Distances Between Formal Concept Analysis Structures

Alexandre Bazin, Giacomo Kahn

Distances Between Formal Concept Analysis Structures

Alexandre Bazin[a], Giacomo Kahn[b]

[a]*LIRMM, CNRS, Université de Montpellier, Montpellier, France*
[b]*Université Lumière Lyon 2, INSA Lyon, Université Claude Bernard Lyon 1, Université Jean Monnet Saint-Etienne, DISP UR4570, Bron, 69676, France*

Abstract

In this paper, we study the notion of distance between the most important structures of formal concept analysis: formal contexts, concept lattices, and implication bases. We first define three families of Minkowski-like distances between these three structures. We then present experiments showing that the correlations of these measures are low and depend on the distance between formal contexts.

Keywords: Pattern Mining, Formal Concept Analysis, Concept Lattices, Implication Bases, Ordinal Data Science

1. Introduction

Formal Concept Analysis (FCA [9]) is a mathematical framework that allows extracting patterns called concepts from data in the form of objects described by attributes, and organises them in an ordered structure called a concept lattice. Concept lattices are then used for exploratory search [12, 11], conceptual navigation [17, 1], and other applications – see [13] for a survey. The framework also handles implications between sets of attributes, that can be summarised by implication bases. In FCA, formal contexts, concept lattices and sets of implications are three representations of – or points of view on – the same entity and all three of them are well known, well studied, and well used in various fields of data mining [15, 14, 6].

We are interested in distances between these FCA structures. Given two data tables on the same objects and attributes, how far apart are the structures that are extracted from them? In this paper, we define three families of distances: one between formal contexts, one between implication bases, and one between concept lattices. For formal contexts, we consider the context as a set of pairs (the incidence relation) and use set-based analogues of Minkowski distances to define the *factual distance*. For concept lattices and implication bases, we consider the structures as representations of, respectively, the derivation operators and the closure operator

of the corresponding context and propose Minkowski-like distances: the *conceptual distance* and the *logical distance*. We show that these distances are metrics and we provide algorithms to compute them. We experimentally study the correlations between those distances on formal contexts that are *closer* or *farther apart* and observe that these correlations depend on the factual distance.

There are multiple expected applications for this work. The most direct one would be the comparison of concept lattices or implication bases, for instance to study the differences in the variability in different software product lines [3]. Another application would be in distance-based machine learning, for instance in the clustering of entities represented by binary relations. More broadly, these distances could be used to define complexity indicators in triadic or polyadic datasets [19] as the relative distance of each *slice* of context to every other, or even define a notion of trajectory in iterative processes such as Relational Concept Analysis [16, 2]. Additionally, this is a contribution to Ordinal Data Science, as defined in its manifesto [18], and it seemed fun to do.

The paper follows a classic structure: in Section 2 we define the necessary notions of FCA and distances, then we introduce our distances between FCA structures and the algorithms to compute them in Section 3. In Section 4 we experiment on the new distances: we study the correlations between them and compare them together and with Domenach's dissimilarity measure [7] on concept lattices.

2. Preliminaries

2.1. Formal Concept Analysis

Formal Concept Analysis (FCA) is a mathematical framework based on lattice theory that aims at structuring the information contained in the relation between *objects* and their *attributes* [9]. It is centered around the notion of *formal context*.

Definition 1 (Formal context)**.** *A* formal context *is a triple* $(\mathcal{O}, \mathcal{A}, \mathcal{R})$ *in which* $\mathcal{O}$ *is a set of* objects*,* $\mathcal{A}$ *is a set of* attributes *and* $\mathcal{R} \subseteq \mathcal{O} \times \mathcal{A}$ *is a binary relation between objects and attributes. We say that* the object o is described the attribute a *when* $(o, a) \in \mathcal{R}$*.*

Formal contexts can be represented as crosstables.

A formal context $\mathcal{C}$ gives rise to two *derivation operators*, both usually noted $\cdot'$, defined as:

$$\cdot' : \mathcal{P}(\mathcal{A}) \mapsto \mathcal{P}(\mathcal{O})$$

$$A' = \{o \in \mathcal{O} \mid \forall a \in A, (o, a) \in \mathcal{R}\}$$

	a_1	a_2	a_3	a_4	a_5
o_1	×	×			
o_2		×	×	×	
o_3		×		×	×
o_4			×		×
o_5				×	×

Figure 1: A formal context with five objects and five attributes

$$\cdot' : \mathcal{P}(\mathcal{O}) \mapsto \mathcal{P}(\mathcal{A})$$
$$O' = \{a \in \mathcal{A} \mid \forall o \in O, (o, a) \in \mathcal{R}\}$$

where $\mathcal{P}(X)$ denotes the powerset of X.

For instance, in the Fig. 1 context, $\{a_2, a_4\}' = \{o_2, o_3\}$ and $\{a_1\}'' = \{a_1, a_2\}$. Both compositions $\cdot''$ form Galois connections and are thus closure operators. Throughout this paper, when in the presence of two different formal contexts $\mathcal{C}_1$ and $\mathcal{C}_2$, we shall use $\cdot'^i$ and $\cdot''^i$ to denote the derivation and closure operators of context $\mathcal{C}_i$.

Definition 2 (Formal concept)**.** *In a formal context* $(\mathcal{O}, \mathcal{A}, \mathcal{R})$*, a* formal concept *is a pair* (E, I) *in which* E *is a set of objects,* I *is a set of attributes and such that* $E = I'$ *and* $I = E'$*. As such,* $I = I''$ *and* $E = E''$ *are both closed sets. We call* E *the* extent *and* I *the* intent *of the concept.*

Visually, concepts correspond to maximal rectangles of crosses in the context's crosstable, up to permutation of rows and columns. In the Fig. 1 context, the pair $(\{o_2, o_3\}, \{a_2, a_4\})$ is a concept while the pair $(\{o_3, o_4\}, \{a_5\})$ is not as $\{a_5\}' = \{o_3, o_4, o_5\}$. Concepts can be ordered by the inclusion relation on their extents, i.e. $(E_1, I_1) \leq (E_2, I_2) \Leftrightarrow E_1 \subseteq E_2$. As per the basic theorem of formal concept analysis [9], the set of all concepts of a context $\mathcal{C}$ ordered in such a way forms a complete lattice called the *concept lattice of* $\mathcal{C}$. Additionally, all complete lattices are isomorphic to the concept lattice of some context.

