 0.49	1.02	0.30
	− 0.50	0.50	1.00	− 0.55	0.13	2.07	− 0.35	0.30	0.84
Σ_3	1.00	0.70	0.70	1.19	0.79	0.87	0.86	0.57	0.63
	0.70	1.00	0.70	0.79	1.00	0.81	0.57	0.76	0.59
	0.70	0.70	1.00	0.87	0.81	1.25	0.63	0.59	0.91

Journal : Large 42979	Article No : 67	Pages : 22	MS Code : 67	Dispatch : 21-1-2020

Table 12 Simulated datasets using multivariate Student-t distributions Scenarios 13–16

	True			GPDC			TPDC		
$\boldsymbol{\mu}_1$	0.00	0.00	0.00	−0.35	0.35	−0.04	−0.06	0.06	0.01
$\boldsymbol{\mu}_2$	−7.00	7.00	0.00	−7.22	7.28	0.20	−7.15	7.17	0.15
$\boldsymbol{\mu}_3$	−7.00	0.00	7.00	−6.97	0.12	6.95	−6.97	0.02	7.05
$\boldsymbol{\Sigma}_1$	1.00	0.00	0.00	3.87	−2.17	−0.03	1.22	−0.33	−0.00
	0.00	1.00	0.00	−2.17	3.67	−0.13	−0.33	1.18	−0.06
	0.00	0.00	1.00	−0.03	−0.13	2.15	−0.00	−0.06	0.90
$\boldsymbol{\Sigma}_2$	16.00	−8.00	−8.00	5.07	−2.80	−2.48	3.94	−2.09	−1.86
	−8.00	16.00	8.00	−2.80	5.56	2.46	−2.09	4.22	1.84
	−8.00	8.00	16.00	−2.48	2.46	4.51	−1.86	1.84	3.55
$\boldsymbol{\Sigma}_3$	2.25	1.57	1.57	2.05	1.22	1.19	1.01	0.70	0.74
	1.57	2.25	1.57	1.22	2.70	0.89	0.70	1.09	0.76
	1.57	1.57	2.25	1.19	0.89	2.98	0.74	0.76	1.18
$\boldsymbol{\mu}_1$	0.00	0.00	0.00	−0.15	0.02	0.15	−0.01	0.00	0.02
$\boldsymbol{\mu}_2$	−7.00	7.00	0.00	−7.09	6.93	0.29	−7.12	7.09	0.13
$\boldsymbol{\mu}_3$	−7.00	0.00	7.00	−7.01	−0.01	7.03	−7.01	−0.00	7.02
$\boldsymbol{\Sigma}_1$	1.00	0.00	0.00	2.83	−0.15	−0.99	1.03	−0.04	−0.14
	0.00	1.00	0.00	−0.15	1.62	0.04	−0.04	0.82	−0.01
	0.00	0.00	1.00	−0.99	0.04	2.56	−0.14	−0.01	0.99
$\boldsymbol{\Sigma}_2$	16.00	−8.00	−8.00	5.42	−3.00	−2.19	2.84	−1.57	−1.29
	−8.00	16.00	8.00	−3.00	7.28	1.35	−1.57	3.40	1.13
