EMO 2021: International Conference on Evolutionary Multi-Criterion Optimization

On the Interaction between Distance Functions and Clustering Criteria in Multi-objective Clustering

Adán José-García & Julia Handl

March 28-31, 2021

Table of contents

Introduction

Introduction

Multi-criterion algorithms for clustering have gained traction due to their ability to cater to diverse cluster properties.

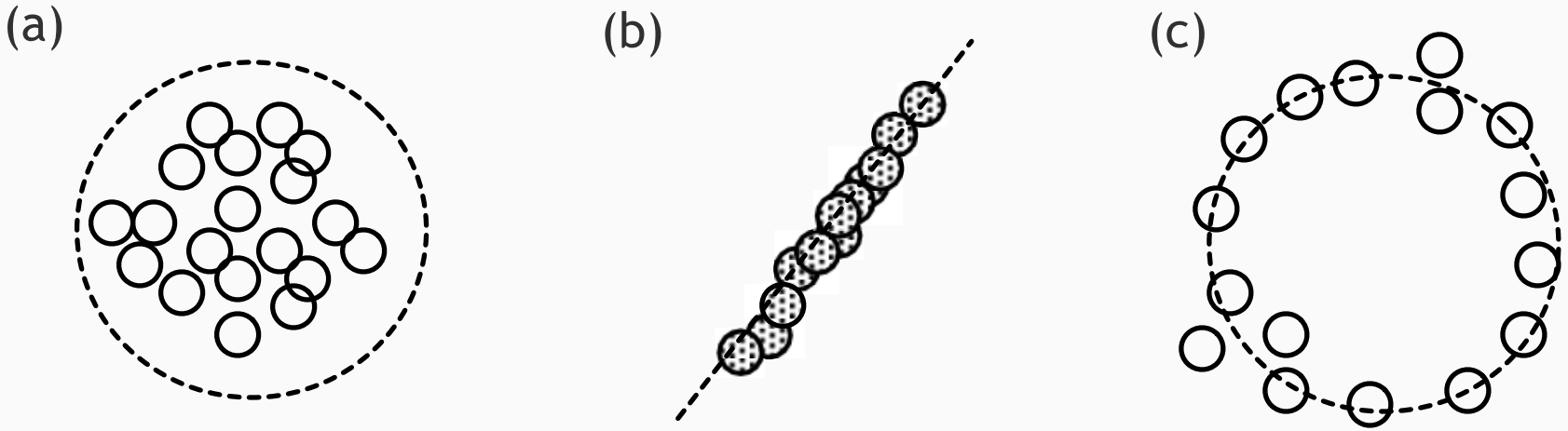


Illustration of three different cluster types: (a) compact cluster, (b) elongated cluster, and (c) arbitrarily-shaped cluster.

Background

Background: Compactness-based distance functions

The Euclidean distance

$$d_{\mathrm{EUC}}(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^{I} (x_i - y_i)^2}$$

where $\mathbf{x}, \mathbf{y} \in \mathbf{X}$ and x_i , y_i are the i-th coordinates of $\mathbf{x}$ and $\mathbf{y}$.

