A distributional framework for evaluation, comparison and uncertainty quantification in soft clustering

Andrea Campagner, Davide Ciucci, Thierry Denœux

A Distributional Framework for Evaluation, Comparison and Uncertainty Quantification in Soft Clustering

Andrea Campagner[a], Davide Ciucci[b], Thierry Denœux[c,d]

[a]*IRCCS Istituto Ortopedico Galeazzi, Milano, Italy*
[b]*Dipartimento di Informatica, Sistemistica e Comunicazione, Università degli Studi di Milano-Bicocca, Milano, Italy*
[c]*Université de technologie de Compiègne, CNRS, UMR 7253 Heudiasyc, Compiègne, France*
[d]*Institut universitaire de France, Paris, France*

Abstract

In this article, we propose and study a general framework for comparing and evaluating soft clusterings, viewed as a form of uncertainty quantification for clustering tasks, with the aim of studying the uncertainty that arises from such a comparison. Our proposal is based on the interpretation of soft clustering as describing uncertain information, represented in the form of a mass function, about an unknown, target hard clustering. We present a general construction to this purpose, called *distributional measures*, whereby any evaluation measure can be naturally extended to soft clustering. We study the theoretical properties of the proposed approach, both in terms of computational complexity and metric properties. Furthermore, we study its relationship with other existing proposals providing, in particular, necessary and sufficient conditions for the equivalence between distributional measures and the recently proposed framework of *transport-theoretic measures*. We also propose sampling-based approximation algorithms with convergence guarantees, making it possible to apply the proposed method to real-world datasets. Finally, we apply the distributional measures in a computational experiment in order to illustrate their usefulness.

Key words: Clustering analysis, Soft clustering, Evaluation, Validation, Comparison

1. Introduction

Clustering [36] refers to the act of (as well as to the algorithms for) partitioning a set of objects into groups, called *clusters*, supposed to encode some unknown classification defined according to a given property. Depending on the formalism that is adopted to represent the mentioned partitioning of the data, one can distinguish two main families of clusterings (and, consequently, of clustering algorithms), namely: *hard clustering* and *soft clustering*.

In the case of hard clustering, the assignment of objects to clusters is one-to-one: each object is precisely assigned to one and only cluster. In the case of soft clustering [32, 22], by

Email address: a.campagner@campus.unimib.it (Andrea Campagner[a])

contrast, the assignment of objects to clusters is affected by uncertainty, which is explicitly represented in the partitioning through some uncertainty quantification formalism. Popular approaches in this category include: probabilistic and fuzzy clustering [35], in which the assignment is represented through a fuzzy partition; three-way and rough clustering [31, 45, 7], in which the assignment is represented through a rough partition; possibilistic clustering [26], in which the assignment is represented through a possibilistic partition; and evidential clustering [19, 17, 15], in which the assignment is represented through an evidential partition: thus, it is the most general formalism among the mentioned ones.

When clustering is applied to real data, with the aim of discovering interesting relationships among a given set of instances, one of the most critical steps is *clustering evaluation* [44], that is, the evaluation of the obtained results. In this context, *internal validation* criteria refer to measures that evaluate the goodness of a given clustering based only on characteristics of the clustering itself, such as its intra-cluster homogeneity or inter-cluster separatedness, while *external validation* criteria refer to evaluation approaches and metrics that compare two different clusterings of the data at hand. While internal validation criteria can be important as *goodness–of–fit* measures, only external validation criteria can be used to objectively assess the quality of a given clustering [34], whenever the reference clustering to be compared with it is interpreted as the *ground truth*. For this reason, external validation is of fundamental importance in the evaluation of newly developed clustering algorithms[1]: we will focus solely on this family of measures, which we simply refer to as *validation measures*.

While several external validation measures have been proposed in the context of hard clustering (including, among others, the Rand index [33], the mutual information [43], or the partition distance [11]), how to properly evaluate the results of a clustering analysis is much less clear in the case of soft clustering methods. Indeed, the interplay between the uncertainty represented in the object-cluster assignments and the errors in the given partitioning (in comparison with the given ground truth partitioning of the data) makes the evaluation of such soft clustering algorithms particularly difficult [9]. For this reason, the development of evaluation measures for soft clustering has largely focused on the extension of common measures [1], notably the Rand index [10, 18, 23, 24], while a general approach to extend other comparison measures to soft clustering has so far been largely lacking.

As a way to bridge this gap, in a series of recent articles [8, 9] we proposed some principles for developing new evaluation metrics for soft clustering, based on the intuition that any given soft clustering can be represented as a distribution (in particular, a mass function) that encodes the belief about an underlying, but unknown, true hard clustering. Indeed, an evidential clustering (as the most general form of soft clustering among the ones considered in

[1] The described application is actually one of the two most common use cases for external validation measures: we have developed a new clustering algorithm and we want to assess its quality by comparing its results on some datasets for which we *know* the ground truth partitioning. That is, basically, we have a collection of supervised datasets that we use to benchmark the newly developed algorithm as a way to evaluate whether, on new datasets, it will similarly work well. The other major application of external validation measures is instead to compare two different clusterings of the same data, in order to assess whether these identify similar clustering structures.

this article) associates with each instance a mass function over the clusters, which represents the uncertainty about the assignment of that specific instance. Obviously, such instance-wise uncertainty quantification can be transformed to a partition-wise uncertainty quantification by simply transferring the mass functions for the single instances to a global mass function over hard clusterings: this latter mass function is considered as a representation of the uncertainty (determined by the evidential clustering algorithm that generated the evidential clustering at hand) about a set of *compatible* hard clusterings of the data, one of which is assumed to be *optimal* (i.e., the hard clustering, among the ones compatible with the given soft clustering, which is as similar as possible to the true, unknown clustering of the data).

In particular, we noted that one can distinguish two different purposes in such an evaluation. First, to objectively evaluate the quality of a soft clustering with respect to a known ground truth partitioning: since any given soft clustering, as mentioned above, can be represented as a distribution of hard clusterings, this amounts to finding bounds on the quality, with respect to the given ground truth, of the hard clusterings compatible with the given soft clustering. Second, to compare two different soft clusterings[2] and to quantify the uncertainty arising from their comparison: that is, to compare all the hard clusterings compatible with the given soft clusterings, and to propagate the degree of uncertainty represented by the two soft clusterings to the quality comparison. Based on these two principles, we proposed a general mathematical framework [9] aimed at addressing the former need mentioned above (i.e. evaluating the result of a soft clustering algorithm in comparison with a given ground truth). This approach, grounded on *optimal transport theory* and a novel representation of soft clustering in *distributional* terms, allows us to extend any validation measure from the hard clustering setting to the soft clustering one using a principled approach.

In this article, which is an extension of our previous conference paper [8], we continue our study of validation measures for soft clustering. After recalling our previous contributions in [9], we provide three main new contributions. First, we propose an alternative approach for the comparison of soft clusterings, which aims at comprehensively evaluating the uncertainty represented by a soft clustering, through the definition of so-called *distributional measures*. Second, we draw connections between the transport-theoretic measures introduced in [9] and the framework of distributional measures, showing their equivalence in the special but important case in which one of the two clusterings to be compared is a hard clustering. Finally, we propose algorithms (both exact and approximation ones) for computing distributional measures and we show their application to a sample collection of datasets, so as to evaluate their practical applicability and illustrate the kind of reasoning they enable.

[2] A possible use case for the mentioned comparison is when the two soft clusterings are obtained from two different clustering algorithms, applied to the same set of data. In this case, our aim would be to assess whether the underlying hard clusterings compatible with the soft clusterings represent similar clustering structures. However, one could also consider to compare two clusterings obtained by the same algorithm, but with different hyper-parameter settings, or different randomization choices taken through the course of the clustering process. In this case, our aim would be to evaluate the stability of the compatible hard clusterings. In both cases, the existence of a hard clustering C compatible with both the soft clusterings could be taken as evidence that C indeed represents some latent characteristic of the data.

The rest of the paper is organized as follows. The necessary background on soft clustering and evaluation measures will first be recalled in Section 2. Our new framework will be then introduced in Section 3, and approximation methods will be presented in Section 4. Finally, illustrative experiment will be reported in Section 5, and Section 6 will conclude the paper.

2. Background and Related Work

General background on clustering will first be exposed in Section 2.1. The distributional representation introduced in [9] will then be recalled in Section 2.2, and an overview of evaluation measures for soft clustering will be provided in Section 2.3.

2.1. Background on Clustering

Let $X = \{x_1, \ldots, x_n\}$ be a collection of objects. A hard clustering is a partitioning of objects in X into groups, called *clusters*. Formally, a hard clustering can be represented by a surjective mapping $C : X \to \Omega$, where $\Omega = \{\omega_1, \ldots, \omega_k\}$ is a set of clusters. This representation is called "object-based". We denote with $\mathscr{C}_\Omega = \{C : X \to \Omega\}$ the set of hard clusterings having Ω as codomain. Given a hard clustering $C \in \mathscr{C}_\Omega$, we can define the equivalence relation $[C] = \{(x, x') \in X \times X : C(x) = C(x')\}$, called the relational representation of C. Two hard clusterings C_1, C_2 are said to be equivalent iff $[C_1] = [C_2]$; we then write $C_1 \sim C_2$. We denote by $\mathscr{C}_\pi = \mathscr{C}_\Omega \backslash \sim$ the quotient of $\mathscr{C}_\Omega$ with respect to $\sim$.

Example 2.1. *Let $X = \{x_1, \ldots, x_5\}$ and $\Omega = \{\omega_1, \omega_2, \omega_3\}$. Then, the clustering $C : X \mapsto \Omega$ defined by $C(x_1) = C(x_5) = \omega_1$, $C(x_2) = C(x_3) = \omega_2$ and $C(x_4) = \omega_3$ is a hard clustering. For simplicity, we will also represent C as the tuple $c = (1, 2, 2, 3, 1)$, where $c_i = j$ iff $C(x_i) = \omega_j$. One can equivalently describe C in terms of its relation representation:*

$$[C] = \{(x_1, x_5), (x_5, x_1), (x_2, x_3), (x_3, x_2), (x_1, x_1), (x_2, x_2), (x_3, x_3), (x_4, x_4), (x_5, x_5)\}.$$

As mentioned in Section 1, in soft clustering, objects are no longer restricted to fully belong to a single cluster: instead, they can partially belong to multiple clusters, where a partial assignment represents the uncertainty in the cluster assignments. Though different formalisms have been proposed to represent such uncertainty in a clustering, here we will focus on the general framework of evidential clustering, in which the uncertainty about the assignment of objects to clusters is represented in the form of a Dempster-Shafer mass function [19, 3, 47]. Formally, using the object-based representation, an evidential clustering is a set $M = \{m_x\}_{x \in X}$, where each m_x is a *mass function*, i.e., a function $m_x : 2^\Omega \mapsto [0, 1]$ such that $\sum_{A \subseteq \Omega} m_x(A) = 1$.

If the mass functions m_x are *logical* (i.e. they have only one focal set[3].), then the collection $R = \{m_x\}_{x \in X}$ is said to be a rough clustering. As in the case of hard clustering, multiple equivalent formalisms can be used to represent a rough clustering. Indeed, a rough clustering can be seen equivalently as a set-valued function $R : X \to 2^\Omega$, where $R(x) = A$

[3]We recall that A is a focal set if $m_x(A) > 0$.

with $m_x(A) = 1$. Similarly, a rough clustering can also be represented as a set of hard clusterings [6]. To this end, a hard clustering C is said to be *compatible* with R if $\forall x \in X$, it holds that $C(x) \in R(x)$: then, the rough clustering R can be equivalently understood as representing the collection of hard clusterings compatible with it, i.e., $C(R) = \{C : C \text{ is compatible with } R\}$.

Example 2.2. *Let X and Ω be defined as in Example 2.1. Then $M = \{m_{x_i}\}_{x_i \in X}$ defined as $m_{x_1}(\omega_1) = 1$; $m_{x_2}(\omega_2) = 1$; $m_{x_3}(\{\omega_2, \omega_3\}) = m_{x_3}(\Omega) = 0.5$; $m_{x_4}(\omega_2) = 1$; $m_{x_5}(\Omega) = 0.5, m_{x_5}(\omega_1) = m_{x_5}(\omega_2) = m_{x_5}(\omega_3) = \frac{1}{6}$ is an evidential clustering.*

Example 2.3. *Let X and Ω be defined as in Example 2.1. Then R defined by $R(x_1) = \omega_1$, $R(x_2) = \omega_2$, $R(x_3) = \{\omega_2, \omega_3\}$, $R(x_4) = \omega_3$, $R(x_5) = \Omega$ is a rough clustering. Using the tuple-based notation introduced in Example 2.1, R can be equivalently represented as the set*

$$C(R) = \{(1, 2, 3, 3, 3), (1, 2, 3, 3, 2), (1, 2, 3, 3, 1), (1, 2, 2, 3, 3), (1, 2, 2, 3, 2), (1, 2, 2, 3, 1)\}.$$

For simplicity, we can represent $C(R)$ as $(1, 2, \{2, 3\}, 3, \{1, 2, 3\})$.

In addition to hard and rough clusterings, evidential clustering also generalizes other forms of soft clustering. Indeed, when all the mass functions m_x are Bayesian (i.e., it holds that $\sum_{\omega \in \Omega} m_x(\{\omega\}) = 1$), then the collection $F = \{m_x\}_{x \in X}$ is a fuzzy clustering. Similarly, if all m_x are consonant (i.e., the focal sets of m_x are nested, that is $\forall A, B$ such that $m_x(A), m_x(B) > 0$, either $A \subseteq B$ or $B \subseteq A$), then the collection $P = \{m_x\}_{x \in X}$ is a possibilistic clustering. As in the case of rough clustering, both for fuzzy clustering and possibilistic clustering, an alternative and simpler representation can be given in terms of cluster membership vectors $F = \{\mu_x\}_{x \in X}$. In possibilistic clustering it is assumed that for all $x \in X$, $\max_{\omega \in \Omega} \mu_x(\omega) \leq 1$, while in fuzzy clustering we assume that for all $x \in X$, $\sum_{\omega \in \Omega} \mu_x(\omega) = 1$.

Example 2.4. *Let X and Ω be defined as in Example 2.1. Then,*

1. *F defined as $\mu_{x_1}(\omega_1) = 1$, $\mu_{x_2}(\omega_2) = 1$, $\mu_{x_3}(\omega_2) = \mu_{x_3}(\omega_3) = 0.5$, $\mu_{x_4}(\omega_3) = 1$ and $\mu_{x_5}(\omega_1) = \mu_{x_5}(\omega_2) = \mu_{x_5}(\omega_3) = \frac{1}{3}$ is a fuzzy clustering;*
2. *P defined as $\mu_{x_1}(\omega_1) = 1$, $\mu_{x_2}(\omega_2) = 1$, $\mu_{x_3}(\omega_2) = \mu_{x_3}(\omega_3) = 1$, $\mu_{x_4}(\omega_3) = 1$ and $\mu_{x_5}(\omega_1) = \mu_{x_5}(\omega_2) = 1, \mu_{x_5}(\omega_3) = 0.8$ is a possibilistic clustering.*

As for the case of hard clustering, we can define a relational representation also for the case of evidential clustering: we refer the reader to [9, 18] for additional details on this alternative representation of soft clusterings.