	−8.00	8.00	16.00	−2.19	1.35	5.99	−1.29	1.13	3.04
$\boldsymbol{\Sigma}_3$	2.25	1.57	1.57	1.89	1.24	1.31	1.43	0.94	1.01
	1.57	2.25	1.57	1.24	1.69	1.24	0.94	1.31	0.96
	1.57	1.57	2.25	1.31	1.24	1.96	1.01	0.96	1.48
$\boldsymbol{\mu}_1$	0.00	0.00	0.00	−0.42	0.51	−0.16	−0.30	0.31	−0.08
$\boldsymbol{\mu}_2$	−4.00	4.00	0.00	−4.50	4.61	0.31	−4.37	4.41	0.34
$\boldsymbol{\mu}_3$	−4.00	0.00	4.00	−3.83	0.40	3.57	−3.96	0.09	4.05
$\boldsymbol{\Sigma}_1$	1.00	0.00	0.00	3.44	−1.68	−0.32	1.95	−0.79	−0.07
	0.00	1.00	0.00	−1.68	3.25	0.21	−0.79	1.84	−0.06
	0.00	0.00	1.00	−0.32	0.21	2.63	−0.07	−0.06	1.50
$\boldsymbol{\Sigma}_2$	16.00	−8.00	−8.00	4.40	−2.35	−2.20	2.98	−1.53	−1.41
	−8.00	16.00	8.00	−2.35	5.01	2.21	−1.53	3.31	1.41
	−8.00	8.00	16.00	−2.20	2.21	3.83	−1.41	1.41	2.68
$\boldsymbol{\Sigma}_3$	2.25	1.57	1.57	2.70	0.49	0.35	1.02	0.62	0.66
	1.57	2.25	1.57	0.49	3.49	0.71	0.62	1.19	0.64
	1.57	1.57	2.25	0.35	0.71	4.04	0.66	0.64	1.35
$\boldsymbol{\mu}_1$	0.00	0.00	0.00	−0.44	0.26	0.17	−0.27	0.16	0.13
$\boldsymbol{\mu}_2$	−4.00	4.00	0.00	−2.50	2.03	0.77	−2.42	2.15	0.36
$\boldsymbol{\mu}_3$	−4.00	0.00	4.00	−3.11	0.13	3.09	−3.10	0.09	3.20
$\boldsymbol{\Sigma}_1$	1.00	0.00	0.00	3.12	−0.84	−0.85	1.78	−0.48	−0.50
	0.00	1.00	0.00	−0.84	2.19	0.23	−0.48	1.29	0.08
	0.00	0.00	1.00	−0.85	0.23	2.47	−0.50	0.08	1.41
$\boldsymbol{\Sigma}_2$	16.00	−8.00	−8.00	3.66	−2.07	−0.63	2.46	−1.40	−0.64
	−8.00	16.00	8.00	−2.07	4.99	0.16	−1.40	3.27	0.14
	−8.00	8.00	16.00	−0.63	0.16	3.83	−0.64	0.14	2.48
$\boldsymbol{\Sigma}_3$	2.25	1.57	1.57	1.49	0.66	0.82	0.90	0.52	0.63
	1.57	2.25	1.57	0.66	1.53	0.57	0.52	0.83	0.50
	1.57	1.57	2.25	0.82	0.57	1.82	0.63	0.50	1.03