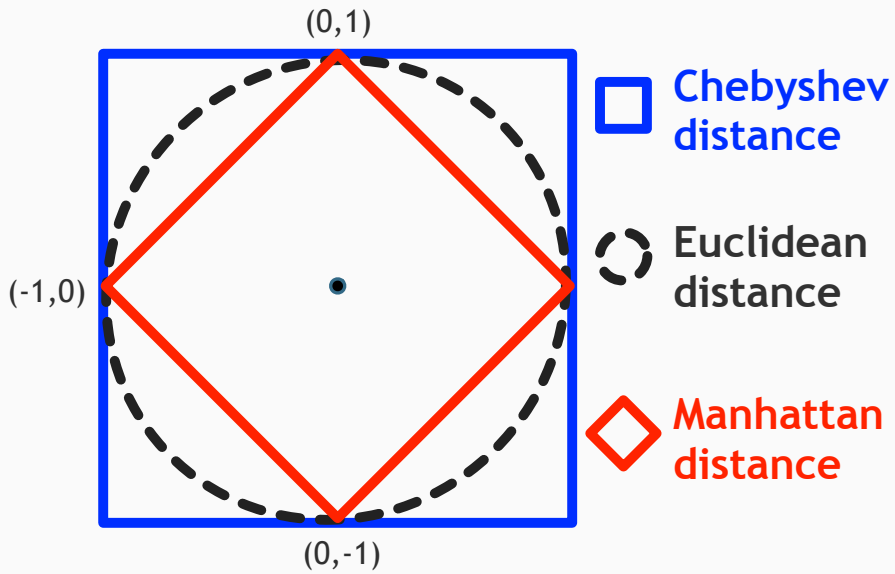


Illustration of the Minkowski distance

Background: A connectivity-based distance function

The maximum edge distance (MED) is defined over the Minimum Spanning Tree (MST) of the graph representing the original pair-wise dissimilarities in a given dataset.