Finally, we note that, as we mentioned in Section 1, since a soft clustering describes the uncertainty in regard to an underlying (unknown) hard clustering, two forms of uncertainty can be distinguished. First, *partial assignment* (or *conflict*), i.e., the fact that for two sets of clusters $A, B \subset \Omega$ with $A \cap B = \emptyset$ it may happen that masses $m_x(A)$ and $m_x(B)$ are both positive: intuitively, partial assignment refers to the fact that the mass functions in the given evidential clustering encode evidence that is partially conflicting due to aleatory uncertainty

about the assignment of object x to the clusters. Second, *ambiguity*, i.e., the assignment of some mass to non-singleton events (that is, $\exists A \subseteq \Omega$ such that $|A| > 1$ and $m_x(A) > 0$): intuitively, ambiguity describes the inability to exactly determine to which cluster an object belongs, and can thus be considered as a way to represent epistemic uncertainty. It is easy to observe that in a fuzzy clustering only partial assignment is relevant, since all the mass is assigned to the singletons, while in the case of rough clustering, only ambiguity is present. By contrast, the evidential clustering formalism is flexible enough to represent both types of uncertainty.

2.2. Distributional Representation of Soft Clustering

As a way to more comprehensively represent the uncertainty underlying any given soft clustering, in [9] we proposed a novel representation of soft clustering, alternative to both the object-based and relational ones. This representation is based on the observation, illustrated above, that any rough clustering R can be represented as a set $C(R)$ of hard clusterings. Noting that such a set of hard clusterings can be interpreted as a Boolean possibility distribution over the space of hard clusterings, this representation can be extended to general soft clustering. The intuitive idea underlying this *distributional representation* is that any soft clustering can be understood as a belief function [12, 16, 37] over the collection of hard clusterings: depending on the specific class of soft clustering considered, a corresponding special class of belief functions is used as a representation formalism. Consequently, for example, a fuzzy clustering can be represented as a probability (i.e., a Bayesian belief function) distribution over hard clusterings, while a possibilistic clustering can be represented as a possibility distribution (i.e., a consonant belief function) over hard clustering. More generally, an evidential clustering can be interpreted as general mass function over hard clustering or, equivalently, a probability distribution over rough clusterings (as each rough clustering can be represented as a set of hard clusterings).

Formally, given an evidential clustering M, obtained by any evidential clustering algorithm applied to the data X, we consider the following probability distribution over rough clusterings:

$$m_M(R) = \prod_{x \in X} m_x(R(x)), \tag{1}$$

which can also be seen as a Dempster-Shafer mass function over hard clusterings. Intuitively, any evidential clustering M can be interpreted as a way to quantify the uncertainty about the assignment of instances to clusters: indeed, M associates with each $x \in X$ a mass function m_x, such that, for any set of clusters $A \subseteq \Omega$, $m_x(A)$ represents the evidence supporting the assignment of x to one of the clusters in A. Then, the distributional representation m_M of M transforms the above mentioned instance-wise uncertainty representation into a clustering-wise uncertainty representation: given any rough clustering R (i.e., a set of hard clusterings), $m_M(R)$ represents the evidence supporting the statement that the optimal clustering of the data lies among the hard clusterings in R. Given an evidential clustering M we denote by $\mathcal{F}(M)$ the collection of focal rough clusterings of M, that is $\mathcal{F}(M) = \{R : m_M(R) > 0\}$, where m_M is the distributional representation of M. We say that two evidential clusterings M_1, M_2 are *equivalent*, denoted $M_1 \sim_e M_2$, if

$\forall R_1 \in \mathcal{F}(M_1), R_2 \in \mathcal{F}(M_2)$, $\forall C_1 \in C(R_1), C_2 \in C(R_2)$ it holds that $C_1 \sim C_2$ (see Section 2.1). We denote with $\mathscr{M}_\Omega$ the set of all evidential clusterings with Ω being the set of clusters.

The distributional representation for rough, fuzzy and possibilistic clusterings can then be obtained as special cases of (1), by restricting the collection of focal rough clusterings. Indeed, in the case of rough clustering, m_M is logical and assigns all the evidence to a single rough clustering (i.e. $|\mathcal{F}(m_M)| = 1$). In the case of a fuzzy clustering $F = \{\mu_x\}_x$, where $\mu_x : \Omega \to [0, 1]$ is a probability distribution, the focal rough clusterings are all singletons (i.e., hard clusterings); we can, thus, define $Pr_F(C) = \prod_{x \in X} \mu_x(C(x))$. Finally, given a possibilistic clustering P and any t-norm $\wedge$, P can be represented as a possibility distribution over hard clusterings $\mathsf{Poss}_P(C) = \bigwedge_{x \in X} \mu_x(C(x))$, which corresponds to a consonant mass function over rough clusterings. We define $\mathscr{R}_\Omega$ (resp., $\mathscr{F}_\Omega$, $\mathscr{P}_\Omega$) by restricting $\mathscr{M}_\Omega$ to the set of rough (resp., fuzzy, possibilistic) clusterings.

Example 2.5. *Let F, P, M be the soft clusterings defined in Examples 2.4 and 2.2. Then, Pr_F is defined as:*

$$Pr_F((1,2,2,3,1)) = Pr_F((1,2,2,3,2)) = Pr_F((1,2,2,3,3)) = \\ Pr_F((1,2,3,3,1)) = Pr_F((1,2,3,3,2)) = Pr_F((1,2,3,3,3)) = \frac{1}{6}.$$

Similarly, $Poss_P$ is defined as

$$Poss_P((1,2,3,3,3)) = Poss_P((1,2,2,3,3)) = 0.8,$$
$$Poss_P((1,2,3,3,2)) = Poss_P((1,2,3,3,1)) = 1,$$
$$Poss_P((1,2,2,3,2)) = Poss_P((1,2,2,3,1)) = 1.$$

Finally, m_M is defined as

$$m_M((1,2,\{2,3\},3,\Omega)) = m_M((1,2,\Omega,3,\Omega)) = 0.25,$$
$$m_M((1,2,\{2,3\},3,1)) = m_M((1,2,\{2,3\},3,2)) = m_M((1,2,\{2,3\},3,3)) = \frac{1}{12},$$
$$m_M((1,2,\Omega,3,1)) = m_M((1,2,\Omega,3,2)) = m_M((1,2,\Omega,3,3)) = \frac{1}{12}.$$

Remark. *Let M be any evidential clustering. Even though the distributional representation m_M explicitly quantifies the uncertainty about clusterings compatible with M, it can also be understood as implicitly quantifying the uncertainty about* properties *of such compatible clusterings. A relevant example in this sense, would be to consider the* number of clusters*: indeed, m_M implicitly encodes information about the uncertainty concerning different possible values for the number of clusters in the unknown, true clustering of the data. For example, let $X = \{x_1, x_2, x_3\}$ and $\Omega = \{\omega_1, \omega_2, \omega_3\}$. Let M be the evidential clustering given by:*

$$m_{x_1}(\{\omega_1\}) = 1$$
$$m_{x_2}(\{\omega_2\}) = 1$$
$$m_{x_3}(\{\omega_1, \omega_2\}) = m_{x_3}(\Omega) = 0.5.$$

Then, the distributional representation m_M of M is given by:

$$m_M((1,2,\{1,2\})) = 0.5,$$
$$m_M((1,2,\Omega)) = 0.5.$$

It can be easily seen that we can compute the belief *and* plausibility *that the number of clusters is equal to 2 as follows:*

$$Bel(\textit{2 clusters}) = \sum_{R\in\mathcal{F}(M):\forall C\in R, C:X\to A \textit{ with } |A|=2} m_M(R) = m_M((1,2,\{1,2\})) = 0.5,$$
$$Pl(\textit{2 clusters}) = \sum_{R\in\mathcal{F}(M):\exists C\in R, C:X\to A \textit{ with } |A|=2} m_M(R) = m_M((1,2,\{1,2\})) + m_M((1,2,\Omega)) = 1.$$

Similarly, we could compute $Bel(1\ cluster) = Bel(3\ clusters) = 0$, as well as $Pl(1\ cluster) = 0$, $Pl(3\ clusters) = 0.5$.

2.3. Clustering Comparison Measures

Several measures have been defined to compare hard clusterings. For example, the widely used Rand index is defined, for any two hard clusterings C_1, C_2, as the proportion of object pairs that are either assigned to the same cluster by C_1 and C_2, or are assigned to different clusters by both C_1 and C_2, i.e.,

$$\mathsf{Rand}(C_1,C_2) = \frac{|\{(x,y)\in X^2 : (x,y)\in([C_1]\cap[C_2])\cup([C_1]^c\cap[C_2]^c)\}|}{|X|^2}. \tag{2}$$

The *partition distance* [11] is defined as the minimum number of objects to be moved to transform C_1 into C_2 (or, equivalently, C_2 into C_1); it can be computed as

$$d_\pi(C_1,C_2) = \frac{1}{|X|-1}\min_w \frac{1}{2}\sum_i |\omega_i^1 \Delta \omega_{w(i)}^2|, \tag{3}$$

where w is a permutation function, and Δ is the symmetric difference operator. Yet another clustering comparison measure is *variation of information* [43], defined as

$$\mathrm{VI}(C_1,C_2) = H(C_1) + H(C_2) - 2\sum_{\omega_i^1\in\Omega_1}\sum_{\omega_j^2\in\Omega_2} p_{ij}\log\frac{p_{ij}}{p_i^1\cdot p_j^2}, \tag{4}$$

where $p_i^1 = |\{x\in X : C_1(x) = \omega_i^1\}|/|X|$, and similarly for each $\omega_i^2\in\Omega_2$, while $p_{ij} = |\{x\in X : C_1(x) = \omega_i^1 \text{ and } C_2(x) = \omega_j^2\}|/|X|$, and $H(C_k) = \sum_{i=1}^{|\Omega|} p_i^k \log\frac{1}{p_i^k}$.

We remark that, even though $1-\mathsf{Rand}$, d_π and VI are defined as mappings in the form $d : \mathscr{C}_\Omega\times\mathscr{C}_\Omega\to\mathbb{R}$ (that is, as mappings that take as input two object-based representations of clusterings and return a real number), they can easily be seen to be metrics on $\mathscr{C}_\pi$: that

is, given two clusterings $C_1, C_2 : X \to \Omega$ with $C_1 \sim C_2$ (i.e., $[C_1] = [C_2]$), then it holds that $d(C_1, C_2) = 0$. Indeed, this is evident for the Rand index (as its definition explicitly refers to the relational representation of the clusterings to be compared), and it can be seen to hold true also for the partition distance and variation of information by noting that these two do not change when given two clusterings that are equivalent up to a relabeling of clusters. Notice that this property implies that $1 - \mathsf{Rand}$, d_π and VI are pseudo-metrics on $\mathscr{C}_\Omega$, but in general it holds that, given C_1, C_2, their value is equal to 0 iff $C_1 \sim C_2$. We refer to measures d that satisfy this property as *hard clustering metrics*.

Several extensions of the above mentioned measures has been proposed in the soft clustering setting, for fuzzy clustering [1, 2, 5, 10, 23, 24, 46], rough clustering [6, 20], possibilistic clustering [1, 2] and evidential clustering [18]: we refer the reader to [9] for a more comprehensive survey of previous works on the development of soft clustering evaluation measures. While such previous research was devoted to the definition and evaluation of specific evaluation measures, in [9] we proposed a general framework to extend any hard clustering validation measure to the setting of evidential clustering based on the distributional representation of soft clustering recalled in Section 2.2.

Before recalling this approach, we first describe some general requirements that a validation measure for soft clustering should meet. As briefly explained in the introduction, any such measure may have two different purposes:

1. On the one hand, it may aim to provide reliable bounds on the similarity (or, dually, the distance) between any given soft clustering and a ground truth partitioning of the data, taking into account the uncertainty represented by the soft clusterings to be compared. In the simplest case, wherein the given ground truth partitioning is a hard clustering, this corresponds to finding optimistic and pessimistic bounds on the similarity between the ground truth partitioning and the unknown, optimal hard clustering underlying the soft clustering to be compared. More generally, when also the ground truth partitioning is a soft clustering [4], this corresponds to bounding the similarity between the true, unknown hard clustering underlying the given soft ground truth and the soft clustering to be compared to it;

2. On the other hand, the objective may be to comprehensively characterize the uncertainty in the given soft clusterings and, henceforth, the uncertainty emerging from their comparison. That is, to comprehensively assess the similarities among all pairs of hard clusterings compatible with the two given soft clusterings and, most importantly, to lift the uncertainty represented in the soft clusterings to the results of such a comparison (i.e., to obtain a distribution over similarity values corresponding to the outcomes of the above comparison), so as to holistically evaluate the similarity among the clustering structures compatible with the two soft clusterings being compared.

While both approaches aim at providing an indication about the quality and information in a soft clustering, they are largely orthogonal and hence correspond to different properties

[4] Such a soft ground truth can arise, for example, in the context of semi-supervised learning, so that we evaluate a clustering algorithm based on a partial labelling of data.

that should be required for a given validation measure to satisfy that aim. In [9] we focused on the former of the above objectives, and remarked that it can be met by a pair of validation measures, representing an interval of values: the lower bound of the interval should quantify the compatibility between two soft clusterings to be compared, i.e. whether there exists a hard clustering which is compatible with both soft clusterings, while the upper bound should quantify their equality. To provide a formal translation of these principles, it is required that the lower bound be a consistency measure, while the upper bound should be a metric (or, dually, a similarity). Intuitively, the lower bound disregards the ambiguity in the soft clusterings to be compared and only focuses on the conflict among them, in the most optimistic case; by contrast, the upper bound equates ambiguity to an error or conflict and hence provides a more conservative validation assessment.

The approach proposed in [9] is based on the Wasserstein construction from Optimal Transport theory [42], which is used to compute a measure of distance between the distributional representations of the two soft clusterings to be compared: such measures are called *transport-theoretic measures*. To briefly recall this approach, we note that every evidential clustering can be represented as a distribution over rough clusterings. Let d be a normalized hard clustering metric. This measure is extended to the setting of rough clustering through the following pair of measures,

$$d_0^R(R_1, R_2) = d_R^l(R_1, R_2), \tag{5}$$

$$d_1^R(R_1, R_2) = d_H(C(R_1), C(R_2)), \tag{6}$$

where

$$d_R^l(R_1, R_2) = \min\{v \in \mathbb{R} : \exists C_1 \in C(R_1), C_2 \in C(R_2) \text{ such that } v = d(C_1, C_2)\},$$

and d_H is the Hausdorff distance between $C(R_1)$ and $C(R_2)$ based on d. Intuitively, d_0^R represents an optimistic lower bound on the distance between R_1, R_2, computed by taking the two hard clusterings C_1, C_2 (compatible with R_1, R_2) such that their distance is minimal. By contrast, d_1^R can be represented as the dual upper bound, under the constraint that the obtained measure is a metric (hence, the use of the Hausdorff distance rather than the max operator). These measures can then be extended to the case of evidential clustering by applying the Wasserstein construction as follows:

$$\begin{aligned} d_\alpha^M(M_1, M_2) = \min_\sigma & \sum_{(R_1,R_2)\in\mathcal{F}(M_1)\times\mathcal{F}(M_2)} \sigma(R_1, R_2) d_\alpha^R(R_1, R_2) \\ \text{such that} & \sum_{R_2\in\mathcal{F}(M_2)} \sigma(R_1, R_2) = m_{M_1}(R_1) \\ & \sum_{R_1\in\mathcal{F}(M_1)} \sigma(R_1, R_2) = m_{M_2}(R_2) \\ & \sum_{(R_1,R_2)\in\mathcal{F}(M_1)\times\mathcal{F}(M_2)} \sigma(R_1, R_2) = 1 \\ & \forall (R_1, R_2) \in \mathcal{F}(M_1) \times \mathcal{F}(M_2), \sigma(R_1, R_2) \geq 0 \end{aligned} \tag{7}$$

where $\alpha \in \{0, 1\}$ and σ is a probability distribution over $\mathcal{F}(M_1) \times \mathcal{F}(M_2)$. Intuitively, computing $d_\alpha^M(M_1, M_2)$ amounts to finding a joint distribution σ (indeed, the third and fourth constraint in Eq. (7) amount to requiring that σ is a probability distribution) whose marginals coincide with the two distributional representations of M_1, M_2 (as required by the first and second constraint in Eq. (7)) and such that the expected distance (computed as d_α^R, defined above) between rough clusterings compatible with M_1, M_2 is minimized. Then, d_0^M represents a lower bound on the distance between (hard clusterings compatible with) M_1 and M_2, while, dually, d_1^M represents an upper bound for the same comparison. Thus, the above construction allows to lift a base distance over rough clusterings to a distance over general distributions (i.e., mass functions) over hard clusterings.