Journal : Large 42979 | Article No : 67 | Pages : 22 | MS Code : 67 | Dispatch : 21-1-2020

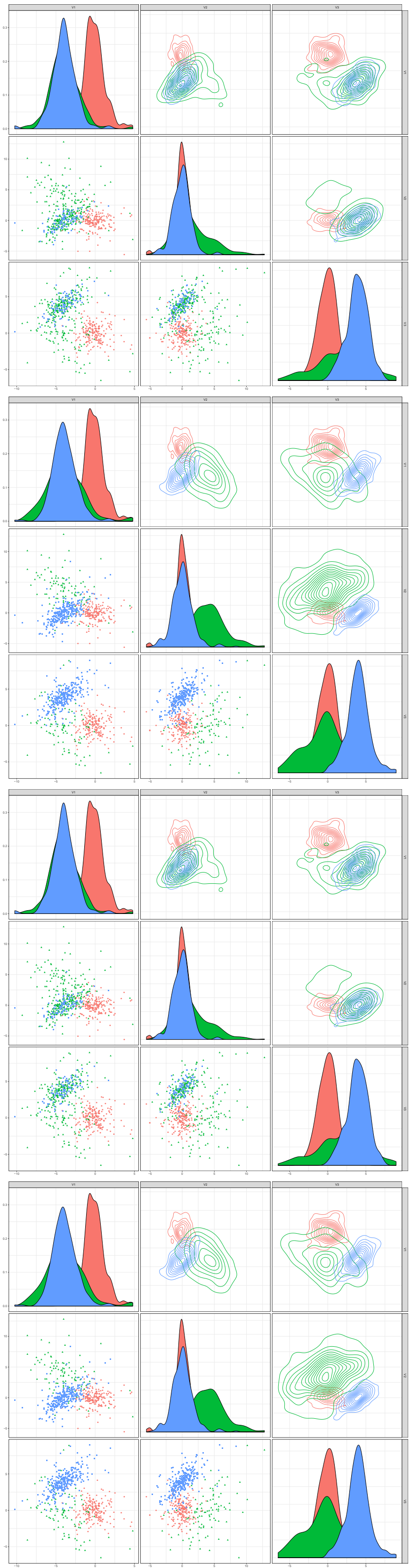

Fig. 5 Simulated datasets using multivariate Gaussian distributions Scenarios 1–4, each color and symbol representing a different cluster

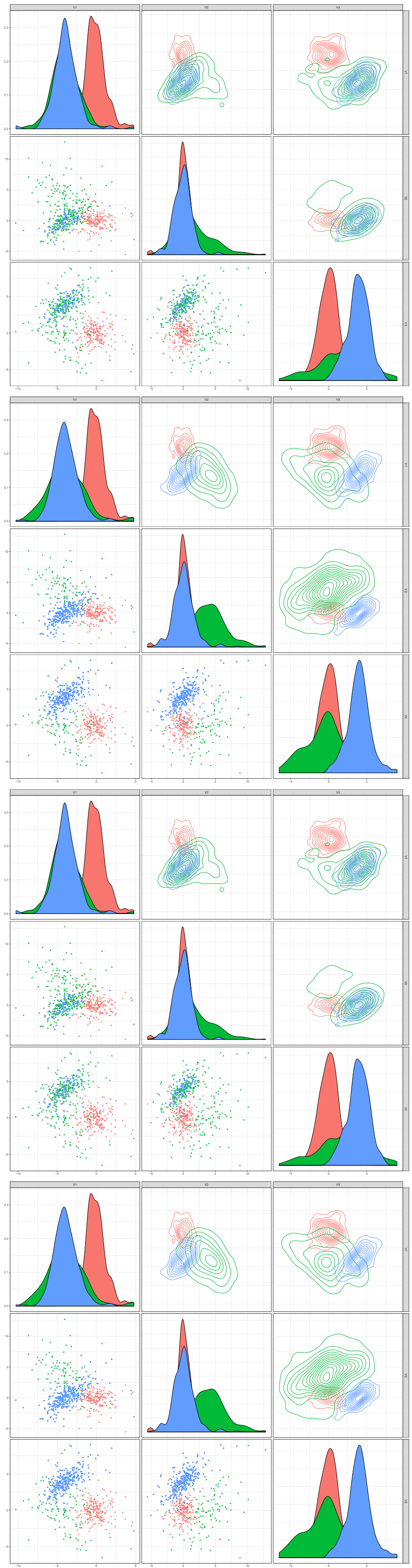

Fig. 6 Simulated datasets using multivariate Gaussian distributions Scenarios 5–8, each color and symbol representing a different cluster

Journal : **Large 42979**	Article No : **67**	Pages : **22**	MS Code : **67**	Dispatch : **21-1-2020**