$$d_{\text{MED}}(i,j) = \max\{e \in E_{P_{ij}}\} \tag{1}$$

where

- P_{ij} is the component of the MST that represents the path between data points i and j,
- $E_{P_{ij}}$ is the set of edges in P_{ij}, and
- e are the elements in $E_{P_{ij}}$.

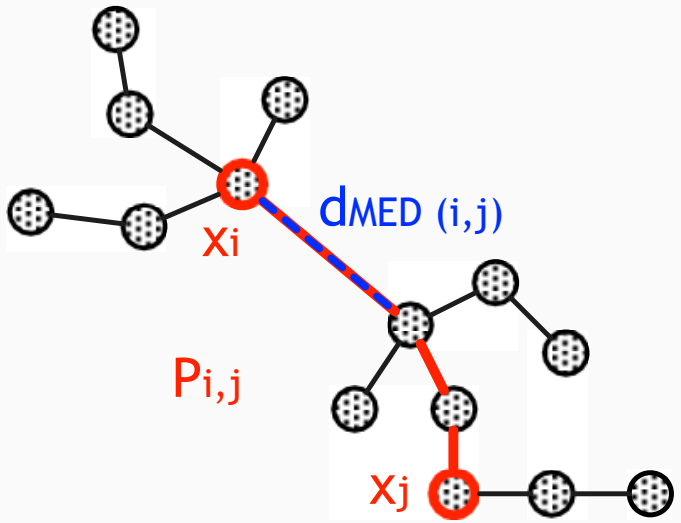
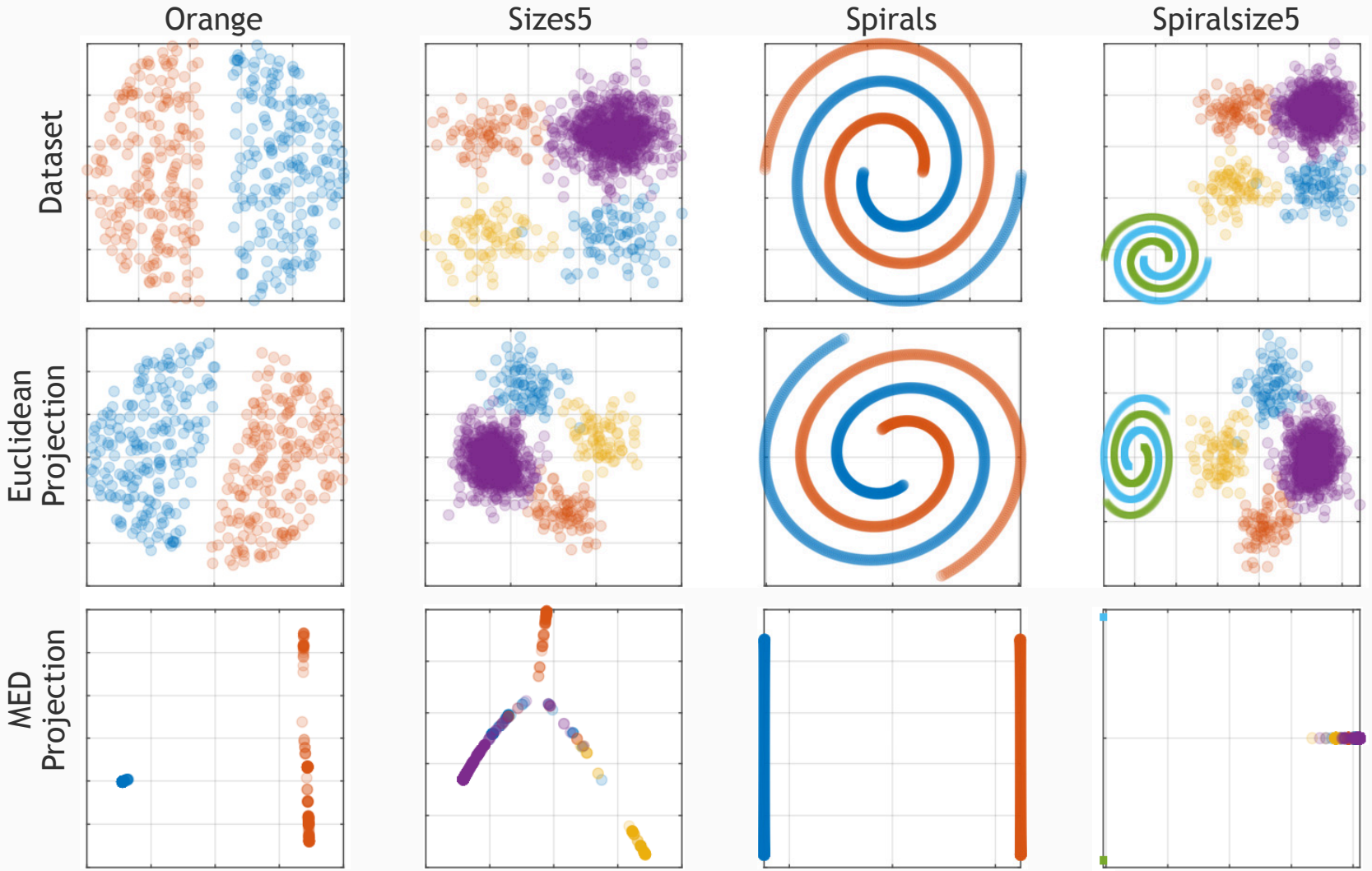