Definition 3 (Implications)**.** *In a formal context* $(\mathcal{O}, \mathcal{A}, \mathcal{R})$*, an* implication *is a pair of attribute sets* (X, Y)*, usually noted* $X \to Y$*. An implication* $X \to Y$ holds *in the context when* $X' \subseteq Y'$ *or, equivalently,* $Y \subseteq X''$*. In other words, the implication holds when all the objects described by* X *are also described by* Y*.*

In the Fig. 1 context, the implications $\{a_1\} \to \{a_1, a_2\}$ and $\{a_3, a_4\} \to \{a_2\}$ hold while the implication $\{a_3\} \to \{a_5\}$ does not. For simplicity's sake, we thereafter say "$X \to Y$" instead of "$X \to Y$ holds". Some implications can be inferred from others through Armstrong's axioms:

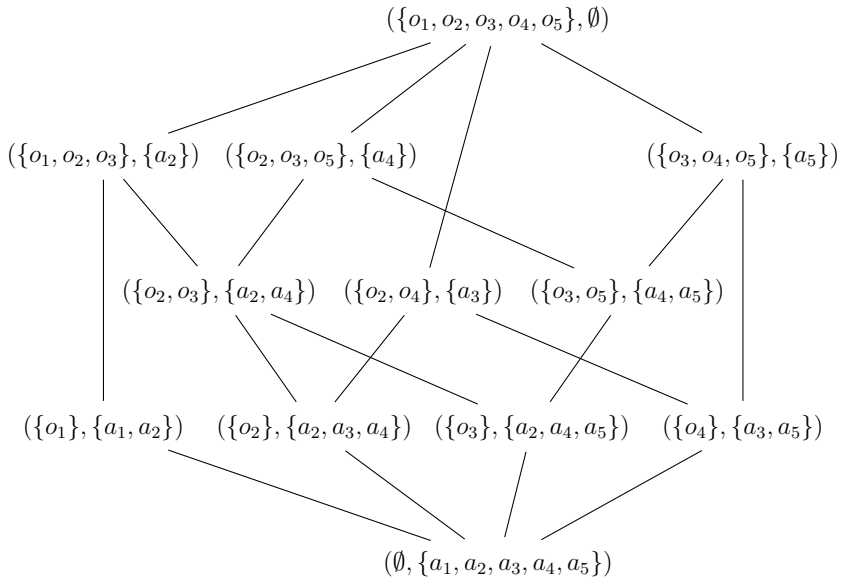


Figure 2: Concept lattice of the formal context depicted in Fig. 1.

- if $Y \subseteq X$, then $X \to Y$ (*Reflexivity*)
- if $X \to Y$, then $X \cup Z \to Y \cup Z$ for all attribute sets Z (*Augmentation*)
- if $X \to Y$ and $Y \to Z$, then $X \to Z$. (*Transitivity*)

Definition 4 (Implication base)**.** *An* implication base *of a formal context is an implication set $\mathcal{I}$ such that the set of implications that can be inferred from $\mathcal{I}$ through Armstrong's axioms is the set of all implications that hold in the context.*

Several implication bases with interesting properties exist in the literature [5, 4]. In this paper, we are interested in only one.

Definition 5 (Proper Premises)**.** *Let $(\mathcal{O}, \mathcal{A}, \mathcal{R})$ be a formal context and a an attribute. A* proper premise of *a is an inclusion-minimal attribute set X such that $X \to \{a\}$, i.e. there is no $Y \subset X$ such that $Y \to \{a\}$.*

In the Fig. 1 example, the set $\{a_2, a_4\}$ is a proper premise of the attribute a_5 as no proper subset of $\{a_2, a_4\}$ implies $\{a_5\}$. The set of all implications $X \to \{a\}$ where a is an attribute and X is one of its proper premises forms an implication base.

Definition 6 (Logical closure)**.** *Let $\mathcal{I}$ be an implication base. The* logical closure *of an attribute X by $\mathcal{I}$, noted $X^{\mathcal{I}}$, is defined as the biggest $Y \supseteq X$ such that $X \to Y$ can be inferred from $\mathcal{I}$.*

For instance, the logical closure of the attribute set $\{a_1, a_3\}$ by the implication base $\mathcal{I} = \{\{a_1\} \to \{a_2\}, \{a_2, a_3\} \to \{a_4\}\}$ is $\{a_1, a_3\}^{\mathcal{I}} = \{a_1, a_2, a_3, a_4\}$. The logical closure, as its name indicates, is a closure operator. If $\mathcal{C}$ is a formal context and $\mathcal{I}$ an implication base of $\mathcal{C}$, then $\cdot^{\mathcal{I}} = \cdot''$.