In the following example, adapted from [9], we illustrate the computation of the transport-theoretic measures on the soft clusterings introduced in the previous examples.

Example 2.6. *Let C, F, P, M be the clusterings defined in Examples 2.1, 2.4, 2.2, and assume that $d = 1 - \mathsf{Rand}$, where Rand is the Rand index over hard clusterings. Then, it holds that*

- *For all $\alpha \in \{0, 1\}$, $d_\alpha^M(C, F) = 0.267$;*
- $d_0^M(C, P) = 0$, $d_1^M(C, P) = 0.48$;
- $d_0^M(C, M) = \frac{1}{12}$, $d_1^M(C, M) = 0.442$.

3. A General Framework for Soft Clustering Evaluation Measures

As shown in Section 2.3, most of the research on comparison measures for soft clustering has focused on the analysis of some specific indices [24, 18]. Furthermore, we have described a recent proposal for a general framework to extend validation and comparison measures to the setting of evidential clustering, aimed at addressing the need for measures that enable the objective comparison of two clusterings. In this section, we propose another framework aimed at the second goal described in the introduction, namely comprehensively describing the uncertainty in the assessment of similarity among a pair of soft clusterings and arising from their comparison.

3.1. Distributional Measures

Let d be a hard clustering metric[5]. Since, as shown in Section 2.2, any soft clustering can be seen as a distribution (in general, a mass function) over hard clusterings, an intuitive approach to comprehensively evaluate the uncertainty arising from the comparison of two soft clusterings would be to extend d to a distribution-valued function, providing a quantification

[5]As previously highlighted in Section 2.3, the fact that d is a hard clustering metric implies that the metric properties introduced in Section A are defined with respect to the equivalence relation $\sim$. Starting from this section, when we will study the metric properties of extensions of d to evidential clustering, we will assume metric properties to be defined with respect to the relation $\sim_e$.

of the uncertainty about the full range of similarities among the hard clusterings compatible with the two soft clusterings to be compared. The intuition behind this approach is based on the definition of soft clustering as representing a clustering with some uncertainty affecting our knowledge with respect to the assignment of objects to clusters, as described in Section 2.2. Thus, it is natural to require that an evaluation measure for soft clustering should transfer this uncertainty to the possible outcomes of the evaluation.

Therefore, a measure over rough clusterings would provide, given two rough clusterings R_1, R_2, a set of values, representing all possible distances between hard clusterings compatible with R_1, R_2. Similarly, given two fuzzy clusterings F_1, F_2, a measure would provide a probability distribution over possible distance values; while given two possibilistic clustering P_1, P_2 a measure would provide instead a possibility distribution over distance values. More generally, a measure over evidential clusterings would provide a mass function over possible values of d. Thus, the existence of a small distance value associated with a positive mass would denote the existence of a clustering structure compatible with both the soft clusterings being compared: thus, lower values of d could be interpreted as evidence that the above mentioned clustering structure shares some characteristics with the true, unknown clustering of the data. Formally, we define the *distributional measure*, based on d, between two rough clusterings as the set

$$d_R(R_1, R_2) = \{d(C_1, C_2) : C_1 \in C(R_1) \text{ and } C_2 \in C(R_2)\}, \tag{8}$$

and, for two general evidential clusterings (with fuzzy and possibilistic clustering as special cases), the mass function on $\mathbb{R}^+$ defined by[6]

$$\forall V \subset \mathbb{R}^+,\ d_M(M_1, M_2)(V) = \sum_{R_1, R_2 : d_R(R_1, R_2) = V} m_{M_1}(R_1) \cdot m_{M_2}(R_2), \tag{9}$$

It is easy to observe that d_M is a generalization of d_R. Indeed, in the specific case where M_1, M_2 are rough it holds that $d_M(M_1, M_2)(d_R(M_1, M_2)) = 1$. We also note that, when M_1, M_2 are fuzzy clusterings, the distributional measure d_M can be simplified to a probability distribution denoted by the symbol d_F, since all mass is assigned to singletons. In particular, the computation of d_F can be simplified as the probability distribution $d_F(F_1, F_2)$ on $\mathbb{R}^+$ defined by

$$\forall v \in \mathbb{R}^+,\ d_F(F_1, F_2)(v) = \sum_{C_1, C_2 : d(C_1, C_2) = v} Pr_{F_1}(C_1) \cdot Pr_{F_2}(C_2),$$

[6]We note that Eq. (9) resembles to the *conjuctive rule of combination* [38] (or, unnormalized Dempster's rule) $\textcircled{\cap}$ defined, for two mass functions m_1, m_2 as: $m_1 \textcircled{\cap} m_2(A) = \sum_{B \cap C = A} m_1(B) \cdot m_2(C)$. Indeed, d_M can be obtained by first computing the conjunctive combination of (the cylindrical extensions to $\mathcal{F}(M_1) \times \mathcal{F}(M_2)$ of) m_{M_1}, m_{M_1} as $m_{M_1} \textcircled{\cap} m_{M_2}(R_1, R_2) = m_{M_1}(R_1) \cdot m_{M_2}(R_2)$, and then computing the restriction of $m_{M_1} \textcircled{\cap} m_{M_2}$ to $\mathbb{R}$ given by $d_M(M_1, M_2)(V) = (m_{M_1} \textcircled{\cap} m_{M_2} \downarrow \mathbb{R})(V) = \sum_{R_1, R_2 : d_R(R_1, R_2) = V} m_{M_1} \textcircled{\cap} m_{M_2}(R_1, R_2)$. We note, furthermore, that in the given setting the conjuctive rule of combination coincides with Dempster's rule, that is $\textcircled{\cap} = \bigoplus$: indeed, it is easy to see that $d_M(M_1, M_2)(\emptyset) = 0$.

We also denote with d_P the possibility distribution on $\mathbb{R}^+$ obtained from d_M when the two evidential clusterings M_1, M_2 to be compared are possibilistic clusterings. It can be computed as

$$\forall v \in \mathbb{R}^+,\ d_P(P_1, P_2)(v) = \sum_{V \in \mathcal{F}(d_M(P_1,P_2)): v \in V} d_M(P_1, P_2)(V),$$

where $\mathcal{F}(d_M(P_1, P_2)) = \{V \subseteq \mathbb{R}^+ : d_M(P_1, P_2)(V) > 0\}$ is the collection of focal sets of the distributional measure $d_M(P_1, P_2)$.

Intuitively, $d_M(M_1, M_2)(V)$, where $V \subseteq \mathbb{R}^+$ is a set of distance values, can be interpreted as the mass of belief assigned to the statement "The true value of the distance between the two optimal hard clusterings underlying M_1, M_2 is within the set V". Therefore, d_M provides a complete representation of the possible distance values that arise when comparing hard clusterings compatible with M_1, M_2, obtained by transferring the mass functions over hard clustering defined by M_1, M_2 to a mass function over distance values.

From the point of view of its metric properties, it is clear that d_M is not a metric: indeed, d_M is not even defined as a single-valued function, but rather as a distribution, since its main aim is to quantify the uncertainty arising from the comparison between two evidential clusterings. However, d_M satisfies the properties stated in the following theorems.

Proposition 3.1. *Function d_M is symmetric, i.e., for any two evidential clusterings M_1 and M_2, $d_M(M_1, M_2) = d_M(M_2, M_1)$.*

Proof. The result directly follows from the definition of d_M. □

Theorem 3.1. *$d_M(M_1, M_2)(\{0\}) = 1$ iff $M_1 \sim_e M_2$, i.e., $\forall (R_1, R_2) \in \mathcal{F}(M_1) \times \mathcal{F}(M_2)$, $\forall C_1 \in C(R_1)$, $\forall C_2 \in C(R_2)$, it holds that $C' \sim C$. In particular, if M_1, M_2 are equivalent hard clusterings, then $d_M(M_1, M_2)(\{0\}) = 1$.*

Proof. The result directly follows from the definition of d_M. □

Theorem 3.2. *Once defined*

$$Bel_{M_1,M_2}(\{v\}) = \sum_{R_1,R_2: d_R(R_1,R_2)=\{v\}} m_{M_1}(R_1) \cdot m_{M_2}(R_2) \tag{10}$$

$$Pl_{M_1,M_2}(\{v\}) = \sum_{R_1,R_2: v \in d_R(R_1,R_2)} m_{M_1}(R_1) \cdot m_{M_2}(R_2). \tag{11}$$

then, it holds that

- *$\forall v \in \mathbb{R}$, the mappings $(M_1, M_2) \mapsto Bel_{M_1,M_2}(\{v\})$ and $(M_1, M_2) \mapsto Pl_{M_1,M_2}(\{v\})$ that assign to each pair of evidential clusterings (M_1, M_2) the corresponding belief and plausibility values, satisfy (M2) and (M3) (see Section A);*
- *$Bel_{M_1,M_2}(\{0\}) = 1$ iff $M_1 \sim_e M_2$;*

- $Pl_{M_1,M_2}(\{0\}) = 1$ *iff* $\forall (R_1, R_2) \in \mathcal{F}(M_1) \times \mathcal{F}(M_2)$, $\exists C_1 \in C(R_1), C_2 \in C(R_2)$ *such that* $C_1 \sim C_2$.

Proof. The result directly follows from the definition of d_M. □

Corollary 3.1. *Let M be an evidential clustering and C a hard clustering. Then, it holds that $d_M(M, C)(\{0\}) = 1$ iff $\forall R \in \mathcal{F}(M)$, $\exists C' \in C(R)$ such that $C' \sim C$. Moreover, it holds that*

$$d_M(M, C)(V) = \sum_{R \in \mathcal{F}(M): d_R(R,C) = V} m_M(R), \tag{12}$$

$$Bel_{M,C}(\{v\}) = \sum_{\{C'\} \in \mathcal{F}(M): d(C',C) = \{v\}} m_M(\{C'\}), \tag{13}$$

$$Pl_{M,C}(\{v\}) = \sum_{R \in \mathcal{F}(M): v \in d_R(R,C)} m_M(R). \tag{14}$$

As a consequence of the previous result, the value $d_M(M_1, M_2)(\{0\})$ can be interpreted as the mass of belief assigned to the hypothesis that the unknown optimal hard clustering structures underlying M_1 and M_2 are the same (have a distance equal to 0). Indeed, $d_M(M_1, M_2)$ assigns full belief to 0, if and only if M_1, M_2 are totally compatible. In particular, simple equality between M_1, M_2 does not suffice to obtain $d_M(M_1, M_2)(\{0\}) = 1$, unless M_1, M_2 are both hard clusterings. Thus, we see that even though d_M was not developed with the aim of enabling an objective comparison between the two evidential clusterings M_1, M_2, its restriction to the value 0 satisfies some reasonable metric properties that enable its use also in this application context, as a potential, more informative alternative to the transport-theoretic framework described in Section 2.3.

In regard to computational complexity, it is easy to show that computing d_M (resp., d_R, d_F, d_P) is computationally easy with respect to the size of the distributional representation introduced in the previous section. Indeed, the computation of the above measures simply requires one to enumerate the collection of compatible hard clusterings of the two soft clusterings to be compared, along with the respective mass function values. At the same time, it is easy to show that computing d_M (resp., d_R, d_F, d_P) is computationally intractable with respect to the size of both the object-based and relational representations (which are the representations most commonly adopted by soft clustering algorithms).

Theorem 3.3. *The problem of computing d_M (resp., d_R, d_F, d_P) has complexity $O(k^m)$, where $m = |\{x \in X : \exists R \in \mathcal{F}(M)$ such that $|R(x)| \neq 1\}|$ and $k = |\Omega|$ is the number of clusters. More precisely, d_R can be computed in constant amortized time (i.e., $d_R \in O(CAT(k, m))$), while d_F, d_P, d_M can be computed in at most linear amortized time (i.e. $d_M \in O(LAT(k, m))$).*

Proof. For the first part of the result, we note that, given a rough clustering R, the size of $C(R)$ is in the worst case exponential in the size of R [6]. Thus, the statement for d_R follows. Similar considerations can be applied to the cases of d_F, d_P and d_M. We note that

in all cases the size of the distributional representation is on the order of $O(k^m)$, therefore for this latter representation the problem of computing d_R (resp., d_F, d_P, d_M) has at most quadratic complexity with respect to the size of the input.

For the second part, we note that enumeration of all partitions of a set can be performed in constant amortized time [39], therefore the same holds for computing d_R. For the case of d_F, d_P, d_M, the same algorithm used for computing d_R can be used as a sub-routine. Specifically, in this case, for each hard clustering we also need to determine its mass. This can be easily performed in time $O(nk)$, which, assuming $n > k$, is linear in the size of the problem (in the worst case, where $n = k$, it is quadratic). □

3.2. Interval representation

In the previous section we introduced distributional measures as a comprehensive framework to evaluate the uncertainty arising from the comparison among two soft clusterings. Even though this framework satisfies some intuitively appealing properties, we have also shown that computing the distributional measures is computationally intractable, as it has complexity that is in general exponential in the size of the given soft clustering to be compared (although the complexity is polynomial in the size of their distributional representations). A possible solution to the above mentioned intractability would be to consider, instead of the complete description of the distributional measures, a compact representation (that is, a *summary*) of it, thus bridging the two aims for a clustering quality measure described above in the introduction.