Illustration of the MED distance.

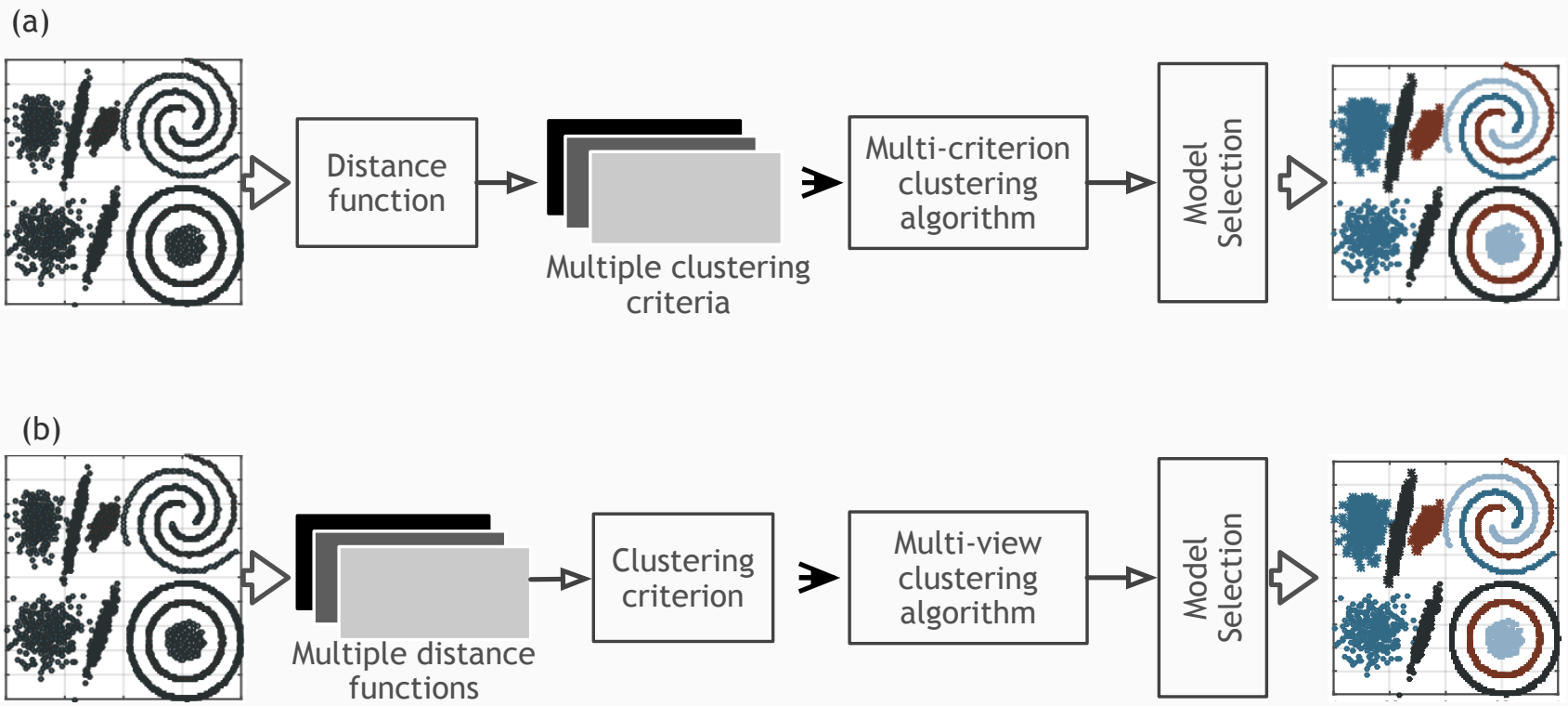
Background: The effect of different distance functions



The effect of different distances. (top) Original data; (centre) Embedding of the associated Euclidean distance; (bottom) Embedding of the MED distance.

Multi-criterion versus multi-view clustering

Multi-criterion versus multi-view clustering



Multi-objective clustering approaches: (a) multi-criterion clustering, and (b) multi-view clustering.

A multi-criterion clustering algorithm: Δ-MOCK

Δ-MOCK is a multi-criterion evolutionary algorithm with the following main characteristics [1]:

- uses a Pareto-based optimiser (NSGA-II).
- uses a flexible representation (the locus-based adjacency scheme).
- makes no prior assumptions on the cluster number.
- formulates multi-objective clustering as a bi-objective problem.

A multi-criterion clustering algorithm: Δ-MOCK

Δ-MOCK's clustering criteria:

- The criterion of compactness is defined as:

$$\mathrm{Var}(\mathbf{C}) = \frac{1}{N} \sum_{k=1}^{K} \sum_{i \in c_k} d(i, \mu_k)^2 \tag{2}$$

  where N is the number of data points, K is the number of clusters, c_k and μ_k are the partition and the centroid of cluster k, respectively.

- The criterion of connectivity is defined as:

$$\mathrm{Conn}(\mathbf{C}) = \sum_{i=1}^{N} \sum_{l=1}^{L} \rho(i, l) \tag{3}$$

  where L is a parameter specifying the size of the neighbourhood, and $\rho(i, l) = \frac{1}{l}$ is a penalty term incurred for $l \leqslant L$.

A multi-view clustering algorithm: MVMC

MVMC is a multi-view clustering approach with the following main characteristics [3]:

- uses a decomposition-based optimiser (MOEA/D [4]).
- achieves scalability with respect to the number of views through the use of a many-objective optimiser.
- uses a prototype-based representation.
- requires the desired number of clusters as input.
- focuses on a single clustering criterion but aims to optimise this with respect to every data view.