2.2. Metrics

A *metric* on a set S is a function of distance between the elements of S satisfying the following axioms:

- $f(x, x) = 0$
- $f(x, y) > 0$ when $x \neq y$, (positivity)
- $f(x, y) = f(y, x)$, (symmetry)
- $f(x, z) \leq f(x, y) + f(y, z)$. (triangular inequality)

In this paper, we make use of two families of metrics between vectors and sets so as to build our own metrics between FCA structures. The first is the well-known family of Minkowski distances between vectors $X = (x_1, \ldots, x_n)$ and $Y = (y_1, \ldots, y_n)$ defined as

$$D_p(X, Y) = \sqrt[p]{\sum_{i=1}^{n} |x_i - y_i|^p}.$$

The second is the family of normalised set-based analogues of Minkowski distances [10] defined, for two sets X and Y, as

$$d_{2,p}(X, Y) = \frac{\sqrt[p]{(|X| - |X \cap Y|)^p + (|Y| - |X \cap Y|)^p}}{|X \cap Y| + \sqrt[p]{(|X| - |X \cap Y|)^p + (|Y| - |X \cap Y|)^p}}.$$

3. Distances Between FCA Structures

3.1. Aim

We aim at proposing distance measures between FCA structures. This is not a brand new endeavor. Distance measures between formal contexts can be obtained by considering contexts as being any more widely known structures, such as bipartite graphs or hypergraphs, and using existing measures for these structures. Similarity measures between concept lattices have already been studied [7]. However, these are not sufficient. What we want is a set of three measures that can be used to compare

two entities in their three different forms (context, lattice and implication base) and the knowledge of how these three measures relate to each others. In this paper, we suppose that all pairs of structures we compare use the same objects and attributes.

In this section, we define families of distances for each of the usual structures of FCA, and show that they are metrics. The three families are based on the normalised set-based analogues of Minkowski distances $d_{2,p}$ [10]. In Section 4, we provide experimental results on the interaction of those distances.

3.2. Distance Between Contexts

As we only consider contexts on the same sets of objects and attributes, the distance between the contexts depends only on their incidence relations. Hence, we define our distance measure between formal contexts as a distance between binary relations seen as sets of pairs.

Definition 7. *Let* $\mathcal{C}_1 = (\mathcal{O}_1, \mathcal{A}_1, \mathcal{R}_1)$ *and* $\mathcal{C}_2 = (\mathcal{O}_2, \mathcal{A}_2, \mathcal{R}_2)$ *be two formal contexts. The* factual distance (FD) *between* $\mathcal{C}_1$ *and* $\mathcal{C}_2$ *is defined as*

$$FD_p(\mathcal{C}_1, \mathcal{C}_2) = d_{2,p}(\mathcal{R}_1, \mathcal{R}_2).$$

The two formal contexts depicted in Fig. 3 have a factual distance of ≈ 0.13.

	a_1	a_2	a_3	a_4
o_1	×	×	×	×
o_2	×	×	×	
o_3	×	×		
o_4	×			

	a_1	a_2	a_3	a_4
o_1	×	×	×	×
o_2	×	×		×
o_3	×	×		
o_4	×			

Figure 3: Two chain contexts. The two contexts have a factual distance of ≈ 0.13 with $p = 2$.

As $d_{2,p}$ is a metric, the factual distance is a metric.

3.3. Distance Between Concept Lattices

We consider concept lattices as pairs of functions that map sets of objects to the set of attributes they have in common and sets of attributes to the set of objects they all describe, i.e. we see concept lattices as representations of the derivation operators $\cdot'$. If (E, I) is a concept, then all subsets of E that are not subsets of lower neighbours in the lattice are mapped to I and reciprocally. This is notationally easier to express in terms of the derivation operators associated with the formal context of the lattice, which might not be explicitly given: object sets O are mapped to O'. As such, we define our distance measure between concept lattices as a distance measure between the derivation operators.

Definition 8. *Let $\mathcal{L}_1, \mathcal{L}_2$ be the two concept lattices of two contexts $\mathcal{C}_1$ and $\mathcal{C}_2$ with the same sets of objects $\mathcal{O}$ and attributes $\mathcal{A}$. We define the* lattice object distance *as*

$$LOD_{p,q}(\mathcal{L}_1, \mathcal{L}_2) = \frac{\sqrt[p]{\sum_{o \in \mathcal{O}} d_{2,q}(\mathcal{P}(\{o\}'^1), \mathcal{P}(\{o\}'^2))^p}}{\sqrt[p]{|\mathcal{O}|}}$$

and the lattice attribute distance *as*

$$LAD_{p,q}(\mathcal{L}_1, \mathcal{L}_2) = \frac{\sqrt[p]{\sum_{a \in \mathcal{A}} d_{2,q}(\mathcal{P}(\{a\}'^1), \mathcal{P}(\{a\}'^2))^p}}{\sqrt[p]{|\mathcal{A}|}}.$$

The conceptual distance *(CD) between $\mathcal{L}_1$ and $\mathcal{L}_2$ is then defined as*

$$CD_{p,q}(\mathcal{L}_1, \mathcal{L}_2) = min(LOD_{p,q}(\mathcal{L}_1, \mathcal{L}_2), LAD_{p,q}(\mathcal{L}_1, \mathcal{L}_2)).$$

Figure 4 depicts the two chain concept lattices of the two contexts in Fig. 3. Even though they are isomorphic, their conceptual distance is ≈ 0.33 with $p = 2$ and $q = 1$.

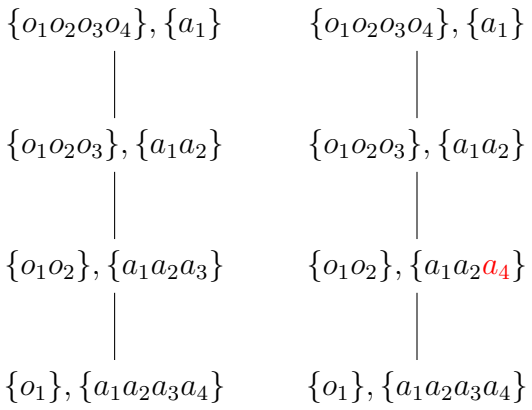


Figure 4: The two concept lattices of the Fig. 3 contexts. These have a conceptual distance of $CD_{2,1} \approx 0.33$ with $p = 2$ and $q = 1$. A small difference in the intents leads to a non-zero distance, even on isomorphic lattices with the same extents.

The conceptual distance takes its values in $[0, 1]$ and is a metric, satisfying the following axioms:

1. $CD(x, x) = 0$
2. $CD(x, y) > 0$ when $x \neq y$, (positivity)

3. $CD(x, y) = CD(y, x)$, (symmetry)
4. $CD(x, z) \leq CD(x, y) + CD(y, z)$. (triangular inequality)

These follow directly from the fact that $d_{2,p}$ is a metric.

