UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Coarse-refinement dilemma: on generalization bounds for data clustering

Yule Vaz

Tese de Doutorado do Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional (PPG-CCMC)

ICMC USP SÃO CARLOS

Yule Vaz

Coarse-refinement dilemma: on generalization bounds for data clustering

Doctoral dissertation submitted to the Institute of Mathematics and Computer Sciences – ICMC-USP, in partial fulfillment of the requirements for the degree of the Doctorate Program in Computer Science and Computational Mathematics. *FINAL VERSION*

Concentration Area: Computer Science and Computational Mathematics

Advisor: Prof. Dr. Rodrigo Fernandes de Mello

USP – São Carlos
November 2020

Yule Vaz

Dilema do sobre-refinamento: limites de generalização para agrupamento de dados

Tese apresentada ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP, como parte dos requisitos para obtenção do título de Doutor em Ciências – Ciências de Computação e Matemática Computacional. *VERSÃO REVISADA*

Área de Concentração: Ciências de Computação e Matemática Computacional

Orientador: Prof. Dr. Rodrigo Fernandes de Mello

USP – São Carlos
Novembro de 2020

ACKNOWLEDGEMENTS

Foremost, I would like to express my gratitude to my advisor and co-advisor, Prof. Rodrigo Fernandes de Mello and Carlos Henrique Grossi, who guide me in this project and in the academic life, being my friends afterall.

I would also like to thanks my friends of the research group: Ricardo Araújo Rios, Tatiane Nogueira Rios, Lucas de Carvalho Pagliosa, Adriele G. Biase, Daniel Moreira Cestari, Fausto Guzzo da Costa, Martha Dais Ferreira, Felipe Simões Lage Gomes Duarte, Cassio M. M. Pereira, and Paulo Henrique Ribeiro Gabriel for the great and fruitful conversations and debates.

Thank you my friends of São Carlos city, José Victor Ferronato Bueno, Marco Murillo, Laura Sodré Galvão Garcia, Wesley Colpani, Gleiciane Paiva, Thales Bonini, Alexandre Negrão, Rodrigo Barreto, Danilo Mendes Dias, Gabriel Ribeiro Camargo Ferreira, José Teixeira, Felipe Coelho, Rafael Ferrer, Rodrigo Carneiro Rodrigues, Rodrigo Brunelli, Thomaz Alberto, Rodrigo Correa, and the people of Solene's house, for sharing so many stories with me and supporting me in this path.

I also express my best regards to my friends of Americana and Campinas city, Leonardo Henrique Macieu, Guilherme Sanches, Thiago Aoqui, Eduardo Maia, Marcos Scantamburlo, Lucas Rosolen, Alan Wagner Gabriel, Allan Carmo de Palma, the people of São Pedro's neighborhood in Americana, Thomaz and Carol Uehara, thank you to share so many good moments.

Finally, I am extremely grateful for my parents Paulo Roberto Saura Vaz and Regiane Aparecida Brandão, as also for the other members of my family, who have supported, loved, cared and taught me in my life.

“Numquam ponenda est pluralitas sine necessitate.”
(William of Ockham)

ABSTRACT

VAZ, Y. **Coarse-refinement dilemma: on generalization bounds for data clustering**. 2020. 146 p. Tese (Doutorado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2020.