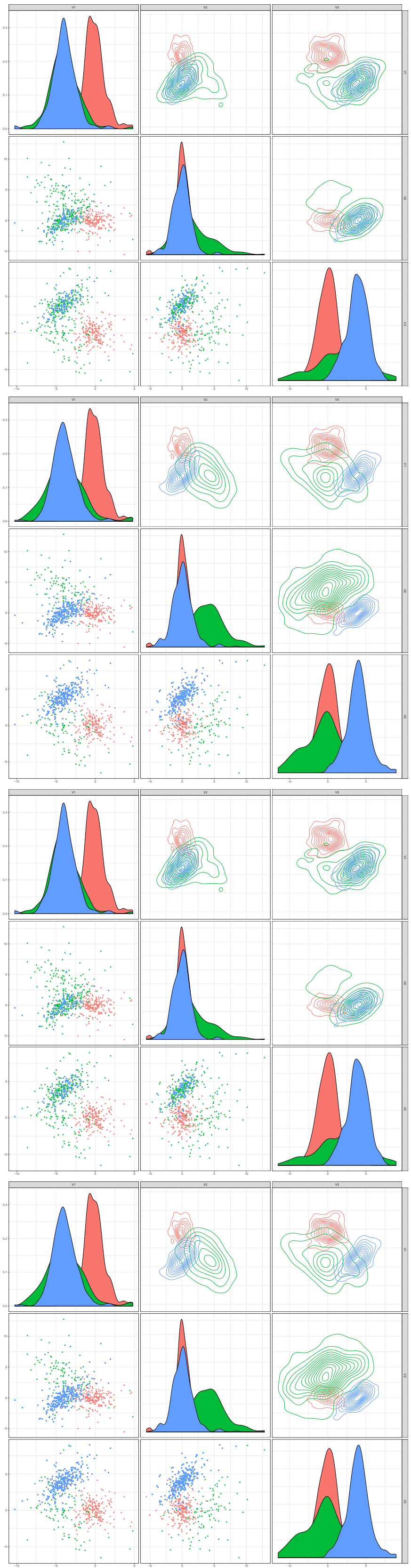

Fig. 7 Simulated datasets using multivariate Gaussian distributions Scenarios 9–12, each color and symbol representing a different cluster

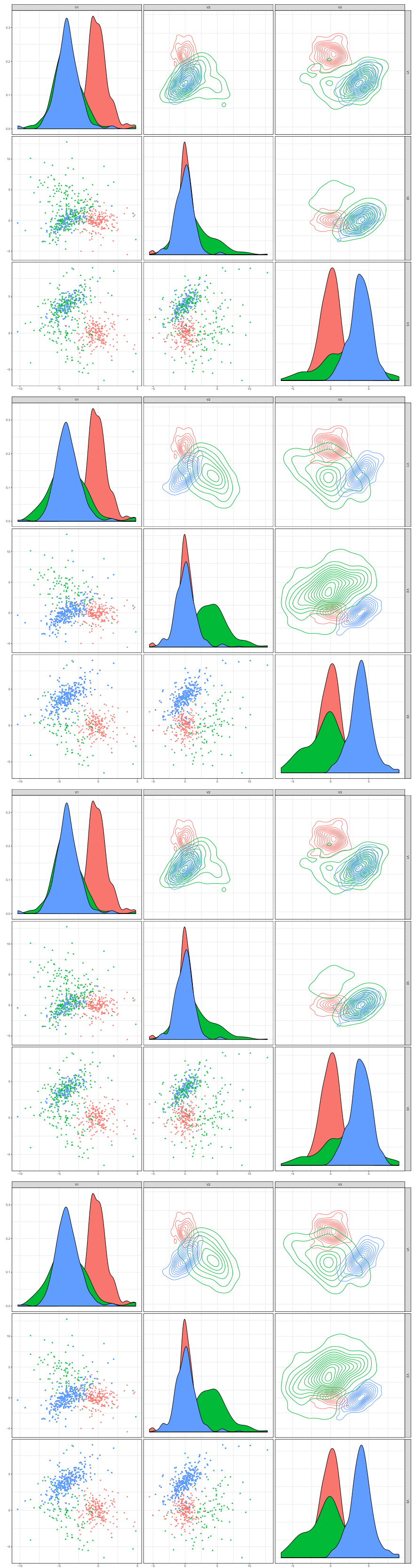

Fig. 8 Simulated datasets using multivariate Gaussian distributions Scenarios 13–16, each color and symbol representing a different cluster