Computing the conceptual distance is quite easy: for each object o, find its introducer concepts in both lattices to obtain $\{o\}'^1$ and $\{o\}'^2$. Then, computing $d_{2,q}(\mathcal{P}(\{o\}'^1), \mathcal{P}(\{o\}'^2))$, the distance between the sets of attribute sets that contain o in their derivation, is straightforward. Algorithm 1 follows this principle. Finding the introducer concept of an object in the lattice $\mathcal{L}_i$ can be done in $O(|\mathcal{L}_i|)$ so the time complexity of Algorithm 1 is in $O((|\mathcal{O}| + |\mathcal{A}|) \times max(|\mathcal{L}_1|, |\mathcal{L}_2|))$.

Algorithm 1: $CD_{p,q}$

Input: Two concept lattices $\mathcal{L}_1$ and $\mathcal{L}_2$, p and q
Output: $CD_{p,q}(\mathcal{L}_1, \mathcal{L}_2)$

$LOD = 0$
foreach *object* o **do**
  (E_1, I_1) = the introducer of o in $\mathcal{L}_1$
  (E_2, I_2) = the introducer of o in $\mathcal{L}_2$
  $LOD = LOD + (\sqrt[q]{(2^{|I_1|} - 2^{|I_1 \cap I_2|})^q + (2^{|I_2|} - 2^{|I_1 \cap I_2|})^q})^p$
$LOD = \sqrt[p]{LOD} / \sqrt[p]{|\mathcal{O}|}$
$LAD = 0$
foreach *attribute* a **do**
  (E_1, I_1) = the introducer of a in $\mathcal{L}_1$
  (E_2, I_2) = the introducer of a in $\mathcal{L}_2$
  $LAD = LAD + (\sqrt[q]{(2^{|E_1|} - 2^{|E_1 \cap E_2|})^q + (2^{|E_2|} - 2^{|E_1 \cap E_2|})^q})^p$
$LAD = \sqrt[p]{LAD} / \sqrt[p]{|\mathcal{A}|}$
return $min(LOD, LAD)$

3.4. Distance between Implication Bases

For our distance measure between implication bases, we consider implication bases as functions mapping attribute sets X to attribute sets $Y = \{y \mid X \to \{y\}\}$, i.e. we see implication bases as representations of the closure operator $\cdot''$ on attributes. Note that, from Armstrong's axioms, we can infer that

$$X \to Y \Leftrightarrow \forall y \in Y, X \to \{y\}.$$

$\{a_4\} \to \{a_2, a_3\}$ $\quad$ $\{a_4\} \to \{a_2\}$

$\{a_3\} \to \{a_2\}$ $\quad$ $\{a_3\} \to \{a_2, a_4\}$

$\emptyset \to \{a_1\}$ $\quad$ $\emptyset \to \{a_1\}$

Figure 5: The two proper premises bases of the Fig. 3 contexts. The logical distance, with parameters $p = 2$ and $q = 1$, between these two bases is ≈ 0.23.

Definition 9. *Let $\mathcal{I}_1, \mathcal{I}_2$ be two implications bases on the same attribute set $\mathcal{A}$. For an attribute $a \in \mathcal{A}$ and an implication base $\mathcal{I}$, we denote by $\mathcal{I}^a = \{X \mid a \in X^{\mathcal{I}}\}$ the set of attributes sets that imply a. The* logical distance *(LD) between $\mathcal{I}_1$ and $\mathcal{I}_2$ is then defined as*

$$LD_{p,q}(\mathcal{I}_1, \mathcal{I}_2) = \frac{\sqrt[p]{\sum_{a \in \mathcal{A}} d_{2,q}(\mathcal{I}_1^a, \mathcal{I}_2^a)^p}}{\sqrt[p]{|\mathcal{A}|}}.$$

Fig. 5 depicts the two proper premises implication bases of the contexts in Fig 3. These two implication bases have a logical distance of ≈ 0.23. Indeed, the attribute a_3 is implied by all supersets of $\{a_4\}$ only in the first context and the attribute a_4 is implied by all supersets of $\{a_3\}$ only in the second context.

The logical distance takes its values in $[0, 1]$ and is a metric, satisfying the following axioms:

1. $LD(x, x) = 0$
2. $LD(x, y) > 0$ when $x \neq y$, (positivity)
3. $LD(x, y) = LD(y, x)$, (symmetry)
4. $LD(x, z) \leq LD(x, y) + LD(y, z)$. (triangular inequality)

Just as those for the conceptual distance, these axioms follow from the fact that $d_{2,q}$ is a metric.

To compute the logical distance, one requires the knowledge of all the attribute sets X that imply a given attribute a. This is not explicitly contained in implication bases and retrieving it is the computationally most expensive part of computing the distance. We propose Algorithm 3 to compute the logical distance. We assume that the implication bases are proper premises bases. If this is not the case, other bases can be converted to proper premises bases.

The algorithm treats each attribute a separately. It starts with computing the attribute sets X that are minimal such that $X \to \{a\}$ in both implication bases

($commonPremises$). It then computes the union closure of the set of common premises (U_c), of the set of premises in $\mathcal{I}_1$ (U_1) and of the set of premises in $\mathcal{I}_2$ (U_2), i.e. the minimal sets of attributes sets such that $X, Y \in U_c \Rightarrow X \cup Y \in U_c$ (resp. U_1, U_2). From there, the algorithm associates to each element x of U_c (resp. U_1, U_2) the number of attribute sets that contain x but not its supersets in U_c (i.e. the size of the equivalence classes in the resulting lattice) with Algorithm 2. Algorithm 3 then sums those numbers (sum_c, sum_1 and sum_2) to obtain the numbers of attribute sets containing one of the corresponding premises. As the size of the union closure is bounded by $2^{|\mathcal{A}|}$ (when all singletons are premises), the worst case complexity of Algorithm 3 is in $O(|\mathcal{A}| \times 2^{|\mathcal{A}|})$.