Machine Learning (ML) is typically organized into two main paradigms: (i) the Supervised Machine Learning (SML) to identify patterns from pre-labeled data, in which a loss function is used to adapt the corresponding model; and, (ii) the Unsupervised Machine Learning (UML) to organize data points in the absence of labels, taking similarity relations among elements into account. SML relies on well-consolidated theoretical frameworks, such as the Statistical Learning Theory (SLT) and the Algorithmic Stability (AS) to define assumptions, properties and convergence guarantees, allowing the comparison of different methods and, consequently, their improvements. Complementary, UML has been supported by investigations on Data Clustering (DC) and Hierarchical Clustering (HC) in order to define properties and improve their characterizations. Specifically, Kleinberg stated richness, scale-invariance and partition consistency as the necessary properties to define the DC problem, proving they do not hold simultaneously, while Ackerman, Ben-David and Loker explored other properties such as locality, and Carlsson and Mémoli developed stability and consistency frameworks for HC from metric spaces. To bring an additional contribution to UML, we considered topological spaces to design more general theoretical results given: (i) the invariance on topological spaces, more precisely isomorphism of homology groups, guarantees the properties of scale-invariance, partition consistency and locality; and (ii) this same invariance is inherited along less general spaces, such as the metric, thus allowing a more abstract clustering representation. Taking such invariance into account, we demonstrated that over-refined topologies endowed by DC and HC models lead to non-consistency in terms of their associated homology groups and, on the other hand, over-coarsed topologies devise consistent but unrepresentative homology groups, a phenomenon that we referred to as the Coarse-Refinement Dilemma (CRD). We then formulated DC and HC problems by employing Carlsson and Zomorodian's bidimensional persistent homology, with the first dimension corresponding to the HC levels and the second to the inclusion of new data, thus allowing a probabilistic study based on martingales process and subsequent formalization of generalization bounds. From such results, we contributed with the related work by: (i) defining lower and upper bounds for Carlsson and Mémoli's metric consistency; (ii) showing that Kleinberg's richness axiom must be relaxed otherwise over-refined or over-coarsed clusterings could be obtained; and, finally, (iii) defining unexpected changes in consistent topologies using what we named as Topological Concept Drift (TCD). An extensive set of experiments was performed to analyze the CRD and the TCD, including a brief study of a real-world scenario involving text documents. Results corroborated the usefulness in representing DC and HC problems using topological spaces, in detecting topology changes and the existence of CRD.

Keywords: Data clustering, Hierarchical clustering, Algorithmic stability, Persistent homology, Topological Concept drift.

RESUMO

VAZ, Y. **Dilema do sobre-refinamento: limites de generalização para agrupamento de dados**. 2020. 146 p. Tese (Doutorado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2020.

O Aprendizado de Máquina é tipicamente organizado em dois principais paradigmas: (i) o Aprendizado de Máquina Supervisionado (do inglês *Supervised Machine Learning* – SML) que identifica padrões de dados pré-rotulados, empregando uma função de perda para adaptar o modelo estatístico correspondente; e, (ii) o Aprendizado de Máquina Não-supervisionado (do inglês *Unsupervised Machine Learning* – UML) que organiza dados não rotulados por meio de relações de similaridade. O SML é suportado por arcabouços teóricos bem consolidados tais como a Teoria do Aprendizado Estatístico e a Estabilidade Algorítmica as quais definem suas principais suposições, propriedades e garantias de convergência, permitindo a comparação de diferentes métodos de SML e, consequentemente, seus aperfeiçoamentos. Complementarmente, UML conta com investigações sobre Agrupamento de Dados (do inglês *Data Clustering* – DC) e Agrupamento Hierárquico de Dados (do inglês *Hierarchical Clustering* – HC) definindo propriedades e caracterizando DC e HC de maneira mais adequada. Especificamente, Kleinberg estabeleceu riqueza, invariância à escala e consistência de partição como propriedades necessárias para a definição do problema de DC, provando a impossibilidade de satisfazê-las simultaneamente. Ackerman, Ben-David e Loker exploraram outras propriedades tais como a localidade, e Carlsson e Mémoli relataram resultados sobre a estabilidade e consistência para HC baseando-se na distância entre espaços métricos. A fim de complementar tais trabalhos, considera-se nesta tese de doutorado espaços topológicos para desenvolver um arcabouço teórico mais geral dado que: (i) a invariância de espaços topológicos, isomorfismos de grupos homológicos mais precisamente, garante as propriedades de invariância à escala, consistência de partição e localidade; e (ii) essa mesma invariância é herdada ao longo de espaços menos gerais, tal como o métrico, permitindo uma representação mais abstrata de agrupamento. Nesse contexto, foi demonstrado que topologias sobre-refinadas adotadas por modelos de DC e HC levam à não-consistência dos grupos homológicos associados. Por outro lado, topologias muito grosseiras levam à consistência, porém produzem grupos homológicos sem representatividade; nomeamos tal dilema como *Coarse-Refinement Dilemma* (CRD). Os problemas de DC e HC foram então formulados empregando a homologia persistente bidimensional de Carlsson e Zomorodian, sendo a primeira dimensão correspondente às hierarquias do HC e a segunda às inclusões de novos dados, o que permitiu o estudo probabilístico por meio de processos de martingales e subsequente formalização de um limite de generalização para DC e HC. Por meio desse resultado, complementamos os trabalhos relacionados: (i) definindo limitantes inferiores e superiores para a consistência métrica de Carlsson e Mémoli; (ii) mostrando que o axioma da riqueza de Kleinberg deve ser relaxado, caso contrário agrupamentos sobre-refinados e sobre-grosseiros serão obtidos; e, finalmente, (iii) definindo mudanças inesperadas em topologias consistentes como Mudanças Topológicas de Conceito (do inglês *Topological Concept Drifts* – TCD). Um extenso conjunto de experimentos foi executado para analisar o CRD e o TCD, incluindo um breve estudo de um cenário real envolvendo documentos textuais. Resultados corroboraram com a empregabilidade de espaços topológicos para a representação de problemas de DC e HC, sendo possível detectar mudanças topológicas e a existência do CRD.