A multi-view clustering algorithm: MVMC

Let $\{D_1, \ldots, D_M\}$ denote the M dissimilarity matrices, which represent the different data views and are each considered by a separate objective.

MVMC uses the within-cluster scatter as the optimisation criterion. Let $\mathbf{C}^r$ and $\mathbf{w}^r$ be the partition and weight vector corresponding to the r-th subproblem. The m-th objective of the r-th subproblem is computed as:

$$f_m(\mathbf{C}^r) = \sum_{\mathbf{c}_k \in \mathbf{C}^r} \sum_{i,j \in \mathbf{c}_k} d_m(i,j) \ , \tag{4}$$

where

- $\mathbf{C}^r$ is a consensus clustering solution.
- $d_m(\cdot)$ is the dissimilarity between two points as defined in D_m.

Experimental Setup

Experimental Setup

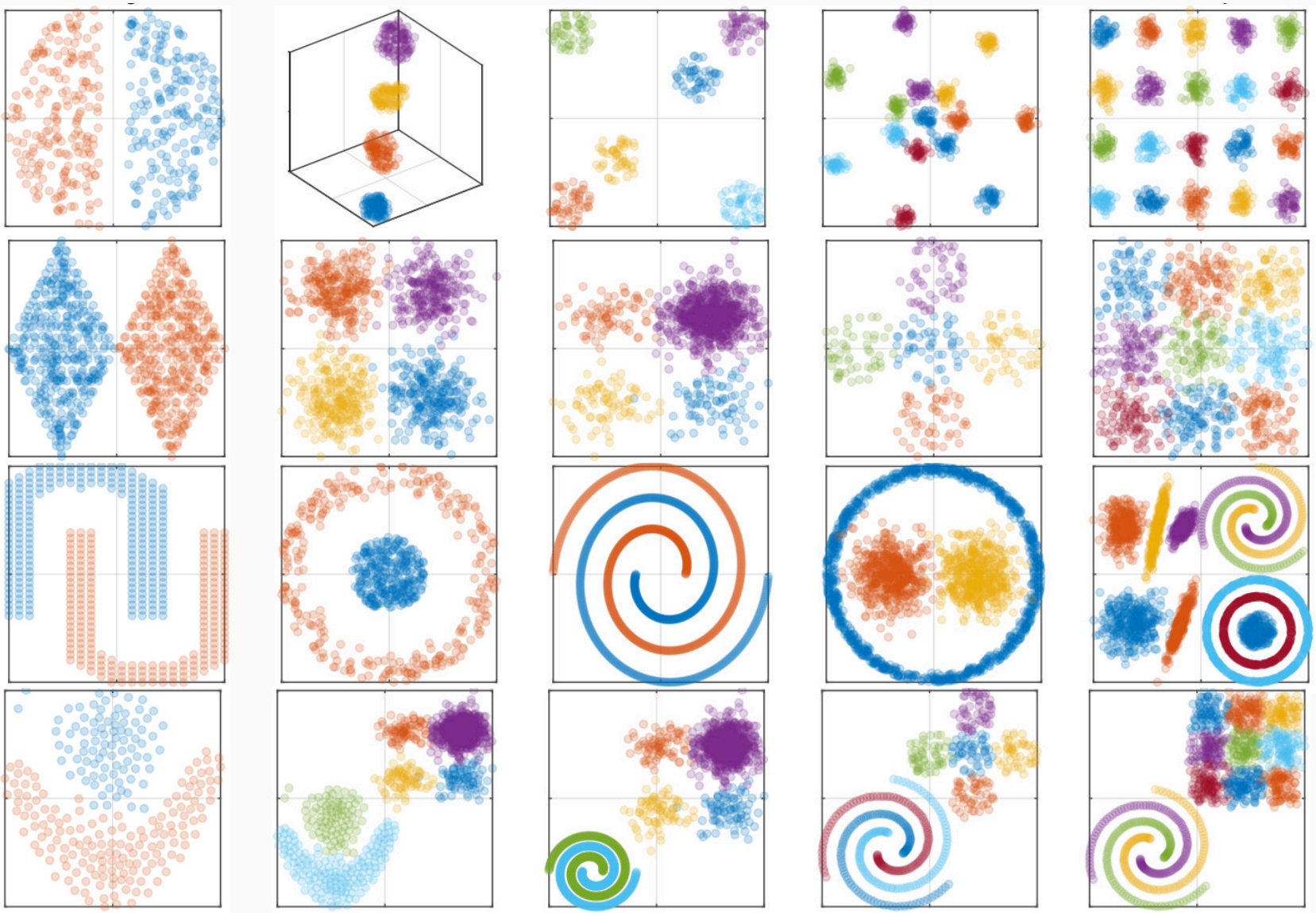


Illustration of the diversity of properties covered by our test datasets.

Experimental Setup

Performance Assessment:

- The Adjusted Rand Index (ARI) is used to assess the clustering performance [2].
- ARI is defined in the range $[\sim 0, 1]$; the larger the value, the better the correspondence between the obtained and the true partition.
- Each run of MVMC and Δ-MOCK produces a set of partitions. From this set, the best solution in terms of ARI is selected.

Reference Methods:

- Single-objective algorithms: k-means and Single-Link (SL).
- A MVMC_{WGS} version that minimises the within-cluster scatter.
- A MVMC_{SIL} version that optimises the Silhouette Width [5].