For the case of rough clustering, it is easy to see that d_R can be summarized as the interval defined by the lower and upper bounds of d_R itself. That is:

$$\langle d^l_R, d^u_R\rangle(R_1, R_2) = \langle \min\{v \in \mathbb{R} : v \in d_R(R_1, R_2)\}, \max\{v \in \mathbb{R} : v \in d_R(R_1, R_2)\}\rangle.$$

We note that this definition satisfies the following properties:

Proposition 3.2. *Let R_1, R_2 be two rough clusterings. Then, $1 - d^l_R$ is a consistency on $\mathscr{R}_\Omega$: in particular $d^l_R(R_1, R_2) = 0$ iff $\exists(C_1, C_2) \in C(R_1) \times C(R_2)$ such that $C_1 \sim C_2$. By contrast, d^u_R satisfies (M1b) , (M2) , (M3) and $d^u_R(R_1, R_2) = 0$ iff $R_1 \sim_e R_2$.*

Proof. Clearly, d^l_R and d^u_R satisfy (M3). Furthermore, it is easy to show that d^l_R satisfies also (M1), while it fails to satisfy (M2). For the case of (M4), consider three rough clusterings R_1, R_2, R_3 such that $\exists C_1 \in C(R_1), C_2 \in C(R_2)$ with $C_1 \sim C_2$, $\exists C_2 \in C(R_2), C_3 \in C(R_3)$ with $C_2 \sim C_3$, while $\nexists C_1 \in C(R_1), C_3 \in C(R_3)$ with $C_1 \sim C_3$. Thus, d^l_R does not satisfy (M4). Claims for d^u_R similarly follow. □

Corollary 3.2. *d^u_R satisfies properties (M1), (M2), (M3) and (M4) iff either R_1 or R_2 is a hard clustering.*

Proof. One side of the implication directly derives from the observation that in case R_1 or R_2 is a hard clustering, d^u_R coincides with the Hausdorff distance between $C(R_1)$ and $C(R_2)$. On the other hand, assume d^u_R satisfies properties (M1), (M2), (M3) and (M4) on some subset of $\mathscr{R}_\Omega$. Then, in particular $d^u_R(R, R) = 0$. Thus, $d_R(R, R) = \{0\} \implies \forall C, C' \in C(R)$ it holds $C \sim C'$ and hence R is a hard clustering. □

As a result of the previous corollary, in the special case where the aim is to evaluate a rough clustering R against a hard clustering C which represents the ground truth partitioning of a data set, then d^u_R is guaranteed to be a metric, enabling the use of pair d^l_R, d^u_R also to the aim of objectively comparing the given rough clustering R with the ground truth hard clustering C. Despite this intuitively appealing property, it is easy to observe that computing $\langle d^l_R, d^u_R \rangle$ is still computationally hard:

Theorem 3.4. *Let R_1, R_2 be two rough clusterings, represented through the object-based representation. Then, the problem of computing $\langle d^l_R, d^u_R \rangle$ is* NP-HARD[7].

Proof. We note that the problems of computing d^l_R and d^u_R can be formulated as 0/1 programming problems. In particular, for the case of d^l_R, it holds that

$$d^l_R(R_1, R_2) = \min d(\{z^1_{ix}\}_{i,x}, \{z^1_{jx}\}_{j,x})$$

$$\begin{aligned}
\text{such that} \quad & \forall x \in X, \sum_{\omega^1_i \in R_1(x)} z^1_{ix} = 1 \\
& \forall x \in X, \sum_{\omega^1_i \notin R_1(x)} z^1_{ix} = 0 \\
& \forall x \in X, \sum_{\omega^2_j \in R_2(x)} z^2_{jx} = 1 \\
& \forall x \in X, \sum_{\omega^2_j \notin R_2(x)} z^2_{jx} = 0 \\
& \forall x \in X, \omega^1_i, z^1_{ix} \in \{0, 1\} \\
& \forall x \in X, \omega^2_j, z^2_{jx} \in \{0, 1\}
\end{aligned}$$

In general, the objective is not guaranteed to be linear: in any case, computational intractability of computing d^l_R follows from the general intractability of 0/1 programming [27]. A similar reduction is also applicable to d^u_R. □

For the cases of fuzzy, possibilistic and, more generally, evidential clustering, we can obtain a similar summarization by applying a decision rule to transform the distribution-valued d_F, d_P, d_M into simpler summary indices [14]. An example of this approach is to compute the following lower and upper expectations:

$$\underline{\mathbb{E}}(d_M)(M_1, M_2) = \sum_{V \subseteq 2^{\mathbb{R}}} \left[d_M(M_1, M_2)(V) \min_{d \in V} d \right] = \mathbb{E}(d^l_R), \tag{15}$$

$$\overline{\mathbb{E}}(d_M)(M_1, M_2) = \sum_{V \subseteq 2^{\mathbb{R}}} \left[d_M(M_1, M_2)(V) \max_{d \in V} d \right] = \mathbb{E}(d^u_R). \tag{16}$$

If M_1, M_2 are two fuzzy clusterings we obtain that $\underline{\mathbb{E}}(d_M) = \overline{\mathbb{E}}(d_M) = \mathbb{E}(d_F)$.

[7]The problem is trivially in P with respect to the distributional representation of R_1, R_2.

Example 3.1. *Let C, F, P, M be the soft clusterings defined in Examples 2.1, 2.4 and 2.2, and let $d = 1 - \mathsf{Rand}$. Then $\mathbb{E}(d_F(C,F)) = 0.267$, $\underline{\mathbb{E}}(d_P(C,P)) = 0, \overline{\mathbb{E}}(d_P(C,P)) = 0.48$, while $\underline{\mathbb{E}}(d_M(C,M)) = \frac{1}{12}, \overline{\mathbb{E}}(d_M(C,M)) = 0.442$. By contrast, if we let $d = d_\pi$, then $\mathbb{E}(d_F(C,F)) = 0.233$, $\underline{\mathbb{E}}(d_P(C,P)) = 0, \overline{\mathbb{E}}(d_P(C,P)) = 0.4$, while $\underline{\mathbb{E}}(d_M(C,M)) = 0.067, \overline{\mathbb{E}}(d_M(C,M)) = 0.367$.*

Similarly to the case of rough clustering, it is easy to show that the following properties hold:

Theorem 3.5. *Let M_1, M_2 be two evidential clusterings. Then $\overline{\mathbb{E}}(d_M)$ satisfies properties (M1b), (M2), (M3) and (M4), and $1 - \underline{\mathbb{E}}(d_M)$ is a consistency on $\mathscr{M}_\Omega$. In particular:*

- *If either M_1 or M_2 is a hard clustering, then $\underline{\mathbb{E}}(d_M)$ satisfies properties (M1) and (M3) and $\overline{\mathbb{E}}(d_M)$ satisfies properties (M1), (M2), (M3) and (M4);*
- *If F_1 and F_2 are two fuzzy clusterings, then $\mathbb{E}(d_F)$ satisfies properties (M1b), (M3) and (M4).*

Proof. Clearly, $\overline{\mathbb{E}}(d_M)$ satisfies (M2), (M3), (M4). Furthermore, it is easy to see that if either M_1 or M_2 is a hard clustering then $\overline{\mathbb{E}}(d_M)$ is equivalent to (7), which was proved to satisfy properties (M1), (M2), (M3) and (M4) in [9]. In particular $\overline{\mathbb{E}}(d_M) = 0$ iff M_1, M_2 are equivalent hard clusterings. For $\underline{\mathbb{E}}(d_M)$, the statement directly derives from Proposition 3.2. The remaining claims follow, respectively, from Corollary 3.2 and the definition of d_F. □

More generally, in the following section, we will show that some interesting relationships hold among the above defined summary indices and the transport-theoretic measures discussed in Section 2.3.

From the computational complexity point of view, it is easy to show that, in general, computing $\underline{\mathbb{E}}(d_M)$ and $\overline{\mathbb{E}}(d_M)$ is at least as hard as computing $\langle d_R^l, d_R^u \rangle$. However, for the case of fuzzy clustering and some base distances d, $\mathbb{E}(d_F)$ can be computed efficiently. Indeed, the computational hardness of computing $\underline{\mathbb{E}}(d_M), \overline{\mathbb{E}}(d_M)$ seems to stem from the ambiguity in the focal rough clusterings:

Proposition 3.3. *Let $d = 1 - \mathsf{Rand}$. Then, $\mathbb{E}(d_F)$ can be computed in time $O(n^2)$.*

Proof. From the definition of d_F, it is easy to show that $\mathbb{E}(d_F) = 1 - \mathsf{Rand}_F$, where Rand_F is the Rand index defined in [23]. □

We leave it as an open problem to characterize the general complexity of computing $\underline{\mathbb{E}}(d_M), \overline{\mathbb{E}}(d_M)$.

3.3. Relations between the Frameworks

It is interesting to study the connection between the distributional measures introduced in this paper, and the transport-theoretic measures introduced in [9]. In particular, in Examples 3.1 and 2.6 we showed that the distributional and transport-based measures provided the same results when one of the two clusterings to be compared was a hard clustering. The following theorem proves that this observation holds in general, as long as we compare a soft clustering with a hard clustering:

Theorem 3.6 ([9], Theorem 3.3)**.** *Let M, C be, respectively, an evidential clustering and a hard clustering. Then $d_0^M = \underline{\mathbb{E}}(d_M)$ and $d_1^M = \overline{\mathbb{E}}(d_M)$.*

Therefore, as a consequence of Theorem 3.6, even though the distributional and transport-based measures have different objectives, they are equivalent if our aim is to compare a soft clustering with a reference hard clustering, at least if we adopt lower and upper expectations as a summarization criterion for the distributional measures. In particular, this result also lends support to this choice, insofar as it allows us to obtain an objective comparison measure. However, we note that the equivalence between the two approaches does not hold in general, as will be shown in the following example, and the transport-based measures should be preferred when the aim is to obtain an objective comparison between two soft clusterings (one of which is assumed to be the ground truth).

Example 3.2. *Let M be the evidential clustering defined in Example 2.2. Then $\underline{\mathbb{E}}(d_M(M, M)) = 0.038$, and $\overline{\mathbb{E}}(d_M(M, M)) = 0.48$. By contrast, for all $\alpha \in \{0, 1\}$, it holds that $d_\alpha^M(M, M) = 0$.*

The following results provide a full characterization of the conditions under which the distributional and transport-theoretic measures coincide, as well as provide bounds for their difference when the conditions are not met:

Theorem 3.7. *Let M_1, M_2 be two evidential clusterings. Let σ^* be the joint distribution over $\mathcal{F}(M_1) \times \mathcal{F}(M_2)$ which minimizes Eq. (7). Then $d_0^M = \underline{\mathbb{E}}(d_M)$ iff one of the following conditions is met:*

1. *$\forall (R_1, R_2) \in \mathcal{F}(M_1) \times \mathcal{F}(M_2)$, $\exists (C_1, C_2) \in C(R_1) \times C(R_2)$ such that $C_1 \sim C_2$;*
2. *$\sigma^* = m_{M_1} \otimes m_{M_2}$, where $(m_{M_1} \otimes m_{M_2})(R_1, R_2) = m_{M_1}(R_1) \cdot m_{M_2}(R_2)$*

In all cases, it holds that $0 \leq \underline{\mathbb{E}}(d_M) - d_0^M \leq |F(M_1)| \cdot |F(M_2)| \cdot || \left[m_{M_1} \otimes m_{M_2}\right] ||$, where $|| \cdot ||$ is the Euclidean norm and $\forall A, B \in 2^\Omega \times 2^\Omega, \left[m_{M_1} \otimes m_{M_2}\right](A, B)$ is defined as

$$\left[m_{M_1} \otimes m_{M_2}\right](A, B) = \begin{cases} 1 - m_{M_1} \otimes m_{M_2}(A, B) & \text{if } m_{M_1} \otimes m_{M_2}(A, B) \leq 0.5 \\ m_{M_1} \otimes m_{M_2}(A, B) & \text{otherwise.} \end{cases} \tag{17}$$

Proof. Let M_1, M_2 be two evidential clusterings and m_{M_1}, m_{M_2} their respective distributional representations. We consider the finite-dimensional vector space $\mathcal{V} = \mathbb{R}^{2^\Omega \times 2^\Omega}$. Each mass function m_{M_i} is represented as a vector in the probability simplex of dimensionality 2^Ω and support $\mathcal{F}(M_i)$. Similarly, we can represent the values of d_0^R as a vector that lies in a $\mathcal{F}(M_1) \times \mathcal{F}(M_2)$-dimensional subspace of $\mathcal{V}$. Then, both d_0^M and $\underline{\mathbb{E}}(d_M)$ can be represented as inner products on $\mathcal{V}$, in particular:

$$d_0^M(M_1, M_2) = \langle \sigma^*, d_0^R \rangle \tag{18}$$

$$\underline{\mathbb{E}}(d_M(M_1, M_2)) = \langle m_{M_1} \otimes m_{M_2}, d_0^R \rangle, \tag{19}$$

where we note that $m_{M_1} \otimes m_{M_2}$, defined as in the statement of the theorem, is the tensor product between m_{M_1}, m_{M_2}.

It is easy to observe that $\underline{\mathbb{E}}(d_M(M_1, M_2)) \geq d_0^M(M_1, M_2)$, thus it holds that $\langle m_{M_1} \otimes m_{M_2}, d_0^R \rangle - \min_\sigma \langle \sigma, d_0^R \rangle \geq 0$. Hence, due to linearity of the inner product, it follows that $\langle m_{M_1} \otimes m_{M_2} - \sigma^*, d_0^R \rangle \geq 0$, where σ^* is defined as in the theorem statement. Due to Cauchy-Schwarz inequality [40], it follows that

$$0 \leq \underline{\mathbb{E}}(d_M(M_1, M_2)) - d_0^M(M_1, M_2) \leq ||m_{M_1} \otimes m_{M_2} - \sigma^*|| \cdot ||d_0^R|| \tag{20}$$

Consequently, $\underline{\mathbb{E}}(d_M(M_1, M_2)) = d_0^M(M_1, M_2)$ iff either $m_{M_1} \otimes m_{M_2} = \sigma^*$ (which is Condition 2 in the theorem) or $||d_0^R|| = 0$: this happens iff Condition 1 in the theorem is satisfied (in which case, $\forall R_1 \in \mathcal{F}(M_1), R_2 \in \mathcal{F}(M_2)$ it holds that $d_0^R(R_1, R_2) = 0$).

The upper bound on the value of $\underline{\mathbb{E}}(d_M(M_1, M_2)) - d_0^M(M_1, M_2)$ follows from above by noting that $||d_0^R|| \leq |\mathcal{F}(M_1) \times \mathcal{F}(M_2)|$ (in particular, it is equal when $\forall R_1 \in \mathcal{F}(M_1), R_2 \in \mathcal{F}(M_2)$ it holds that $d_0^R(R_1, R_2) = 1$) and bounding the value of $||m_{M_1} \otimes m_{M_2} - \sigma^*||$. □

Theorem 3.8. *Let M_1, M_2 be two evidential clusterings. Let σ^* be the joint measure over $\mathcal{F}(M_1) \times \mathcal{F}(M_2)$ which minimizes Eq. (7). Then $d_1^M = \overline{\mathbb{E}}(d_M)$ iff at least one of M_1 and M_2 is a hard clustering.*

Proof. We have already proved that if at least one of M_1, M_2 is a hard clustering then $d_1^M = \overline{\mathbb{E}}(d_M)$. For the converse statement, following a similar argument as to Theorem 3.7 it can be shown that

$$0 \leq \overline{\mathbb{E}}(d_M) - d_1^M = \langle m_{M_1} \otimes m_{M_2}, d_R^u \rangle - \langle \sigma^*, d_1^R \rangle.$$

Since $d_R^u \geq d_1^R$, the rightmost expression is equivalent to

$$\langle m_{M_1} \otimes m_{M_2}, d_R^u - d_1^R \rangle + \langle m_{M_1} \otimes m_{M_2}, d_1^R \rangle - \langle \sigma^*, d_1^R \rangle,$$

which, in turn, is equivalent to

$$\langle m_{M_1} \otimes m_{M_2}, d_R^u - d_1^R \rangle + \langle m_{M_1} \otimes m_{M_2} - \sigma^*, d_1^R \rangle.$$

Therefore, by applying Cauchy-Schwarz inequality to the two inner product terms above, it follows that

$$0 \leq \overline{\mathbb{E}}(d_M) - d_1^M \leq ||m_{M_1} \otimes m_{M_2} - \sigma^*|| \cdot ||d_1^R|| + ||m_{M_1} \otimes m_{M_2}|| \cdot ||d_R^u - d_1^R||. \tag{21}$$

Then, it easily follows that for the above expression to be equal to 0 it is required that one of the following holds:

- $||m_1 \otimes m_2 - \sigma^*|| = 0$ and $||d_R^u - d_1^R|| = 0$, which only holds if at least one of M_1, M_2 is a hard clustering;
- $||d_R^u|| = ||d_1^R|| = 0$, which in turn implies that M_1, M_2 are equivalent hard clusterings.