Journal : **Large 42979**	Article No : **67**	Pages : **22**	MS Code : **67**	Dispatch : **21-1-2020**

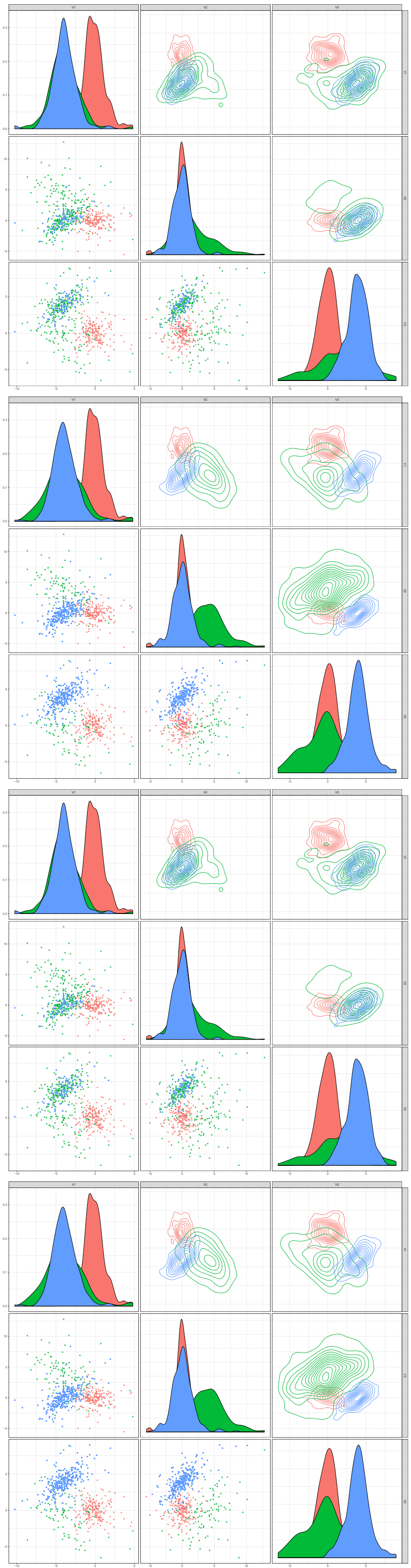

Fig. 9 Simulated datasets using multivariate Student-*t* distributions Scenarios 1–4, each color and symbol representing a different cluster