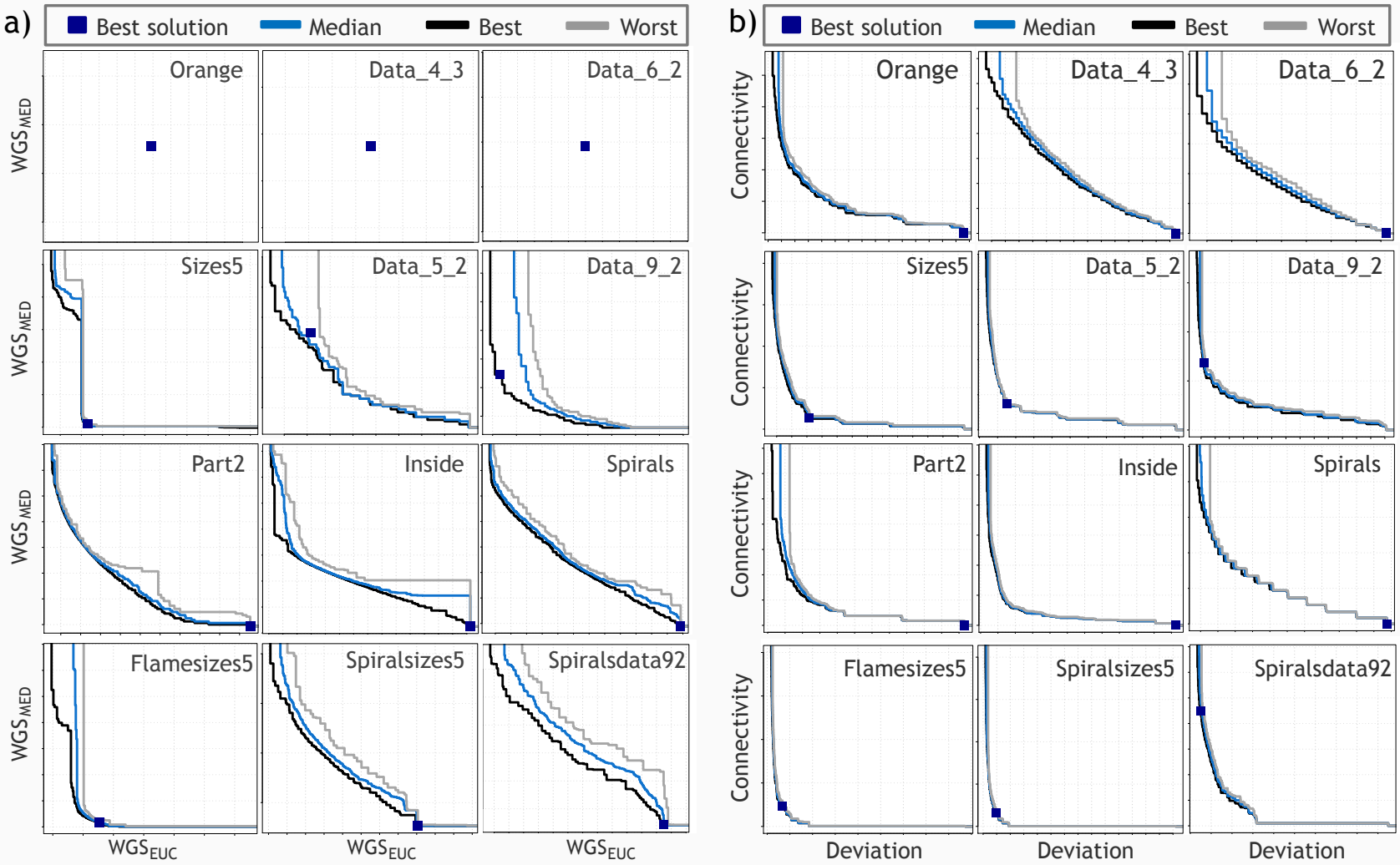
Findings and Results

Findings: Empirical Results

Table 1: ARI values obtained by the different clustering algorithms k-means, SL, δ-MOCK, and MVMC. The best ARI value scored for each dataset has been shaded and highlighted in bold and, additionally, the statistically best ($\alpha = 0.05$) results are highlighted in boldface.

Dataset	N	D	K	k-means$^{\blacktriangle}$	k-means$^{\blacktriangledown}$	SL$^{\blacktriangle}$	SL$^{\blacktriangledown}$	Δ-MOCK	MVMC$^{\blacktriangle\blacktriangledown}_{\text{SIL}}$	MVMC$^{\blacktriangle\blacktriangledown}_{\text{WGS}}$
Orange	400	2	2	**1.000±0.00**	**1.000±0.00**	**1.000**	**1.000**	**1.000±0.00**	**1.000±0.00**	**1.000±0.00**
Data_4_3	400	3	4	**1.000±0.00**	**1.000±0.00**	**1.000**	**1.000**	**1.000±0.00**	**1.000±0.00**	**1.000±0.00**
Data_6_2	300	2	6	**1.000±0.00**	**1.000±0.00**	**1.000**	**1.000**	**1.000±0.00**	**1.000±0.00**	**1.000±0.00**
R15	600	2	15	**0.966±0.05**	0.975±0.00	0.548	0.751	**0.992±0.00**	**0.992±0.00**	**0.993±0.00**
Twenty	1000	2	20	0.966±0.04	**1.000±0.00**	**1.000**	**1.000**	**1.000±0.00**	**1.000±0.00**	**1.000±0.00**
TwoDiamonds	800	2	2	**1.000±0.00**	0.895±0.01	0.000	0.000	**0.999±0.00**	**0.999±0.00**	**1.000±0.00**
Square1	1000	2	4	**0.973±0.00**	0.906±0.01	0.000	0.000	**0.946±0.18**	**0.978±0.00**	**0.979±0.00**