Hence, the theorem follows. □

Thus, as a consequence of the previous results we know that the upper bounds for the distributional and transport-theoretic measures coincide if and only if one of the two clusterings to be compared is a hard clustering. By contrast, the criteria for equivalence between the corresponding lower bounds are much weaker: indeed, it suffices that either the two soft clusterings to be compared are partially compatible (i.e., they are not defined on totally distinct bodies of evidence) or that the solution of the optimal transport problem coincides with the element-wise product of the corresponding distributional measures. As we will show in the next section, these equivalences enable us to use efficient approximation algorithms for the distributional measures also to approximate the transport-theoretic ones.

4. Approximation Methods

In the previous section we proposed distributional measures as a general approach to extend any hard clustering comparison measure to a soft clustering comparison measure. Nonetheless, the computation of these distributional measures is, in general, intractable. For this reason, in this section, we introduce some approximation methods and algorithms, based on a sampling approach, which can be applied to any base distance between hard clusterings.

We start with the case of the summarized representation of d_R, that is, with d_R^l, d_R^u. Given two rough clusterings R_1, R_2, we draw s independent samples $(C_1^1, C_2^1), \ldots, (C_1^s, C_2^s)$ uniformly from $C(R_1), C(R_2)$. Then, we can approximate d_R^l and d_R^u as, respectively, $\hat{d}_R^l = \min_{i\in\{1,\ldots,s\}} d(C_1^i, C_2^i)$ and $\hat{d}_R^u = \max_{i\in\{1,\ldots,s\}} d(C_1^i, C_2^i)$. The algorithm, for the case of d_R^l, is illustrated in Algorithm 1.

Clearly, the following result holds:

Proposition 4.1. *The following bounds hold for any $\epsilon > 0$:*

$$Pr(d_R^u - \hat{d}_R^u > \epsilon) \leq F(d_R^u - \epsilon)^s, \quad Pr(\hat{d}_R^l - d_R^l > \epsilon) \leq \left(1 - F(\epsilon - d_R^l)\right)^s \tag{22}$$

where F is the cumulative distribution function (CDF) of the probability distribution p_R defined as $p_R(t) = \frac{|\{C_1\in C(R_1), C_2\in C(R_2): d(C_1,C_2)=t\}|}{|d_R(R_1,R_2)|}$.

Furthermore, let $\hat{F}$ be the empirical estimator of F obtained from Algorithm 1. Then, for any $\alpha \in (0,1)$, it holds that

$$\left[\hat{d}_R^u, \min\left\{1, \hat{d}_R^u + \left(1 - \hat{F}^{-1}\left(\sqrt[s]{\alpha}\right)\right)\right\}\right]$$

is an asymptotic, one-sided, $1-\alpha$ confidence interval for d_R^u. Similarly, it holds that

$$\left[\max\left\{0, \hat{d}_R^l - \hat{F}^{-1}\left(1 - \sqrt[s]{\alpha}\right)\right\}, \hat{d}_R^l\right]$$

is an asymptotic, one-sided, $1-\alpha$ confidence interval for d_R^l.

Algorithm 1 The procedure to approximate the lower bound for the distributional measure of rough clustering through sampling.

procedure ROUGH_DISTRIBUTIONAL_SAMPLING(R_1: rough clustering, R_2: rough clustering, s: number of samples, d: normalized measure)
  $min \leftarrow \infty$
  for all iterations $it = 1$ to s **do**
    $C_1 \leftarrow \emptyset$
    for all $x \in X$ **do**
      $C_1(x) \leftarrow Uniform(R_1(x))$
      $C_2(x) \leftarrow Uniform(R_2(x))$
    end for
    $val \leftarrow d(C_1, C_2)$
    if $val \leq min$ **then**
      $min \leftarrow val$
    end if
  end for
  return min
end procedure

Proof. The first result follows from the distribution of the order statistics $\hat{d}_R^l, \hat{d}_R^u$. The form of the confidence intervals derives from the first statement and by applying Dvoretzky-Kiefer-Wolfowitz inequality [29] to F. In particular, it holds that

$$Pr(\sup_{t\in\mathbb{R}} |\hat{F}(t) - F(t)| > \delta) \leq 2e^{-2s\delta^2},$$

hence the asymptotic convergence rate is exponential in the sample size s. □

Since for each ϵ, the quantity $F(d_R^u - \epsilon)$ (resp., $F(\epsilon - d_R^u)$) is stricly less than 1, it holds that $Pr(d_R^u - \hat{d}_R^u > \epsilon)$ (resp., $Pr(\hat{d}_R^l - d_R^l > \epsilon)$) decays exponentially with respect to the sample size s. However, we note that the quality of the previously described approximation method largely depends on d_R. In particular, the convergence in (22) is influenced by the *tailedness* of p_R: the heavier the tails of p_R, the lower the approximation error.

For the case of fuzzy clustering, if we use the expected value $\mathbb{E}(d_F)$ to summarize d_F and we use a sampling procedure to estimate $\mathbb{E}(d_F)$ as $\hat{d}_F$ then we can obtain a tail bound by applying Hoeffding's inequality:

Proposition 4.2. *Assume that d is a normalized hard clustering metric. Then:*

$$Pr(|\hat{d}_F - \mathbb{E}(d_F)| \geq \epsilon) \leq 2e^{-2s\epsilon^2} \tag{23}$$

Hence, the deviation between the empirical mean $\hat{d}_F$ and $\mathbb{E}(d_F)$ has exponential decay in the sample size s. Furthermore, for any $\alpha \in (0, 1)$,

$$\hat{d}_F \pm \sqrt{\frac{1}{2s} \log\left(\frac{2}{\alpha}\right)}$$

is a $1-\alpha$ *confidence interval for* $\mathbb{E}(d_F)$.

Algorithm 2 The procedure to approximate the bounds for the distributional measure of evidential clustering through sampling.

procedure EVIDENTIAL_DISTRIBUTIONAL_SAMPLING(M_1: rough clustering, M_2: rough clustering, s: number of samples, d: normalized measure, $i \in \{l, u\}$)
  $res \leftarrow \infty$
  for all iterations $it = 1$ to s **do**
    $R_1 \leftarrow \emptyset$
    $R_2 \leftarrow \emptyset$
    for all $x \in X$ **do**
      $R_1(x) \leftarrow R \sim m_x^{M_1}$
      $R_2(x) \leftarrow R \sim m_x^{M_2}$
    end for
    $val \leftarrow d_R^i(R_1, R_2)$
    $res \leftarrow res + val$
  end for
  return $\frac{res}{s}$
end procedure

A similar result holds also for d_M, using a sampling approach as defined in Algorithm 2:

Proposition 4.3. *Assume that* d *is a normalized hard clustering metric. Let* $\hat{d}_M^l, \hat{d}_M^u$ *be the sample estimates of* $\underline{\mathbb{E}}(d_M), \overline{\mathbb{E}}(d_M)$. *Then:*

$$Pr(|\hat{d}_M^l - \underline{\mathbb{E}}(d_M)| \geq \epsilon) \leq 2e^{-2s\epsilon^2}, \quad Pr(|\hat{d}_M^u - \overline{\mathbb{E}}(d_M)| \geq \epsilon) \leq 2e^{-2s\epsilon^2}. \tag{24}$$

Furthermore, it holds that

$$\hat{d}_M^l \pm \sqrt{\frac{1}{2s}\log\left(\frac{2}{\alpha}\right)}$$

is a $1-\alpha$ *confidence interval for* $\underline{\mathbb{E}}(d_M)$. *The same holds for* $\hat{d}_M^u$ *and* $\overline{\mathbb{E}}(d_M)$.

Proof. The result directly derives from an application of Hoeffding's inequality. □

Given two evidential clusterings M_1, M_2, the previous estimate requires that $\hat{d}_M^l, \hat{d}_M^u$ are computed by sampling pairs R_1, R_2 of rough clusterings from the distributions m_{M_1}, m_{M_2} and then computing the exact values of $d_R^l(R_1, R_2)$ and $d_R^u(R_1, R_2)$. As a consequence of Proposition 3.4, this may not be feasible when $|X|$ is large. In such cases, a possible solution would be to compute $\hat{d}_M^l, \hat{d}_M^u$ by means of nested sampling (i.e, first we sample a rough clustering R from m_M, then we sample a hard clustering C from $C(R)$), as described in Algorithm 3. In this case, however, one should expect a larger approximation error, as described by the following theorem:

Algorithm 3 The procedure to approximate the bounds for the distributional measure of evidential clustering through nested sampling.

procedure EVIDENTIAL_NESTED_SAMPLING(M_1: rough clustering, M_2: rough clustering, s_i: inner samples, s_o: outer samples, d: normalized measure, $i \in \{l, u\}$)
  $res \leftarrow \infty$
  for all iterations $it = 1$ to s_o **do**
    $R_1 \leftarrow \emptyset$
    $R_2 \leftarrow \emptyset$
    for all $x \in X$ **do**
      $R_1(x) \leftarrow R \sim m_x^{M_1}$
      $R_2(x) \leftarrow R \sim m_x^{M_2}$
    end for
    $val_{it} \leftarrow rough\_distributional\_sampling(R_1, R_2, s_i, d)$
    $res \leftarrow res + val_{it}$
  end for
  return $\frac{res}{s}$
end procedure

Theorem 4.1. *Let F be the cumulative distribution function (CDF) of the probability distribution p_M defined as*

$$p_M(t) = Pr_{R_1 \sim m_{M_1}, R_2 \sim m_{M_2}} \frac{|\{C_1 \in C(R_1), C_2 \in C(R_2) : d(C_1, C_2) = t\}|}{|d_R(R_1, R_2)|}.$$

Let s_i^u be such that $s_i^u \geq \frac{\log(\frac{\delta}{s_o})}{\log(F(d_R^u - \frac{\epsilon}{2s_o}))}$. Similarly, let s_i^l be such that $s_i^l \geq \frac{\log(\frac{\delta}{s_o})}{\log(1 - F(\frac{\epsilon}{2s_o} - d_R^l))}$. Then, if $\hat{d}_M^l, \hat{d}_M^u$ are the sample estimates of $\underline{\mathbb{E}}(d_M), \overline{\mathbb{E}}(d_M)$ computed through Algorithm 3, it holds with probability greater than $1 - \delta$ that

$$Pr(|\hat{d}_M^l - \underline{\mathbb{E}}(d_M)| \geq \epsilon) \leq 2e^{-8s_o\epsilon^2}, \quad Pr(|\hat{d}_M^u - \overline{\mathbb{E}}(d_M)| \geq \epsilon) \leq 2e^{-8s_o\epsilon^2}. \tag{25}$$

Furthermore, under the same assumptions as above, it holds that

$$\hat{d}_M^l \pm \sqrt{\frac{1}{8s_o} \log\left(\frac{2}{\alpha}\right)}$$

is an asymptotic $1-\alpha$ confidence interval for $\underline{\mathbb{E}}(d_M)$. The same holds for $\hat{d}_M^u$ and $\overline{\mathbb{E}}(d_M)$.

Proof. If s_i^l, s_i^u are selected as in the statement of the theorem then, by Proposition 4.1, it holds that, with probability greater than $1 - \frac{\delta}{s_o}$ over the sampling of rough clusterings R_1, R_2 from the given evidential clusterings M_1, M_2, $|val - d_R^j(R_1, R_2)| \leq \frac{\epsilon}{2s_o}$, where val is defined as in Algorithm 1 and $j \in \{l, u\}$. In particular, this implies that, with probability greater than $1 - \delta$, all of the iterations of the inner for loop in Algorithm 3 will have an error smaller than

$\frac{\epsilon}{2s_o}$. Thus, $Pr(|\hat{d}^l_M - \underline{\mathbb{E}}(d_M)| \geq \frac{\epsilon}{2}) \leq Pr(\frac{1}{s_o}\sum_{i=1} s_o|val_i - d^l_R(i)| + |\sum_{i=1}^{s_o} d^l_R(i) - \underline{\mathbb{E}}(d_M)| \geq \frac{\epsilon}{2}) \leq Pr(|\sum_{i=1}^{s_o} d^l_R(i) - \underline{\mathbb{E}}(d_M)| \geq \epsilon) \leq 2e^{-8s_o\epsilon^2}$. Similarly, the result holds also for the case of $\hat{d}^u_M, \overline{\mathbb{E}}(d_M)$ and the theorem follows. □

Also, it is easy to show that Algorithms 1 and 3 are computable in polynomial time:

Proposition 4.4. *Let T_d be the complexity of the base measure d. Then, the complexity of Algorithm 1 is $O\left(s(|X||\Omega| + T_d)\right)$. Similarly, the time complexity of Algorithm 3 is $O\left(s_o(|X|2^{|\Omega|} + s_i(|X||\Omega| + T_d))\right)$. In all cases, if $|\Omega| \in O(\log|X|)$, then the above algorithms have time complexity polynomial in $|X|$.*

Finally, we conclude with a remark on the significance of the proposed sampling approximation methods and their theoretical guarantees. Obviously, these methods enable us to efficiently and effectively approximate the values of the interval representation of a distributional measure. Indeed, the above tail bound guarantees ensure that the speed of convergence of these approximations to the corresponding exact results is almost exponential: such a speed-up with respect to the exact interval representaion of distributional measures is particularly significant since Algorithms 1, 2 and 3 can all be trivially implemented as parallel algorithms (indeed, it is easy to see that the approximation methods are embarassingly parallel). Furthermore, the above theoretical results concerning the convergence of the approximation methods provide a way to obtain confidence intervals for interval representation of a distributional measure that will contain, with high probability, the exact value of the latter statistic. Most relevantly, however, in light of the results in Section 3.3, we also know that when the conditions for equivalence between the (interval representations of the) distributional and transport-theoretic measures hold (in particular, when one of the two clusterings to be compared is hard), the sampling approaches presented in this section enable also to approximate the values of the transport-theoretic measures. This is particularly significant because the transport-theoretic measures are computationally hard to compute in their exact form [9]. Though in [9] we presented some approximation methods for specific evaluation measures (namely, the Rand index and the partition distance), we notice that the approaches presented here have the advantage of being fully general (i.e., they can be applied irrespective of the base clustering evaluation measure) and satisfying strong convergence guarantees.

5. Illustrative Experiments

In this section we discuss two sets of experiments, with the aim of illustrating the application of the proposed sampling-based approximations for the distributional measures, as well as to evaluate their accuracy and convergence speed. In particular, in Section 5.1 we will analyze the accuracy and speed with which the sampling-based approximations converge to the exact values of the (interval representations of the) distributional measures. By contrast, in Section 5.2 we will illustrate the applicability of the sampling-based approximations on a collection of real-world datasets.

In both experiments, we considered five different clustering algorithms, namely: k-means (KM), rough k-means (RKM) [31], fuzzy c-means (FCM) [4], possibilistic c-means (PCM) [26] and evidential c-means (ECM) [28]. In regard to the algorithm hyper-parameters, since our goal was only to illustrate the applications of our metrics rather than to provide a bias-free comparison of the considered algorithms, we considered the default values as defined in the corresponding articles. Code (for both the clustering algorithms, as well as the evaluation measures) was implemented in Python (v. 3.8.8), using scikit-learn (v. 0.24.1), numpy (v. 1.20.1) and scipy (v. 1.6.2). The code for all the metrics and implementations of clustering algorithms is publicly available on GitHub[8].