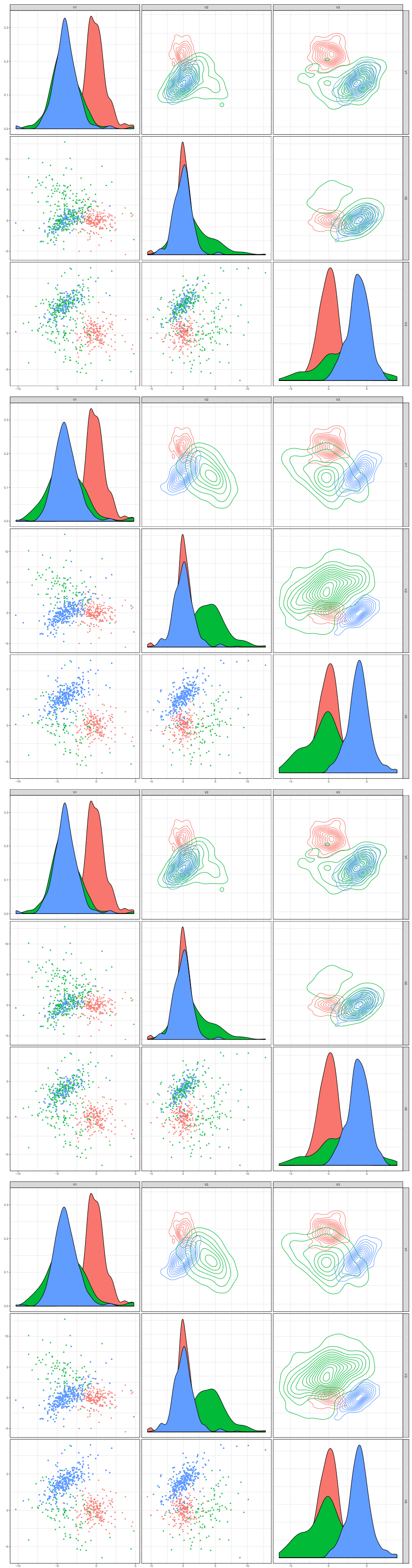

Fig. 10 Simulated datasets using multivariate Student-*t* distributions Scenarios 5–8, each color and symbol representing a different cluster

Journal : Large 42979	Article No : 67	Pages : 22	MS Code : 67	Dispatch : 21-1-2020

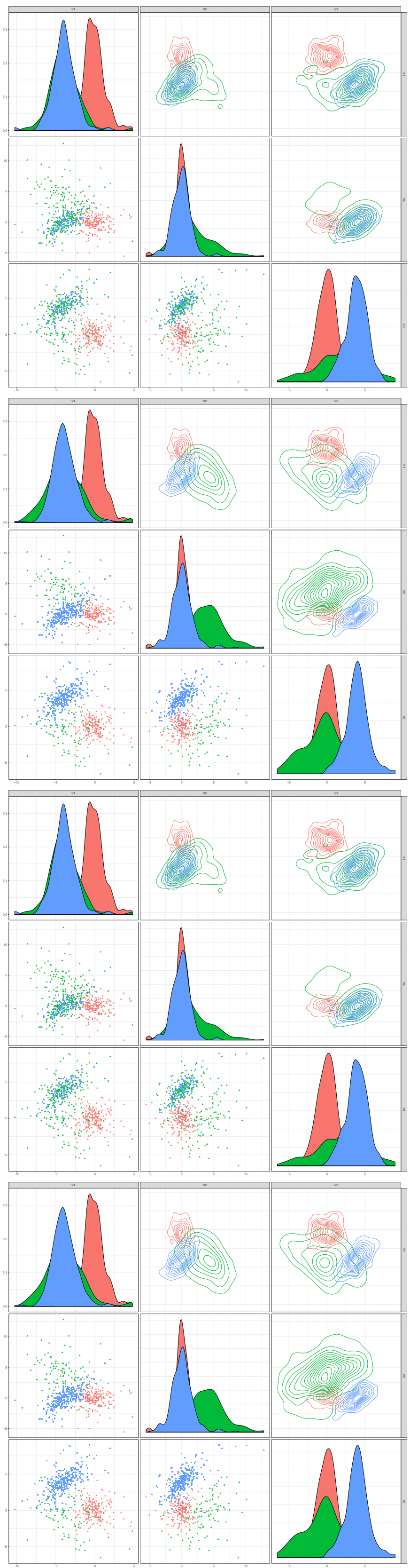

Fig. 11 Simulated datasets using multivariate Student-*t* distributions Scenarios 9–12, each color and symbol representing a different cluster