Size5	1000	2	4	**0.920±0.00**	0.739±0.01	0.025	0.015	**0.961±0.01**	**0.970±0.00**	**0.962±0.00**
Data_5_2	250	2	5	0.870±0.01	0.750±0.02	0.189	0.394	**0.916±0.02**	**0.946±0.04**	**0.930±0.02**
Data_9_2	900	2	9	**0.831±0.00**	0.506±0.03	0.000	0.000	0.748±0.01	0.826±0.01	**0.838±0.01**
Part2	417	2	2	0.265±0.00	**1.000±0.00**	**1.000**	**1.000**	**1.000±0.00**	**1.000±0.00**	**1.000±0.00**
Inside	600	2	2	0.008±0.01	**1.000±0.00**	**1.000**	**1.000**	**1.000±0.00**	**1.000±0.00**	**1.000±0.00**
Spirals	1000	2	2	0.074±0.00	**1.000±0.00**	**1.000**	**1.000**	**1.000±0.00**	**1.000±0.00**	**1.000±0.00**
Ringauss	2000	2	3	0.252±0.00	**0.963±0.02**	0.001	0.001	0.473±0.02	**0.971±0.00**	**0.972±0.00**
Multidist	3012	2	11	0.532±0.03	**0.991±0.00**	0.805	0.805	0.764±0.02	0.967±0.06	**0.984±0.04**
Flame	240	2	2	0.462±0.02	**0.910±0.05**	0.013	0.013	**0.963±0.01**	**0.935±0.03**	**0.967±0.00**
Flamesize5	240	2	6	0.926±0.01	0.824±0.00	0.489	0.657	**0.971±0.01**	**0.948±0.01**	**0.976±0.00**
Spiralsizes5	2000	2	6	0.659±0.00	0.833±0.02	0.555	0.782	**0.987±0.00**	**0.974±0.04**	**0.980±0.02**
Spiralsdata52	562	2	8	0.342±0.00	0.934±0.01	0.772	0.808	0.735±0.03	0.948±0.00	**0.960±0.01**
Spiralsdata92	1212	2	12	0.610±0.02	0.623±0.02	0.128	0.130	0.677±0.02	0.750±0.12	**0.878±0.01**