5.1. Convergence Analysis

In Section 4 we studied the sampling-based approximations for the distributional measures, as a way to side-step the high computational complexity of computing the corresponding exact interval representations (indeed, in Section 3.2 we showed that for all forms of soft clustering, computing the interval representation of the distributional measures is NP-hard). To this aim, we studied the converge of the sampling-based approximation from a theoretical point of view, using tools from concentration inequality theory. However, while these results provide evidence that the sampling-based approximations converge exponentially-fast to their exact versions, they do not provide any practical indication about the selection of the number of samples to draw, as well as (for the case of possibilistic and evidential clustering) about how to select a trade-off between the number of outer and inner samples (see Algorithm 3). To address these limitations, in this section we study the convergence properties of the sampling-based approximations from an empirical point of view.

To this aim, we considered two small synthetic datasets. In each of the two datasets we considered a sample of 50 objects and 2 different clusters: the dimensionality of the dataset was selected so that the computation of the exact versions of the interval representations for the distributional measures could be performed in a reasonable amount of time. The two datasets differ with respect to the degree of overlap between the two clusters: in the first case (no-noise), the clusters have a very limited overlap (the two clusters were generated using the *make_blob* function from the scikit-learn library, with two centers and cluster standard deviation equal to 1.25); by contrast, in the second synthetic dataset (noisy), the two clusters exhibit a significantly larger degree of overlap (the cluster standard deviation was set equal to 2.5). In both cases, we evaluated the degree of error (measured as the absolute difference between the approximated and exact values of the distributional Rand index, i.e. $|Exact(Rand) - Approx.(Rand)|$) and its variation with increasing number of iterations of the sampling-based procedures: for the case of RKM we considered Algorithm 1; for the case of FCM we considered a simple implementation of Monte Carlo sampling; while for the case of PCM and ECM we considered the nested sampling procedure described in Algorithm 3. In all cases, we varied the number of iterations (for PCM and ECM, outer samples) between 10 and 1000; for the case of PCM and ECM we also varied the number of inner samples, with values in $\{1, 5, 10, 50, 100\}$. The results for the no-noise synthetic

[8] `https://github.com/AndreaCampagner/scikit-cautious`

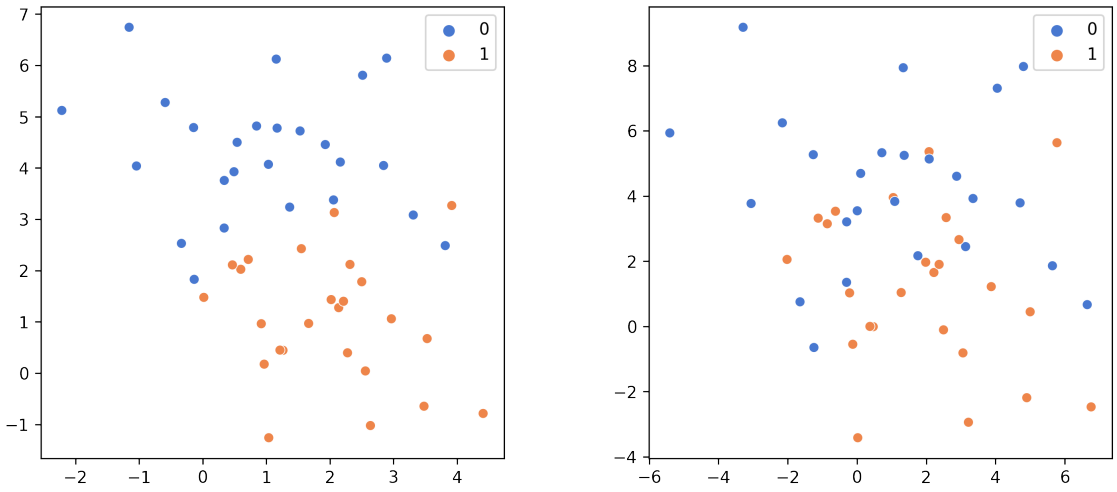

Figure 1: The two synthetic datasets for the converge analysis experiment: (left) the dataset with minimal cluster overlap; (right) the dataset with larger cluster overlap.

dataset, in terms of lower and upper bounds of the distributional measures, are depicted in Figures 2 and 3, for RKM and FCM, and for PCM and ECM, respectively; the results for the noisy synthetic dataset, in terms of lower and upper bounds of the distributional measures, are depicted in Figures 4, for RKM and FCM, and 5, for PCM and ECM.

We can observe that, for the case of rough and fuzzy clustering, the convergence was very rapid, for both the noisy and well separated datasets. In both cases, RKM converged to almost zero error already with 10 iterations of the sampling-based procedure, similarly also FCM exhibited a very rapid convergence to the exact value of the evaluation metrics: in the worst case (which was the approximation of the lower bound of distributional Rand index for the noisy dataset), the error exponentially decreased below 5×10^{-2} in fewer than 100 iterations. By contrast, both the convergence of both PCM and ECM strongly depended on the number of inner samples considered in the nested sampling procedure. When using less than 50 inner samples, the sampling-based approximation did not converge and, after an exponential decrease phase, reached a plateau (whose value decreased with the increasing number of inner samples). By contrast, when using a number of inner samples equal or greater to 50 (and, in particular, in the case of 100 inner samples) the sampling-based approximations for possibilistic and evidential clustering exhibited a convergence pattern similar to that observed in fuzzy clustering. The lack of convergence observed in the case of possibilistic and evidential clustering, when using a small number of inner samples, likely stems from the fact that the nested sampling procedure in Algorithm 3 needs to approximate the pairwise distance between the focal rough sets of the two soft clustering to be compared: if the number of inner samples for this approximation is not sufficiently high, with high probability the internal sampling estimates will significantly deviate from each other, thus

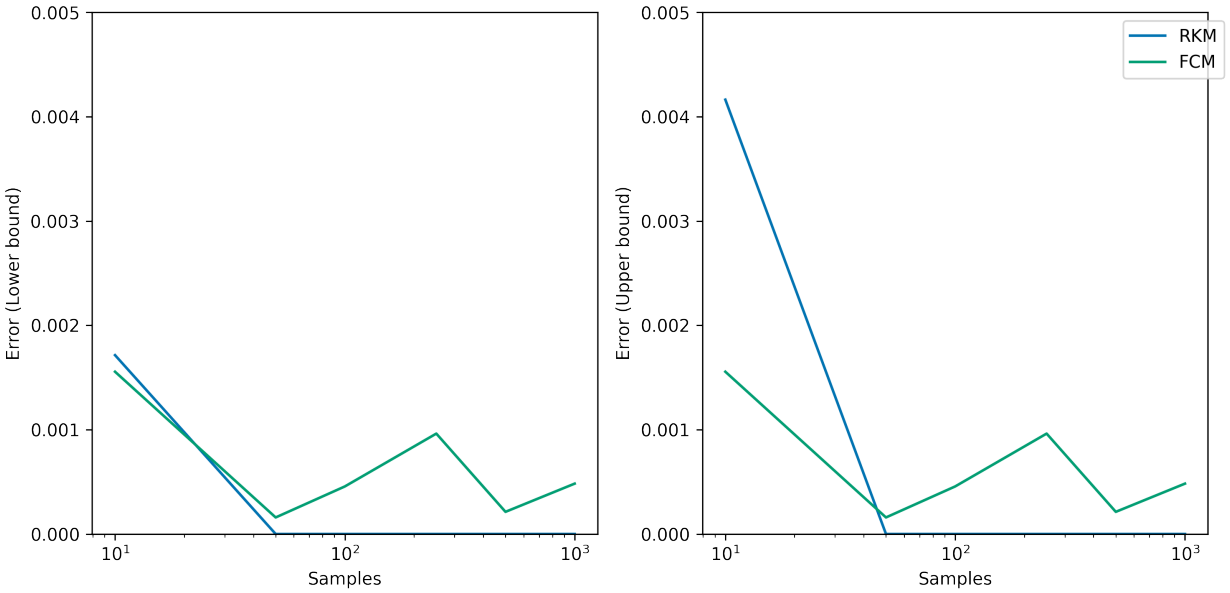

Figure 2: Results of the convergence analysis for the no-noise dataset, for rough clustering and fuzzy clustering.

preventing convergence of the sampling procedure. In fact, also from the theoretical point of view, the convergence result for Algorithm 3 (see Theorem 4.1), which was used for possibilistic and evidential clustering, is much weaker than the corresponding convergence results that hold for the sampling procedures for rough and fuzzy clustering. Indeed, in the former case convergence and error reduction of the sampling procedure can be guaranteed only if the number of inner samples is sufficiently high so as to reach a desirable approximation quality for the inner sampling procedure. Interestingly, however, we note that, for both possibilistic and evidential clustering, the sampling approaches reported a much lower approximation error and more rapid convergence for the upper bound of the distributional measure, as compared with the corresponding lower bound. We plan to further investigate this observation in future work.

The observed results provide some practical indications for the application of the sampling-based approximations in real-world settings:

- For both rough clustering and fuzzy clustering, the sampling-based approximations seem to provide good convergence speed and, even with a small number of samples, can be used as a good proxy for the exact values of the distributional measure, hence enabling its approximation;

- The approximation quality of the nested sampling procedure for possibilistic and evidential clustering is strongly dependent on the number of inner samples: a small number of inner samples may impede convergence, even when a large number of outer samples are used. Thus, when a fixed computational budget is available, inner samples should be preferred to outer samples when executing Algorithm 3;

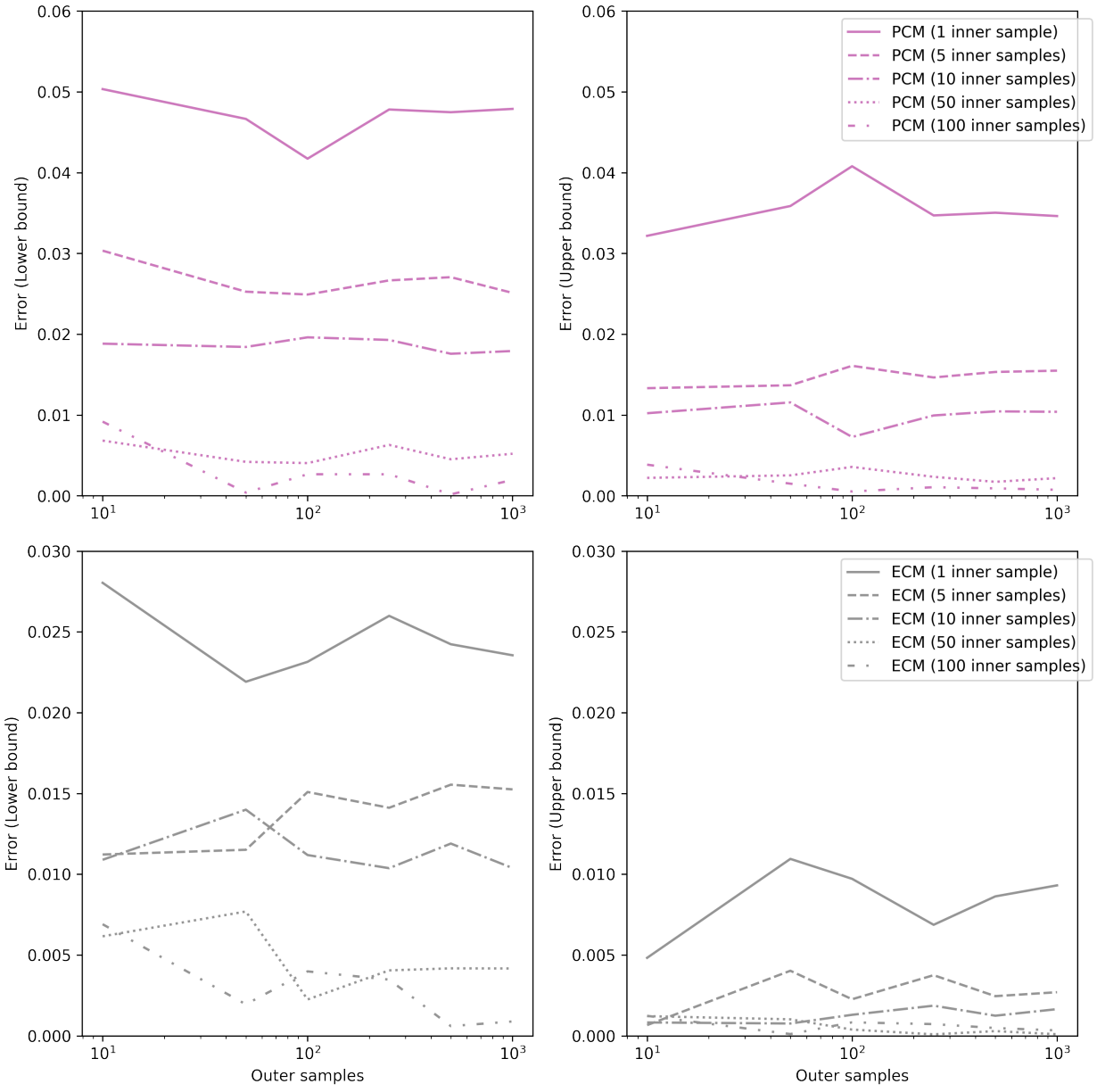

Figure 3: Results of the convergence analysis for the no-noise dataset, for possibilistic clustering and evidential clustering.

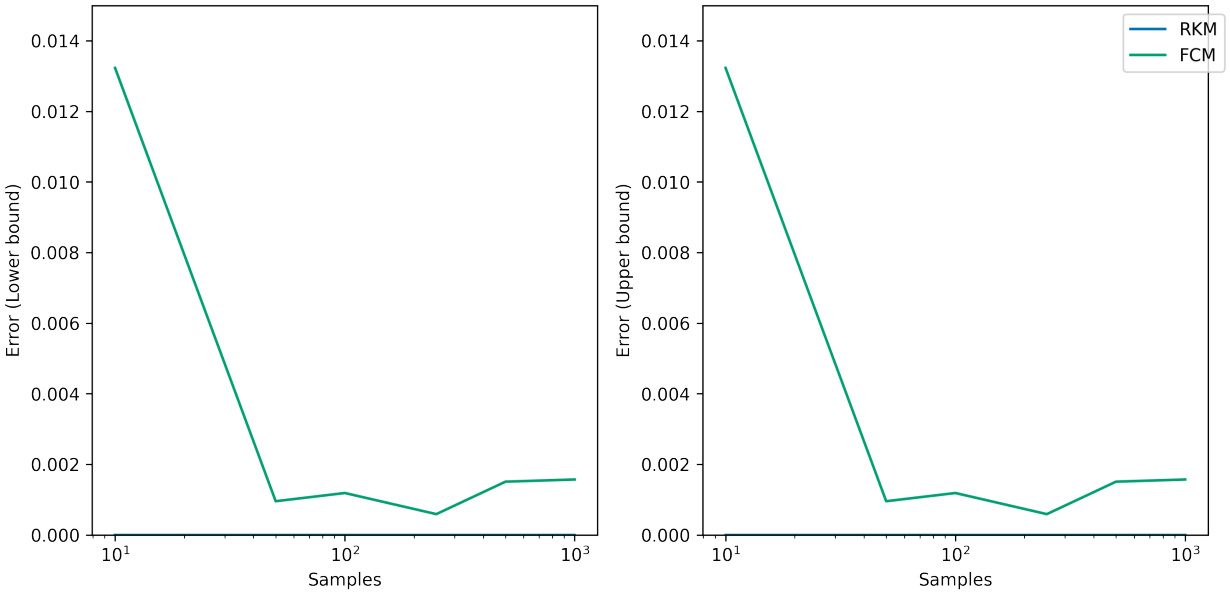

Figure 4: Results of the convergence analysis for the noisy dataset, for rough clustering and fuzzy clustering. The approximation error for RKM was equal to 0 as RKM converged to a hard clustering partition.