Palavras-chave: Agrupamento de dados, Agrupamento hierárquico de dados, Estabilidade algorítmica, Homologia persistente, Mudanças Topológicas de Conceito.

LIST OF FIGURES

LIST OF ABBREVIATIONS AND ACRONYMS

AI	Artificial Intelligence
AS	Algorithmic Stability
CQM	Clustering Quality Measure
CRD	Coarse-Refinement Dilemma
DC	Data Clustering
ERM	Empirical Risk Minimization
HC	Hierarchical Clustering
ML	Machine Learning
MLP	Multilayer Perceptron
NMF	Non-negative Matrix Factorization
PHI	Persistent Homology Index
SC	Shattering Coefficient
SLT	Statistical Learning Theory
SML	Supervised Machine Learning
TCD	Topological Concept Drift
TDA	Topological Data Analysis
TS	Time Series
UML	Unsupervised Machine Learning
VC	Vapnik-Chervonenkis

CONTENTS

CHAPTER

1

INTRODUCTION

Machine Learning (ML) is a branch of Artificial Intelligence (AI) whose main goal is the identification of patterns in complex data which are infeasible or even impossible to be directly assessed by human cognition. This area of computer science provides tools to address different applications, such as disease diagnosis (SAJDA, 2006), genetic characterization (LIBBRECHT; NOBLE, 2015), epidemic forecasting (RAHMAWATI; HUANG, 2016), stock market prediction (PATEL *et al.*, 2015), and brain-computer interfaces (WANG *et al.*, 2018). ML is typically organized into two paradigms: (i) the Supervised Machine Learning (SML) that counts on well-consolidated theoretical frameworks to ensure learning from pre-labeled instances, such as the Statistical Learning Theory (SLT) (VAPNIK, 1995; LUXBURG; SCHÖLKOPF, 2011; MELLO; PONTI, 2018) and the Algorithmic Stability (AS) (BOUSQUET; ELISSEEFF, 2002; MUKHERJEE *et al.*, 2006); and (ii) the Unsupervised Machine Learning (UML) to analyze the structure of data spaces by using similarity measures, whose first theoretical results have been mostly formalized in the last two decades (KLEINBERG, 2002; BEN-DAVID; ACKERMAN, 2009; ACKERMAN; BEN-DAVID; LOKER, 2010; CARLSSON; MÉMOLI, 2010).