Algorithm 2: $sizeEQ$

Input: A set U of premises
Output: $sizeEQ(U)$

```
1  Build a dictionary D mapping each premise P in U to the set of premises
     P_2 ⊃ P
2  sum ← 0
3  over ← false
4  while over = false do
5      over ← true
6      foreach premise P in U do
7          if all P_2 ∈ D(P) have been tagged then
8              |P^≡| ← 2^(|A|−|P|−1) − Σ_{P_2∈D(P)} |P_2^≡|
9              Tag P
10             sum ← sum + |P^≡|
11             over ← false
12 return sum
```

Algorithm 3: LD

Input: Two implication bases $\mathcal{I}_1$ and $\mathcal{I}_2$, p, q

Output: $LD_{p,q}(\mathcal{I}_1, \mathcal{I}_2)$

$Result \leftarrow 0$

foreach *attribute* a **do**

  $commonPremises = min(\{P_1 \cup P_2 \mid P_1 \in \mathcal{I}_1^a, P_2 \in \mathcal{I}_2^a\})$

  $U_c = unionClosure(commonPremises)$

  $U_1 = unionClosure(\mathcal{I}_1^a)$

  $U_2 = unionClosure(\mathcal{I}_2^a)$

  $sum_c = sizeEQ(U_c)$

  $sum_1 = sizeEQ(U_1)$

  $sum_2 = sizeEQ(U_2)$

  $Result = Result + (\sqrt[q]{(sum_1 - sum_c)^q + (sum_2 - sum_c)^q})^p$

return $\sqrt[p]{Result} / \sqrt[p]{|\mathcal{A}|}$

4. Experiments

In all these experiments, we used parameters $p = 2$ and $q = 1$ for all distances. A Python module containing the three distances is publicly available[1].

4.1. Correlation Between distances

$\mathcal{C}_1$	a_1	a_2	a_3	a_4
o_1		×	×	×
o_2	×	×		×
o_3			×	×
o_4	×			×

$\mathcal{C}_2$	a_1	a_2	a_3	a_4
o_1				
o_2				×
o_3				×
o_4		×		

$\mathcal{C}_3$	a_1	a_2	a_3	a_4
o_1		×		×
o_2			×	
o_3	×	×		×
o_4	×	×	×	×

Figure 6: Three formal contexts $\mathcal{C}_1$, $\mathcal{C}_2$ and $\mathcal{C}_3$.

The first question that may come to mind is "how do these distance measures compare to each others?". Let us consider the Fig. 6 example representing three contexts $\mathcal{C}_1$, $\mathcal{C}_2$ and $\mathcal{C}_3$. We denote their concept lattices by $\mathcal{L}_1$, $\mathcal{L}_2$ and $\mathcal{L}_3$ and their proper implication bases by $\mathcal{I}_1$, $\mathcal{I}_2$ and $\mathcal{I}_3$. We compute the factual, conceptual and logical distances between $\mathcal{C}_1$ and the other two and obtain the following results:

[1] https://github.com/Authary/FCAD

$$FD_{2,1}(\mathcal{C}_1, \mathcal{C}_2) = 0.80 \qquad FD_{2,1}(\mathcal{C}_1, \mathcal{C}_3) = 0.58$$
$$CD_{2,1}(\mathcal{L}_1, \mathcal{L}_2) = 0.69 \qquad CD_{2,1}(\mathcal{L}_1, \mathcal{L}_3) = 0.68$$
$$LD_{2,1}(\mathcal{I}_1, \mathcal{I}_2) = 0.15 \qquad LD_{2,1}(\mathcal{I}_1, \mathcal{I}_3) = 0.20$$

We observe that $\mathcal{C}_3$ is factually and conceptually closer to $\mathcal{C}_1$ while $\mathcal{C}_2$ is logically closer. The three measures therefore do not always agree. Are they at least correlated? In order to answer this question, we generated structures in different ways. As contexts are the easiest structure to randomly generate, we explored the following approaches:

- generating random contexts by adding each cross with a probability p
- generating random contexts factually close to a reference context by randomly flipping the truth value of the presence of pairs in the incidence relation with a probability p
- generating a sequence of contexts that get progressively farther from a reference context by iteratively modifying the context

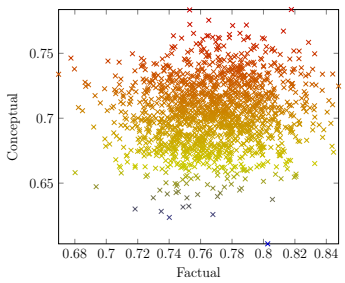
In each case, we computed three correlation coefficients, Pearson, Spearman and Kendall's τ.

4.2. Randomly Generated Contexts

We randomly generated 1500 pairs (A, B) of formal contexts with 50 objects and 10 attributes, with a pair (o, a) having a probability 0.3 of being in the incidence relation. We then computed the distances between the contexts (resp. their associated lattices and implication bases) in each pair. Fig. 7 depicts three diagrams illustrating the relation between the factual (x-axis) and logical (y-axis) distances, the relation between the factual (x-axis) and the conceptual (y-axis) distances and the relation between the conceptual (x-axis) and the logical (y-axis) relations.

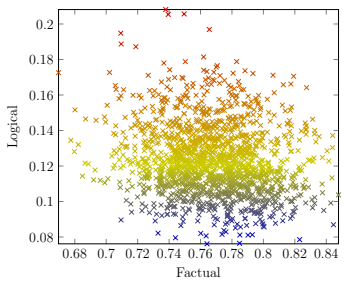
We observe that the three distances appear to be pairwise independent when the contexts are randomly generated in such a way. Fig. 7 also depicts the values of the three correlation coefficients, Pearson, Spearman and Kendall's τ. Their values confirm the independence, with the factual and conceptual distances being very slightly less independent. Note that Pearson measures linear correlation, Spearman assesses monotonic relationships and Kendall's τ measures rank correlation.