Findings: Empirical Results



Comparison of the EAFs computed from the PFAs obtained from all independent execution of (a) MVMC and (b) Δ-MOCK approaches.

Findings: Mathematical Analysis

We analyse Δ-MOCK's objectives in order to understand the similarities and differences between Δ-MOCK and MVMC.

- The similarity in the first objective is straightforward as both algorithms optimise a compactness-based criterion.
- The Δ-MOCK's second objective can be reformulated as a minimisation problem over a non-metric dissimilarity measure $d(i,j) = 1 - s(i,j)$ as:

$$\min \ \text{Conn}^d(\mathbf{C}) = \min \ \sum_{k=1}^{K} \sum_{i \in c_k} \sum_{j \in c_k \setminus i} d(i,j) - \sum_{k=1}^{K} |c_k| \left(|c_k| - 1\right) \quad (5)$$

Discussion and Conclusions

Summary

Our work makes the following main contributions in the context of evolutionary multi-objective clustering:

- For Δ-MOCK: Mathematically, the underlying multi-criterion optimisation problem remains equivalent, but the reformulation explicitly highlights the parallels to a multi-view approach.
- For MVMC: Empirically, competitive results to multi-criterion clustering can be obtained when a multi-view algorithm is used with an appropriate pair of distance measures.
- In general: We showed that the suitable choice of a distance function overcomes the common limitations of prototype-based encodings and compactness-based criteria.

Summary

Get the datasets and source code of this study from

`https://mvc-repository.github.io/`

References i


M. Garza-Fabre, J. Handl, and J. Knowles.
An Improved and More Scalable Evolutionary Approach to Multiobjective Clustering.
IEEE Transactions on Evolutionary Computation, 22(4):515–535, 2018.


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


A. Jose-Garcia, J. Handl, W. Gomez-Flores, and M. Garza-Fabre.
An evolutionary many-objective approach to multiview clustering using feature and relational data.
Applied Soft Computing (under review), 2021.

References ii

Qingfu Zhang and Hui Li.
MOEA/D: A Multiobjective Evolutionary Algorithm Based on Decomposition.
IEEE Transactions on Evolutionary Computation, 11(6):712–731, 2007.

P. J. Rousseeuw.
Silhouettes: A graphical aid to the interpretation and validation of cluster analysis.
Journal of Computational and Applied Mathematics, 20:53–65, 1987.