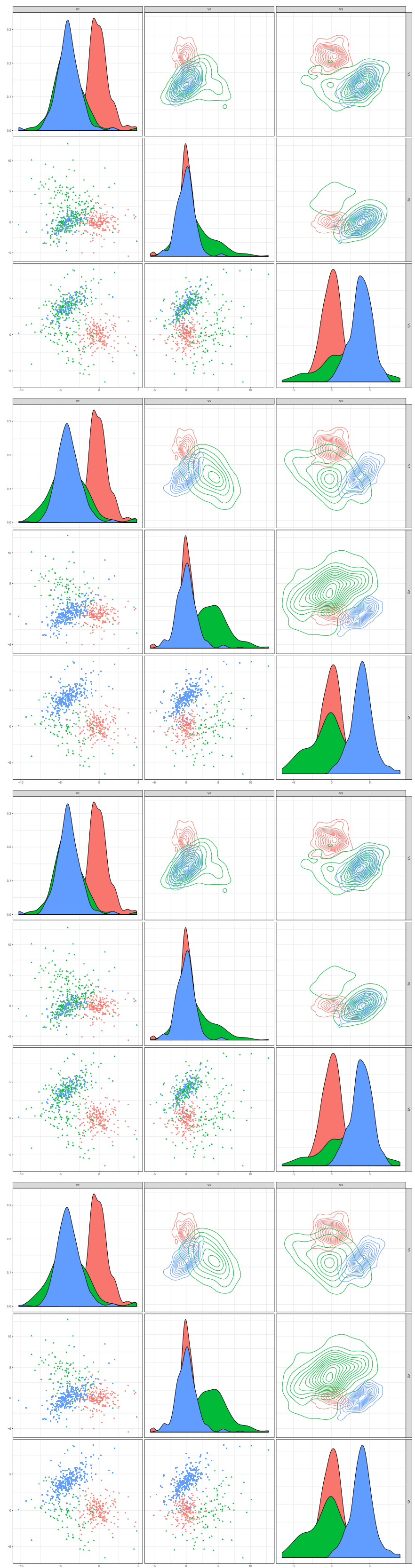

Fig. 12 Simulated datasets using multivariate Student-*t* distributions Scenarios 13–16, each color and symbol representing a different cluster

Journal : **Large 42979**	Article No : **67**	Pages : **22**	MS Code : **67**	Dispatch : **21-1-2020**