Interestingly, all factual distances are between 0.67 and 0.85, suggesting that random generation produces contexts that are far apart.



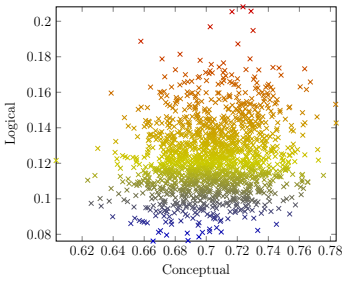
Pearson	Spearman	Kendall
0.09	0.10	0.06

(a) Factual (x-axis) and conceptual (y-axis) distances of random contexts.



Pearson	Spearman	Kendall
−0.16	−0.15	−0.10

(b) Factual (x-axis) and logical (y-axis) distances of random contexts.



Pearson	Spearman	Kendall
0.14	0.13	0.09

(c) Conceptual (x-axis) and logical (y-axis) distances of random contexts.

Figure 7: Randomly generated contexts : correlation between the distance measures.

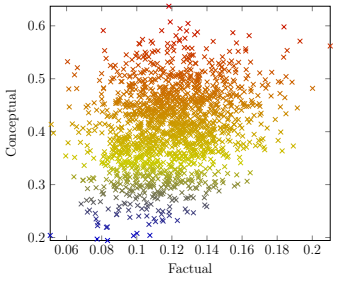
4.3. Randomly Modified Contexts

In a second batch of experiments, we generated 1500 other pairs of contexts such that A is a randomly generated context and B is obtained by randomly modifying A. All contexts contain 50 objects and 10 attributes. The contexts A were generated with a probability 0.3 for each cross. The modified contexts were obtained through the following algorithm: for each $(object, attribute)$ pair, with a probability 0.05, remove the pair from the incidence relation if it belongs to it or add it if it does not. We then computed the distances between the contexts (resp. their associate lattices and implication bases) in each pair. Fig. 8 depicts three diagrams illustrating the relation between the factual (x-axis) and logical (y-axis) distances, the relation between the factual (x-axis) and the conceptual (y-axis) distances and the relation between the conceptual (x-axis) and the logical (y-axis) relations. Fig. 8 also depicts the values of the three correlation coefficients, Pearson, Spearman and Kendall's τ.

Visually, we observe some slight positive correlation between the factual and conceptual distances and between the factual and logical distances. This is in opposition to the previous experiment with randomly generated contexts. All factual distances are below 0.2, suggesting that our modification algorithms successfully produces contexts that are close together. This result, together with the previous one on randomly generated contexts, hints at the correlations between the factual distance and the others being stronger for very close contexts.

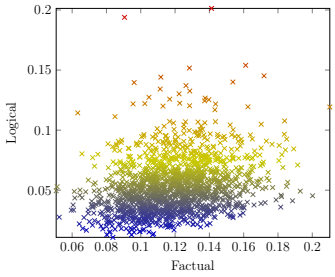
4.4. Variation of Correlation Relative to the Factual Distance

In the previous experiments, the distances seemed to be more correlated for low factual distances. This hinted at differences in correlations depending on factual



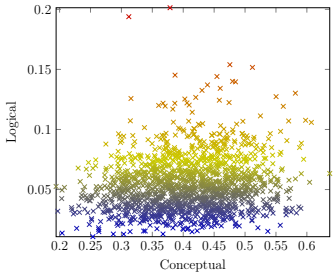
Pearson	Spearman	Kendall
0.25	0.24	0.16

(a) Factual (x-axis) and conceptual (y-axis) distances to a reference context of contexts obtained by randomly modifying the reference.



Pearson	Spearman	Kendall
0.26	0.28	0.19

(b) Factual (x-axis) and logical (y-axis) distances to a reference context of contexts obtained by randomly modifying the reference.



Pearson	Spearman	Kendall
0.14	0.13	0.09

(c) Conceptual (x-axis) and logical (y-axis) distances to a reference context of contexts obtained by randomly modifying the reference.

Figure 8: Randomly modified contexts: correlation between the distance measures (factual,logical), (factual,conceptual) and (conceptual,logical)

distances. Let us check whether this is really the case. For this experiment, we generated pairs (A, B) of contexts such A is a 50×10 randomly generated context and B is obtained by randomly flipping the truth value of each $(object, attribute)$ pair in A with a probability p. We made p vary from 0.025 to 0.5 with 0.025 increments. For each value of p, we generated 1000 pairs of contexts and computed the three distances between A and B. We then computed the three correlation coefficient for each pair of distances. Fig. 9 presents these correlation values (y-axes) for the different values of p (x-axes). As a higher p results in factually more distant contexts, we observe that the correlations between the factual distance and the other two decrease when p increases. This confirms our previous observation that these distances are more correlated for low factual distances.

4.5. Progression from a Reference Context

In a fourth experiment, we generated 200 contexts by starting with a random reference context and iteratively modifying it using the modification algorithm previously described with a probability $p = 0.02$. The goal is to observe the progression of the three distances when contexts get factually farther from the reference context. Fig. 10 depicts the factual, conceptual and logical distances of the 200 contexts to the reference context. We observe that the contexts indeed get progressively factually farther from the reference context until around 50 iterations, at which point the factual distance stabilises. The conceptual and logical distances increase much faster and stabilise around 40 iterations, with more variance in the logical distance.