- In all cases, a number of iterations on the order of $O(mk)$, with m being the number of objects and k the number of clusters, seems to be sufficient for good approximation quality, provided that (for the case of possibilistic and evidential clustering) the number of inner samples is on the order of $O(m)$ and at least greater than 100.

5.2. Experiments on Real Datasets

In this section, we illustrate the use of the proposed metrics using a collection of benchmark datasets from the UCI repository [21], as shown in Table 1.

Table 1: Selected datasets table

Dataset	Results Name	Samples	Features	Classes
Iris	Iris	150	4	3
Wine	Wine	178	13	3
Breast Cancer Wisconsin (Diagnostic)	WDBC	569	30	2
Glass	Glass	214	9	6
Ecoli	Ecoli	336	7	8
Wine Quality (Red)	WineQuality	1599	11	6
Yeast	Yeast	1484	8	10
Optical Recognition of Digits	OptDigits	5620	64	10
Crowdsourced Mapping (Training)	Crowdsource	10545	28	6
Isolet (Training)	Isolet	6238	617	26

For each benchmark dataset, the soft clusterings given as output by each of the algorithms were compared with the ground truth labeling of the corresponding dataset, as

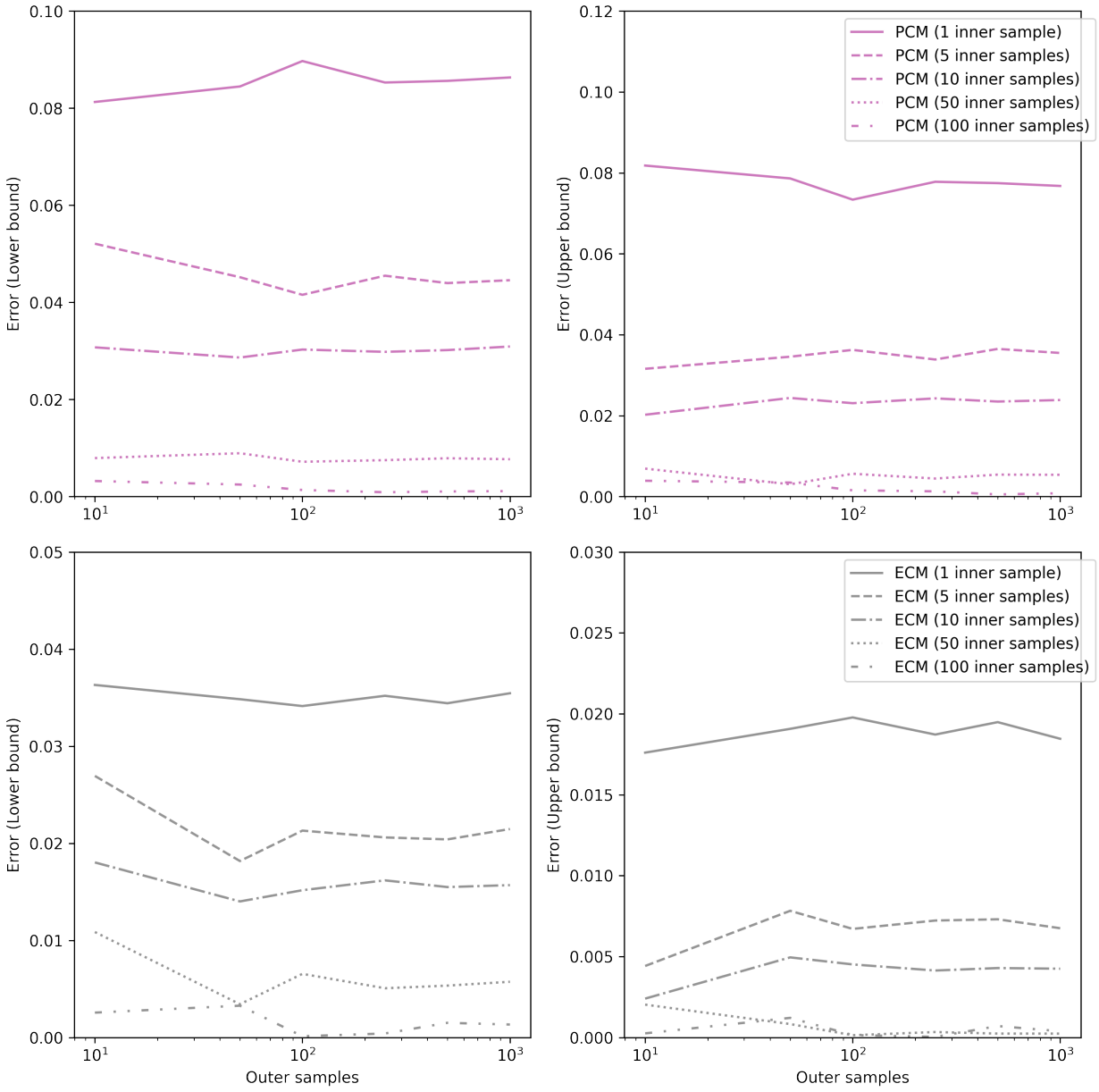

Figure 5: Results of the convergence analysis for the noisy dataset, for possibilistic clustering and evidential clustering.

provided in the UCI repository: the aim of these experiments was to illustrate the application of the proposed distributional measures as a way to evaluate soft clustering algorithms' to reconstruct a true clustering structure (see also Section 1). We considered the distributional generalization of the Rand index and, in particular, its sampling-based approximation as defined in Section 4. We decided to focus only on the sampling-based approximations not only due to the computational intractability of computing the exact values of the distributional measures, but also to illustrate how these approximations enable the application of our metrics to medium and large-scale datasets. For a comparison between the exact and approximated versions of the metrics, we refer the reader to [8, 9]. In regard to the sampling scheme, based also on the results discussed in Section 5.1, for RKM and FCM we considered, for each dataset (with m instances and k classes), Algorithm 1 and standard sampling, respectively, with a number of samples equal to $m \cdot k$; by contrast, for the case of PCM and ECM we considered Algorithm 3, with $m \cdot k$ samples in the outer loop and a constant number of 100 samples in the inner loop.

The results of the experiments are represented in Figure 6, in terms of raw performance values, and in Figure 7, in terms of a Critical Difference diagram [13]. The width of the interval representations (for RKM, PCM and ECM), represented in terms of a Critical Difference diagram, is reported in Figure 8. The running times for the sampling-based approximations of the distributional measures are represented in Figure 9, in terms of raw computing seconds, and in Figure 10, in terms of a Critical Difference diagram. In all cases, significant differences were evaluated through Wilcoxon signed-rank test and Holm correction for multiple testing.

Though our analysis has no pretense of generalizability or of drawing definitive conclusions on the performance of the considered algorithms, it is easy to see that the application of the proposed measures (and in particular, their sampling-based approximations) allows us to compare soft clustering methods that belong to completely different families in a meaningful manner. For example, it is easy to see that FCM and KM generally had comparable performance; except if we consider only the lower bounds on the interval representation of the distributional measures, they were outperformed by all other soft clustering methods, both in terms of average performance as well as (even more relevantly) in terms of the corresponding upper bounds. In this last respect, in particular, ECM significantly outperformed all other methods. This advantage of ECM can easily be seen to stem from the higher degree of ambiguity contained in the evidential clusterings returned by this algorithm: indeed, it can be seen from Figure 8 that ECM had a significantly larger gap between the lower and upper bounds of the distributional measure, as compared to the other considered algorithms. In any case, PCM and RKM also achieved good performances; in particular, RKM is the second best algorithm in terms of the lower bound, as well as the average, of the distributional measure. These results, all together, seem to suggest that allowing uncertainty and, specifically, ambiguity in the output of a clustering algorithm might provide a benefit in terms of performance. We plan to further investigate this conjecture in future work. At the same time, we see that, in general, algorithms which produced more ambiguous clustering also had worse performance in terms of running time required to compute the sampling-based approximations. For example, ECM (which was the algorithm that produced the

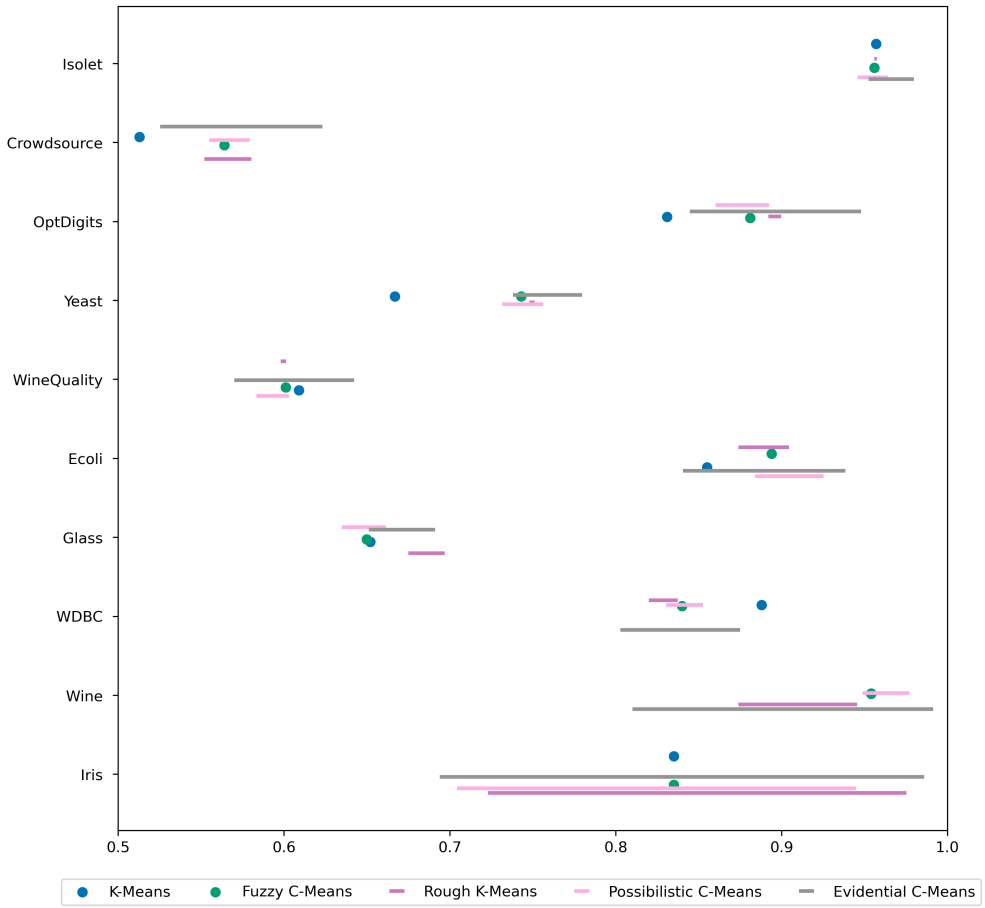

Figure 6: Performance of the considered soft clustering algorithms, as measured by the sampling-based approximation of the distributional measure (using the Rand index as base measure).

most ambiguous clusterings) was associated with the worst running time required for the approximation of the distributional measure, though the difference with respect to PCM was not significantly different. By contrast, FCM (which always produces clustering with no ambiguity) had the best running time: also in this case, it can be seen that computing the sampling-based approximations for RKM did not require significantly more time than for FCM: this may be a consequence of what we previously discussed in regard to Figure 8, as RKM was (among RKM, PCM and ECM) the algorithm which produced the least ambiguous clusterings. Thus, while ambiguity may be helpful to find soft clusterings with good performance, it may have an impact in terms of computational resources needed to compute the corresponding evaluation measures, as more candidate hard clusterings need to be evaluated to obtain sufficiently reliable results.

6. Conclusion

In this article, which is an extension of [8], we proposed and studied a mathematical approach, as well as computational algorithms, making it possible to lift any clustering

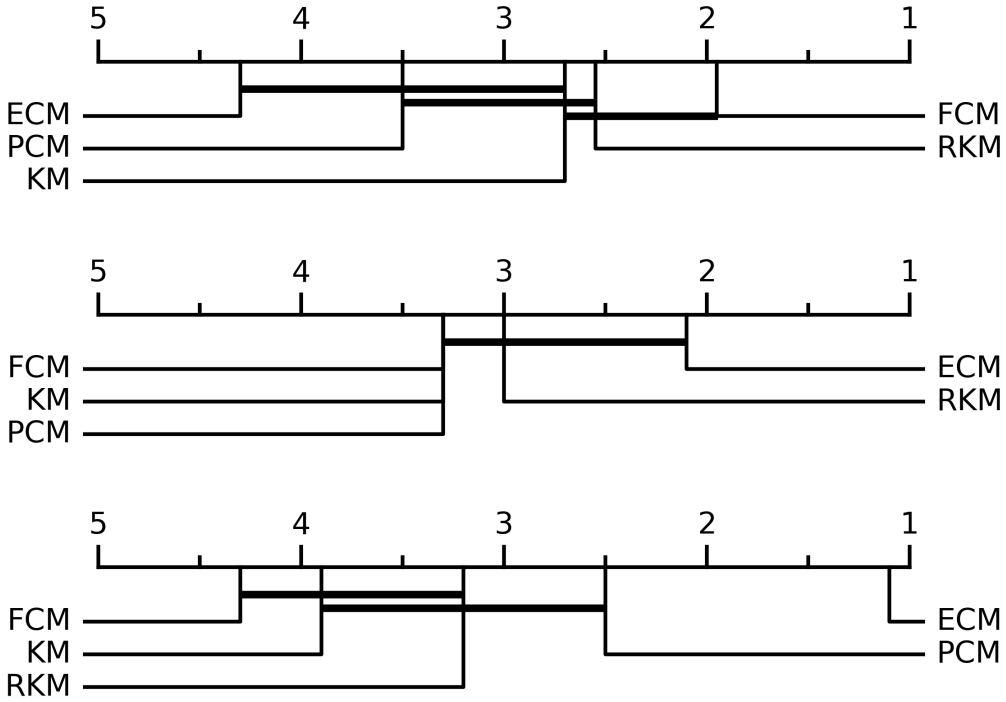

Figure 7: Critical difference diagram of the mean ranks of the considered soft clustering algorithm, in terms of: (top) lower bound on the performance; (middle) average performance; (bottom) upper bound on the performance. Ranks are based on the performance values shown in Figure 6: namely, for each dataset, algorithms were ranked according to their performance, and then the average rank for each algorithm (across all datasets) was computed. Thick horizontal bars denote lack of significant differences among a group of algorithms (Wilcoxon signed-rank test): that is, when two algorithms are connected by a horizontal bar, then, no significant difference was found among their performance.