References

1. Andrews JL, Wickins JR, Boers NM, McNicholas PDT. An r package for model-based clustering and classification via the multivariate t distribution. J Stat Softw. 2018;83:7.
2. Barnett V. Comparative statistical inference. 3rd ed. Hoboken: Wiley; 1999.
3. Ben-Israel A, Iyigun C. Probabilistic d-clustering. J Classif. 2008;25(1):5–26.
4. Bezdek JC, Ehrlich R, Full W. Fcm: the fuzzy c-means clustering algorithm. Comput Geosci. 1984;10(2–3):191–203.
5. Blight B. Estimation from a censored sample for an exponential family. Biometrika. 1970;57:389–95.
6. Browne RP, ElSherbiny A, McNicholas PD. mixture: mixture models for clustering and classification; R package version 1.4. 2015.
7. Buck S. A method of estimation of missing values in multivariate data suitable for use with an electronic computer. J R Stat Soc B. 1960;22:302–6.
8. Chiang M, Mirkin B. Intelligent choice of the number of clusters in k-means clustering: an experimental study with different cluster spreads. J Classif. 2010;27(1):3–40.
9. Dang UJ, Browne RP, McNicholas PD. Mixtures of multivariate power exponential distributions. Biometrics. 2015;71(4):1081–9. https://doi.org/10.1111/biom.12351.
10. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc B. 1977;39(1):1–38.
11. Genz A, Bretz F, Miwa T, Mi X, Leisch F, Scheipl F, Hothorn T. mvtnorm: multivariate Normal and t Distributions; R package version 1.0-8. 2018.
12. Hubert L, Arabie P. Comparing partitions. J Classif. 1985;2(1):193–218.
13. Iyigun C. Probabilistic Distance Clustering. Ph.D. thesis, New Brunswick Rutgers, The State University of New Jersey. 2007.
14. Iyigun C, Ben-Israel A. Probabilistic distance clustering adjusted for cluster size. Prob Eng Inf Sci. 2008;22(04):603–21.
15. Kaufman L, Rousseeuw P. Finding groups in data: an introduction to cluster analysis. New York: Wiley; 1990.
16. Kulin HW, Kuenne RE. An efficient algorithm for the numerical solution of the generalized weber problem in spatial economics. J Reg Sci. 1962;4(2):21–33. https://doi.org/10.1111/j.1467-9787.1962.tb00902.x.
17. Lange KL, Little RJ, Taylor JM. Robust statistical modeling using the t distribution. J Am Stat Assoc. 1989;84(408):881–96.
18. Lee SX, McLachlan GJ. Finite mixtures of multivariate skew t-distributions: some recent and new results. Stat Comput. 2014;24(2):181–202.
19. Lin TI. Robust mixture modeling using multivariate skew t distributions. Stat Comput. 2010;20(3):343–56.
20. MacQueen J. Some methods for classification and analysis of multivariate observations. Proc Fifth Berkeley Symp. 1967;1:281–97.
21. McNicholas SM, McNicholas PD, Browne RP. A mixture of variance-gamma factor analyzers. In: Ahmed SE, editor. Big and complex data analysis: methodologies and applications. Cham: Springer International Publishing; 2017. p. 369–85.
22. Murray PM, McNicholas PD, Browne RB. A mixture of common skew-*t* factor analyzers. Statistics. 2014;3(1):68–82.
23. Newcomb S. A generalized theory of the combination of observation so as to obtain the best result. Am J Math. 1886;8:343–66.
24. Orchard T, Woodbury M. A missing information principle: Theory and applications. In: C.U.o.C.P. Berkley (ed.) Proceedings of the 6th Berkeley Symposium on Mathematical Statistics and Probability; 1972, vol 1, pp. 697–715
25. Punzo A, McNicholas PD. Parsimonious mixtures of multivariate contaminated normal distributions. Biometr J. 2016;58(6):1506–37.
26. R Core Team: R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2018.
27. R Core Team and contributors worldwide: stats: the R Stats Package 2014; R package version 3.1.2. 2014.
28. Rachev ST, Klebanov LB, Stoyanov SV, Fabozzi FJ. The methods of distances in the theory of probability and statistics. Berlin: Springer; 2013.
29. Rainey C, Tortora C, Palumbo F. A parametric version of probabilistic distance clustering. In: Greselin F, Deldossi L, Vichi M, Bagnato L, editors. Advances in statistical models for data analysis, studies in classification, data analysis, and knowledge organization. Cham: Springer; 2019. p. 33–43.
30. Rand WM. Objective criteria for the evaluation of clustering methods. J Am Stat Assoc. 1971;66:846–50.
31. Steinley D. Properties of the Hubert-Arable adjusted Rand index. Psychol Methods. 2004;9(3):386.
32. Tang Y, Browne RP, McNicholas PD. Flexible clustering of high-dimensional data via mixtures of joint generalized hyperbolic distributions. Statistics. 2018;7(1):e177.
33. Theodoridis S, Koutroumbas K. Pattern recognition. 2nd ed. New York: Academic Press; 2003.
34. Tortora C, Franczak BC, Browne RP, McNicholas PD. A mixture of coalesced generalized hyperbolic distributions. J Classif. 2019;36(1):26–57.
35. Tortora C, Gettler Summa M, Marino M, Palumbo F. Factor probabilistic distance clustering (FPDC): a new clustering method for high dimensional data sets. Adv Data Anal Classif. 2016;10(4):441–64.
36. Tortora C, Gettler Summa M, Palumbo F. Factor PD-clustering. In: Berthold UL, Dirk V (eds) Algorithms from and for Nature and Life; 2013, p. 115–123.
37. Tortora C, McNicholas PD. FPDclustering: PD-clustering and factor PD-clustering. R package version 1.4. 2019.

Journal : Large 42979	Article No : 67	Pages : 22	MS Code : 67	Dispatch : 21-1-2020