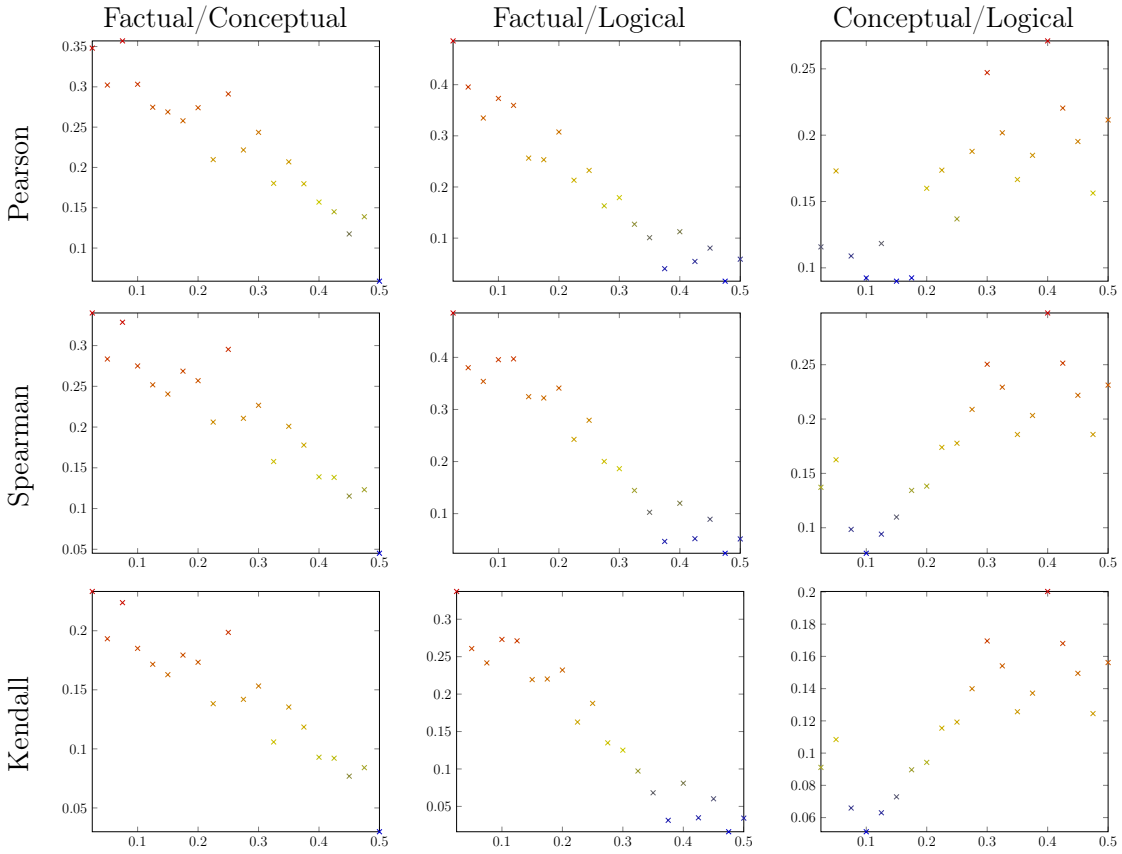
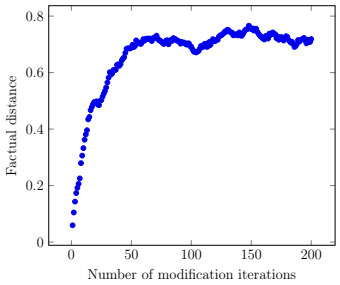


Figure 9: Correlation of the three distance measures for different values of probabilities used in the modification of contexts. Higher probabilities means higher factual distances.

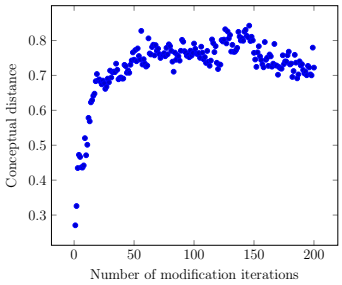
This reinforces the idea of a correlation between the factual distance and the other two for small factual distances. Note that the stabilisation occurs when the factual distance reaches above 0.6, which is the distance range in which randomly generated contexts occur.

4.6. Comparaison with Domenach's Dissimilarity Measure

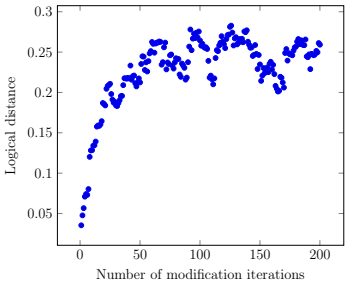
Domenach's dissimilary measure is based on the *overhanging* relation [8] between sets of objects. Two sets are overhanged if one is a subset of the other and their closures are different. To compute a distance between concept lattices, Domenach defines two matrices, M_1 and M_2, based on the overhanging relation of pairs of objets in each concept lattice. The distance is then based on the L_1 norms of those matrices: $\frac{||M_1 - M_2||}{||M_1|| + ||M_2||}$.



(a) Progression of the factual distance.

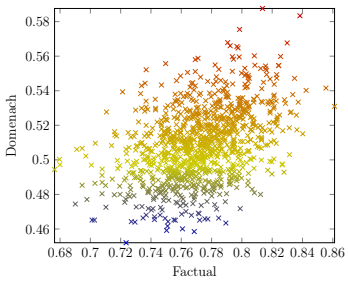


(b) Progression of the conceptual distance.



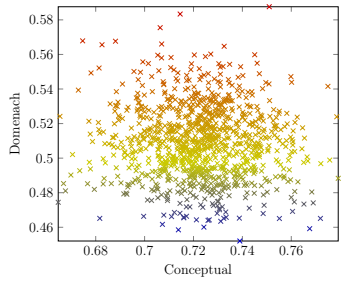
(c) Progression of the logical distance.

Figure 10: Iterative modifications from a reference context: progression of the three distances.



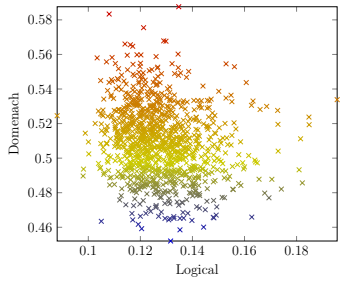
Pearson	Spearman	Kendall
0.41	0.40	0.27

(a) Factual distance (x-axis) and Domenach's dissimilarity (y-axis) of random contexts.