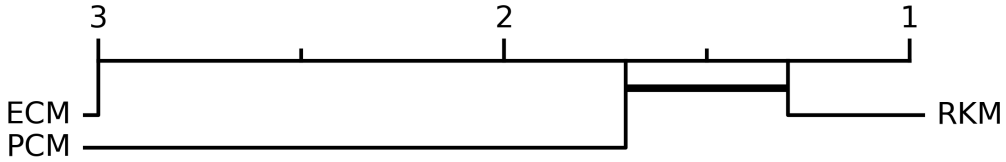

Figure 8: Critical difference diagram of the mean ranks of the considered soft clustering algorithm. Ranks are based on the gap between the upper and lower bound of the performance values shown in Figure 6. Thick horizontal bars denote lack of significant differences among a group of algorithms (Wilcoxon signed-rank test). See also Figure 7 for an extended explanation of critical difference diagrams.

comparison measures from hard clustering to evidential clustering (hence, as special cases, also to rough, fuzzy and possibilistic clustering). We studied the metric and complexity-theoretic properties of the proposed distributional measures. We have shown that computing these measures, or their interval representation, is in general computationally intractable,

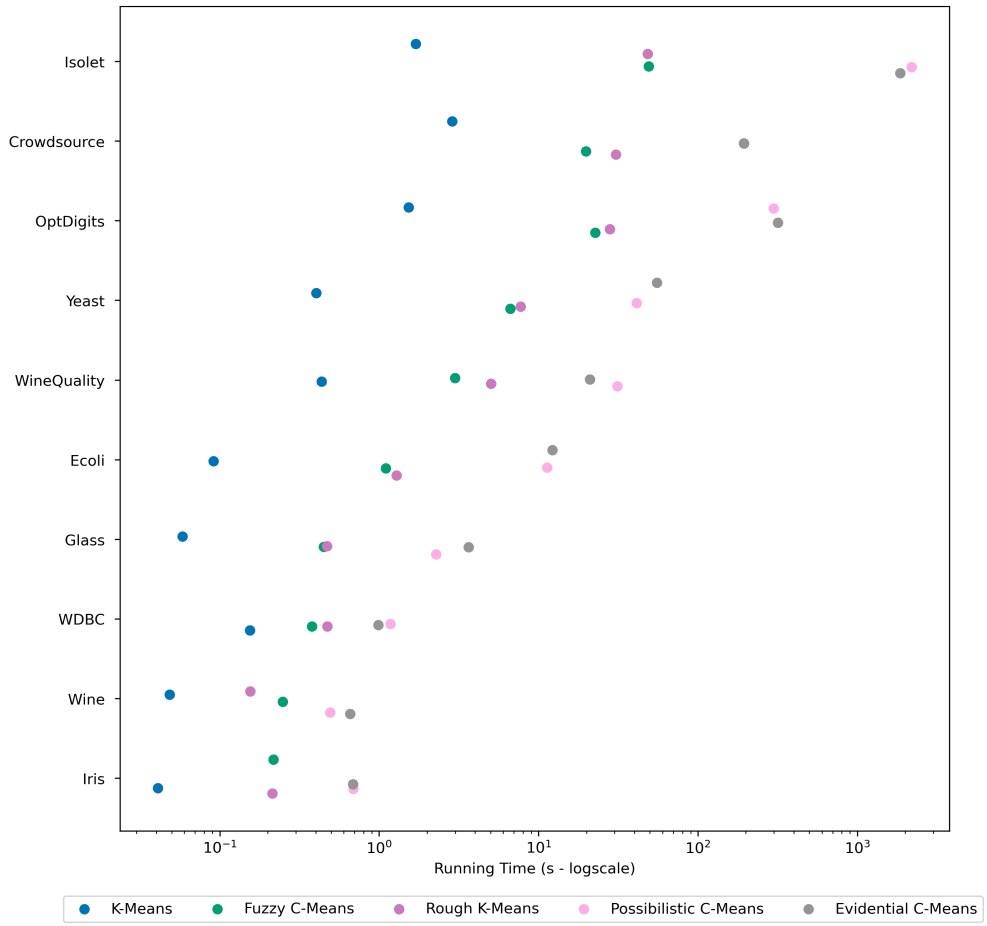

Figure 9: Running time (in seconds) of the sampling-based approximation of the distributional measure (using the Rand index as base measure).

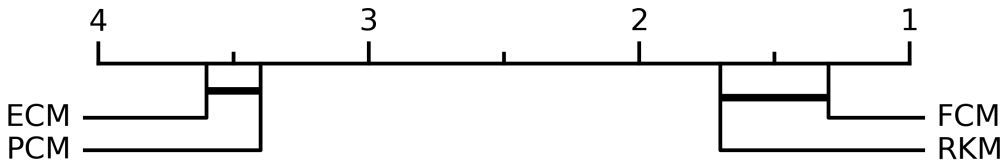

Figure 10: Critical difference diagram of the mean ranks of the running time for the sampling-based approximation of the distributional measure, based on the running time values shown in Figure 9. Thick horizontal bars denote lack of significant differences among a group of algorithms (Wilcoxon signed-rank test)

and we have proposed approximation strategies based on sampling with strong convergence guarantees. We have also provided sufficient and necessary conditions for the equivalence between distributional and transport-theoretic measures [9], setting the ground for a unified theory of soft clustering comparison measures. Finally, we illustrated the application of the

proposed methods through some simple experiments.

We believe that this work makes a step toward the development of general and principled approaches for the comparison of soft clustering algorithms. With this perspective in mind, we deem the following problems to be worthy of further investigation: 1) Generalizing Proposition 3.3 to other base distance measures, and determining whether this result can be extended to evidential clustering; 2) Designing more efficient sampling approaches, both from the theoretical point of view (i.e., new sampling algorithms with convergence bounds that are sharper or simpler than those proved in Section 4) and from the application-oriented one (e.g., by exploiting parallel computing approaches, or state-of-the-art Monte-Carlo engines [30]); 3) Extending our experimental results to more recent and state-of-the-art soft clustering algorithms, so as to delineate a comprehensive picture of their performance.

References

[1] Anderson, D.T., Bezdek, J.C., Popescu, M., et al., 2010. Comparing fuzzy, probabilistic, and possibilistic partitions. IEEE Transactions on Fuzzy Systems 18, 906–918.

[2] Anderson, D.T., Zare, A., Price, S., 2012. Comparing fuzzy, probabilistic, and possibilistic partitions using the earth mover's distance. IEEE Transactions on Fuzzy Systems 21, 766–775.

[3] Antoine, V., Guerrero, J.A., Xie, J., 2021. Fast semi-supervised evidential clustering. International Journal of Approximate Reasoning 133, 116–132.

[4] Bezdek, J.C., 1981. Pattern recognition with fuzzy objective function algorithms. Springer Science & Business Media.

[5] Brouwer, R.K., 2009. Extending the Rand, adjusted Rand and jaccard indices to fuzzy partitions. Journal of Intelligent Information Systems 32, 213–235.

[6] Campagner, A., Ciucci, D., 2019. Orthopartitions and soft clustering: soft mutual information measures for clustering validation. Knowledge-Based Systems 180, 51–61.

[7] Campagner, A., Ciucci, D., Denœux, T., 2022a. Belief functions and rough sets: Survey and new insights. International Journal of Approximate Reasoning 143, 192–215.

[8] Campagner, A., Ciucci, D., Denœux, T., 2022b. A distributional approach for soft clustering comparison and evaluation, in: Belief Functions: Theory and Applications: 7th International Conference, BELIEF 2022, Paris, France, October 26–28, 2022, Proceedings, Springer. pp. 3–12.

[9] Campagner, A., Ciucci, D., Denœux, T., 2023. A general framework for evaluating and comparing soft clusterings. Information Sciences 623, 70–93.

[10] Campello, R.J., 2007. A fuzzy extension of the Rand index and other related indexes for clustering and classification assessment. Pattern Recognition Letters 28, 833–841.

[11] Day, W.H., 1981. The complexity of computing metric distances between partitions. Mathematical Social Sciences 1, 269–287.

[12] Dempster, A., et al., 1967. Upper and lower probabilities induced by a multivalued mapping. Annals of Mathematical Statistics 38, 325–339.

[13] Demšar, J., 2006. Statistical comparisons of classifiers over multiple data sets. The Journal of Machine learning research 7, 1–30.

[14] Denoeux, T., 2019. Decision-making with belief functions: A review. International Journal of Approximate Reasoning 109, 87–110.

[15] Denœux, T., 2021. NN-EVCLUS: neural network-based evidential clustering. Information Sciences 572, 297–330.

[16] Denœux, T., Dubois, D., Prade, H., 2020. Representations of uncertainty in AI: beyond probability and possibility, in: A Guided Tour of Artificial Intelligence Research. Springer, pp. 119–150.

[17] Denœux, T., Kanjanatarakul, O., 2016. Evidential clustering: A review, in: Integrated Uncertainty in Knowledge Modelling and Decision Making, Springer International Publishing, Cham. pp. 24–35.

[18] Denœux, T., Li, S., Sriboonchitta, S., 2017. Evaluating and comparing soft partitions: An approach based on Dempster–Shafer theory. IEEE Transactions on Fuzzy Systems 26, 1231–1244.
[19] Denœux, T., Masson, M.H., 2004. EVCLUS: evidential clustering of proximity data. IEEE Trans Syst Man Cybern B Cybern 34, 95–109.
[20] Depaolini, M.R., Ciucci, D., Calegari, S., et al., 2018. External indices for rough clustering, in: International Joint Conference on Rough Sets, Springer. pp. 378–391.
[21] Dua, D., Graff, C., 2017. UCI machine learning repository. URL: `http://archive.ics.uci.edu/ml`.
[22] Ferraro, M.B., Giordani, P., 2020. Soft clustering. Wiley Interdisciplinary Reviews: Computational Statistics 12, e1480.
[23] Frigui, H., Hwang, C., Rhee, F.C.H., 2007. Clustering and aggregation of relational data with applications to image database categorization. Pattern Recognition 40, 3053–3068.
[24] Hüllermeier, E., Rifqi, M., Henzgen, S., et al., 2011. Comparing fuzzy partitions: A generalization of the Rand index and related measures. IEEE Transactions on Fuzzy Systems 20, 546–556.
[25] Kantorovich, L.V., 1960. Mathematical methods of organizing and planning production. Management science 6, 366–422.
[26] Krishnapuram, R., Keller, J.M., 1993. A possibilistic approach to clustering. IEEE transactions on fuzzy systems 1, 98–110.
[27] Liberti, L., 2019. Undecidability and hardness in mixed-integer nonlinear programming. RAIRO-Operations Research 53, 81–109.
[28] Masson, M.H., Denoeux, T., 2008. ECM: an evidential version of the fuzzy c-means algorithm. Pattern Recognition 41, 1384–1397.
[29] Naaman, M., 2021. On the tight constant in the multivariate Dvoretzky–Kiefer–Wolfowitz inequality. Statistics & Probability Letters 173, 109088.
[30] Patil, A., Huard, D., Fonnesbeck, C.J., 2010. PyMC: Bayesian stochastic modelling in Python. Journal of statistical software 35, 1.
[31] Peters, G., 2014. Rough clustering utilizing the principle of indifference. Information Sciences 277, 358–374.
[32] Peters, G., Crespo, F., Lingras, P., Weber, R., 2013. Soft clustering: Fuzzy and rough approaches and their extensions and derivatives. International Journal of Approximate Reasoning 54, 307–322.
[33] Rand, W.M., 1971. Objective criteria for the evaluation of clustering methods. Journal of the American Statistical association 66, 846–850.
[34] Rendón, E., Abundez, I., Arizmendi, A., Quiroz, E.M., 2011. Internal versus external cluster validation indexes. International Journal of computers and communications 5, 27–34.
[35] Ruspini, E.H., Bezdek, J.C., Keller, J.M., 2019. Fuzzy clustering: A historical perspective. IEEE Comput Intell Mag 14, 45–55.
[36] Saxena, A., Prasad, M., Gupta, A., Bharill, N., Patel, O.P., Tiwari, A., Er, M.J., Ding, W., Lin, C.T., 2017. A review of clustering techniques and developments. Neurocomputing 267, 664–681.
[37] Shafer, G., 1976. A mathematical theory of evidence. Princeton University Press.
[38] Smets, P., 1998. The transferable belief model for quantified belief representation, in: Quantified Representation of Uncertainty and Imprecision. Springer, pp. 267–301.
[39] Stamatelatos, G., Efraimidis, P.S., 2021. Lexicographic enumeration of set partitions. arXiv preprint arXiv:2105.07472 .
[40] Steele, J.M., 2004. The Cauchy-Schwarz master class: an introduction to the art of mathematical inequalities. Cambridge University Press.
[41] Sutherland, W.A., 2009. Introduction to metric and topological spaces. Oxford University Press.
[42] Villani, C., 2021. Topics in Optimal Transportation. American Mathematical Soc.
[43] Vinh, N.X., Epps, J., Bailey, J., 2010. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. The Journal of Machine Learning Research 11, 2837–2854.
[44] Xiong, H., Li, Z., 2018. Clustering validation measures, in: Data Clustering. Chapman and Hall/CRC, pp. 571–606.

[45] Yu, H., 2017. A framework of three-way cluster analysis, in: Rough Sets - Proceedings IJCRS 2017, Springer International Publishing. pp. 300–312.
[46] Zhou, D., Li, J., Zha, H., 2005. A new Mallows distance based metric for comparing clusterings, in: Proceedings of the 22nd international conference on Machine learning, pp. 1028–1035.
[47] Zhou, K., Guo, M., Martin, A., 2022. Evidential prototype-based clustering based on transfer learning. International Journal of Approximate Reasoning 151, 322–343.

A. Metric Spaces

Let X be a set and $\sim$ be a symmetric and transitive relation over X. A *metric* over X is a function $d : X \times X \mapsto \mathbb{R}_+$ such that: (M1) $\forall x \in X$, $d(x, x) = 0$; (M2) $\forall x, y \in X$, $x \nsim y \implies d(x, y) > 0$; (M3) $\forall x, y \in X$, $d(x, y) = d(y, x)$; and (M4) $\forall x, y, z \in X$, $d(x, z) \leq d(x, y)+d(y, z)$. Metric d is *normalized* if $\max_{x,y\in X} d(x, y) = 1$. If d is a normalized metric, then its dual $s = 1 - d$ is called a *similarity* over X. Mapping d is a *pseudo-metric* iff it satisfies (M1), (M3) and (M4); it is a *semi-metric* iff it satisfies (M1), (M2) and (M3); it is a *meta-metric* iff it satisfies (M2), (M3) and (M4) and (M1b) $\forall (x, y) \in X^2$, $d(x, y) = 0 \implies x \sim y$. A *consistency* is a semi-pseudo similarity, i.e. a c function such that $1 - c$ satisfies (M1) and (M3). For additional background on metric structures we refer the reader to [41]. A metric d over X can be extended to a metric over 2^X. The resulting metric d_H is called the *Hausdorff metric* based on d, defined as

$$d_H(A, B) = \max\{\max_{a\in A} d(a, B), \max_{b\in B} d(A, b)\}, \tag{26}$$

where $d(a, B) = \min_{b\in B} d(a, b)$ and $d(A, b) = \min_{a\in A} d(a, b)$. If d is a (pseudo-, meta-, semi-) metric, then d_H satisfies the same properties.

Similarly, a metric d over X can be extended to a metric over the space $\mathcal{P}(X)$ of probability measures over X. The resulting metric, denoted as d_W, is called the *Wasserstein metric* (also known as Kantorovich-Rubinstein metric) [25, 42] based on d. It is formally defined, for any two probability measures Pr_1 and Pr_2 on X, as

$$\begin{aligned} d_W(Pr_1, Pr_2) = \min_{\sigma} & \sum_{(x_1,x_2)\in X^2} \sigma(x_1, x_2) d(x_1, x_2) \\ \text{such that} & \sum_{x_2\in X} \sigma(x_1, x_2) = Pr_1(x_1) \\ & \sum_{x_1\in X} \sigma(x_1, x_2) = Pr_2(x_2) \\ & \sum_{(x_1,x_2)\in X^2} \sigma(x_1, x_2) = 1 \\ & \forall (x_1, x_2) \in X, \sigma(x_1, x_2) \geq 0. \end{aligned} \tag{27}$$

If d is a (pseudo-, meta-, semi-) metric then d_W satisfies the same properties.