Pearson	Spearman	Kendall
−0.05	−0.03	−0.02

(b) Conceptual distance (x-axis) and Domenach's dissimilarity (y-axis) of random contexts.



Pearson	Spearman	Kendall
−0.17	−0.19	−0.13

(c) Logical distance (x-axis) and Domenach's dissimilarity (y-axis) of random contexts.

Figure 11: Randomly generated contexts : correlation between our distance measures and Domenach's dissimilarity.

We compared our distances with Domenach's dissimilarity measure on 1000 pairs of randomly generated contexts. Fig. 11 depicts the results. We observe that Domenach's dissimilarity measure is independent of our conceptual distance and slightly correlated with our factual distance.

5. Conclusion and Perspectives

We presented three distance families between the most important structures in formal concept analysis, *i.e.* formal contexts, concept lattices and implication bases. These structures represent three complementary points of view on the information contained in formal context: the factual, conceptual and logical points of views. We see the distances we studied in this paper as a first step towards the simultaneous

exploitation of the three points of view in the analysis of data. The applications could be distance-based machine learning, both supervised and unsupervised, or the measurement of the complexity of multidimensional data.

Our experiments indicate that, of our distances, only the factual distance is (barely) correlated with the other two and that their correlations depends on the factual distance. The variation of the correlation w.r.t. other distances should also be studied once we better understand how to control the conceptual and logical distances in the generation of data. Our experiments also highlight the need to study the metric spaces induced by the distances, and their relations, as experimental results are insufficient.

Future work includes the extension of these measures to contexts defined on different sets of objects and attributes, and to the polyadic concept analysis framework.

Acknowledgement

This work was partially supported by the ANR SmartFCA project Grant ANR-21-CE23-0023 of the French National Research Agency. The authors thank the members of the SmartFCA project for their advice.

References

[1] Bazin, A., Carbonnel, J., Kahn, G.: On-demand generation of aoc-posets: Reducing the complexity of conceptual navigation. In: Foundations of Intelligent Systems: 23rd International Symposium, ISMIS 2017, Warsaw, Poland, June 26-29, 2017, Proceedings 23. pp. 611–621. Springer (2017)

[2] Bazin, A., Galasso, J., Kahn, G.: Polyadic relational concept analysis. International Journal of Approximate Reasoning **164**, 109067 (2024)

[3] Bazin, A., Huchard, M., Martin, P.: Towards analyzing variability in space and time of products from a product line using triadic concept analysis. In: Proceedings of the 27th ACM International Systems and Software Product Line Conference-Volume B. pp. 85–89 (2023)

[4] Bertet, K., Demko, C., Viaud, J.F., Guérin, C.: Lattices, closures systems and implication bases: A survey of structural aspects and algorithms. Theoretical Computer Science **743**, 93–109 (2018)

[5] Bertet, K., Monjardet, B.: The multiple facets of the canonical direct unit implicational basis. Theoretical Computer Science **411**(22-24), 2155–2166 (2010)

[6] Codocedo, V., Napoli, A.: Formal concept analysis and information retrieval–a survey. In: International Conference on Formal Concept Analysis. pp. 61–77. Springer (2015)

[7] Domenach, F.: Similarity measures of concept lattices. In: Data Science, Learning by Latent Structures, and Knowledge Discovery. pp. 89–99. Springer (2015)

[8] Domenach, F., Leclerc, B.: Closure systems, implicational systems, overhanging relations and the case of hierarchical classification. Mathematical Social Sciences **47**(3), 349–366 (2004)

[9] Ganter, B., Wille, R.: Formal concept analysis: mathematical foundations. Springer Science & Business Media (2012)

[10] Horadam, K.J., Nyblom, M.A.: Distances between sets based on set commonality. Discrete Applied Mathematics **167**, 310–314 (2014)

[11] Huchard, M., Martin, P., Muller, E., Poncelet, P., Raveneau, V., Sallaberry, A.: Rcaviz: Exploratory search in multi-relational datasets represented using relational concept analysis. International Journal of Approximate Reasoning p. 109123 (2024)

[12] Keip, P., Gutierrez, A., Huchard, M., Le Ber, F., Sarter, S., Silvie, P., Martin, P.: Effects of input data formalisation in relational concept analysis for a data model with a ternary relation. In: International Conference on Formal Concept Analysis. pp. 191–207. Springer (2019)

[13] Poelmans, J., Ignatov, D.I., Kuznetsov, S.O., Dedene, G.: Formal concept analysis in knowledge processing: A survey on applications. Expert systems with applications **40**(16), 6538–6560 (2013)

[14] Poelmans, J., Ignatov, D.I., Kuznetsov, S.O., Dedene, G.: Fuzzy and rough formal concept analysis: a survey. International Journal of General Systems **43**(2), 105–134 (2014)

[15] Poelmans, J., Kuznetsov, S.O., Ignatov, D.I., Dedene, G.: Formal concept analysis in knowledge processing: A survey on models and techniques. Expert systems with applications **40**(16), 6601–6623 (2013)

[16] Rouane-Hacene, M., Huchard, M., Napoli, A., Valtchev, P.: Relational concept analysis: mining concept lattices from multi-relational data. Annals of Mathematics and Artificial Intelligence **67**, 81–108 (2013)

[17] Rudolph, S., Săcărea, C., Troancă, D.: Conceptual navigation for polyadic formal concept analysis. In: IFIP International Workshop on Artificial Intelligence for Knowledge Management. pp. 50–70. Springer (2016)

[18] Stumme, G., Dürrschnabel, D., Hanika, T.: Towards ordinal data science. Transactions on Graph Data and Knowledge (TGDK) (2023)

[19] Voutsadakis, G.: Polyadic concept analysis. Order **19**, 295–304 (2002)