1.1 Learning in the supervised scenario

The goal of SML is to predict labels associated with data instances from some unknown joint probability distribution $P(\mathcal{X},\mathcal{Y})$, such that $x \in \mathcal{X}$ is an instance and $y \in \mathcal{Y}$ is its corresponding label. In this sense, SML algorithms attempt to model either classes (classification) or real values (regression) from instances in $P(\mathcal{X},\mathcal{Y})$. In order to devise such model, SML algorithms attempt to minimize some loss function $\ell(x,y)$ by adapting their parameters from a limited dataset $\{(x_1,y_1),\ldots,(x_n,x_n)\} \in \mathcal{X}\times\mathcal{Y}$ sampled from $P(\mathcal{X},\mathcal{Y})$, which is referred to as training set.

The theoretical frameworks of SLT (VAPNIK, 1995; MELLO; PONTI, 2018) and AS (BOUSQUET; ELISSEEFF, 2002; MUKHERJEE *et al.*, 2006) define conditions to ensure learning using the asymptotic approximation (uniform convergence) of the estimate

$\overline{\ell}_{\{(x_1,y_1),\ldots,(x_n,y_n)\}}(x,y)$ to $\mathbb{E}_{(x,y)\in P(\mathcal{X},\mathcal{Y})}[\ell(x,y)]$, as the sample size $n \to \infty$. SLT and AS rely on a set of assumptions among which it is worth to mention that the joint probability distribution $P(\mathcal{X},\mathcal{Y})$ must be constant (fixed or static such as referred in the literature (LUXBURG; SCHÖLKOPF, 2011)), that every pair (x_i,y_i) must be statistically independent from (x_j,y_j), for all $i \neq j$, and uniformly sampled from $P(\mathcal{X},\mathcal{Y})$. In summary, they formalize the effort involved in starting with some initial function until the convergence to the best as possible function, a.k.a. learning model, contained in the algorithm bias.

Regarding the algorithm bias, Vapnik (1995) formalized the Bias-Variance problem which determines the trade-off related to the effort of reducing the empirical risk $R_{\text{emp}}(f) = \frac{1}{n}\sum_{i=1}^{n}\ell(x_i,y_i)$, i.e., the estimated loss in some data sample, while maintaining similar classification results when predicting unseen instances that is referred to as expected risk $R(f) = \mathbb{E}_{(x,y)\in P(\mathcal{X},\mathcal{Y})}[\ell(x,y)]$, for some f contained in the algorithm bias $\mathcal{F}$. In that sense, $\mathcal{F}$ corresponds to some space of admissible functions (MELLO; PONTI, 2018) from which classification models are selected from, such that the space cardinality is proportional to the effort involved in converging to the best as possible solution.

Observe that one supervised algorithm may endow as much complexity as possible so that it would be capable of representing the universe set of functions, thus allowing to fit all training instances. Besides correctly classifying all training data, the model found from such large-scale space of admissible functions would miserably fail whenever used to classify unseen instances given its low probability of representing the unknown joint probability distribution $P(\mathcal{X},\mathcal{Y})$ what is referred to as overfitting (LUXBURG; SCHÖLKOPF, 2011). On the other hand, whenever one over-reduces such space complexity, the supervised algorithm has only access to a non-representative set of functions which is not even capable of modeling the training data. In such scenario, besides the convergence of $R_{\text{emp}}(f)$ to $R(f)$ is held, the best classification function found does not minimize the empirical risk as necessary, thus leading to what is referred to as underfitting (LUXBURG; SCHÖLKOPF, 2011).

For example, suppose the regression problem, illustrated in Figure 1, whose data follows a linear behavior with some additive white noise from which we can conclude that: (i) if the algorithm bias considers a set of high-order polynomial functions, it will certainly fit all training instances, but as it does not endow the necessary linear behavior, it will probably devise substantial errors when forecasting unseen elements (overfitting); otherwise, (ii) if the algorithm bias only takes the average function, then no adequate solution will be given neither over the training set nor over unseen data, but the associated errors will approach one another (underfitting).

This result is the main motivation for the formulation of the Empirical Risk Minimization (ERM) principle (VAPNIK, 1995) responsible for stating the conditions for the minimization of the divergence between the empirical and the expected risks, this is simply defined by the generalization term $G(f) = |R_{\text{emp}}(f) - R(f)|$ which must converge to $G \to 0$ as the sample size

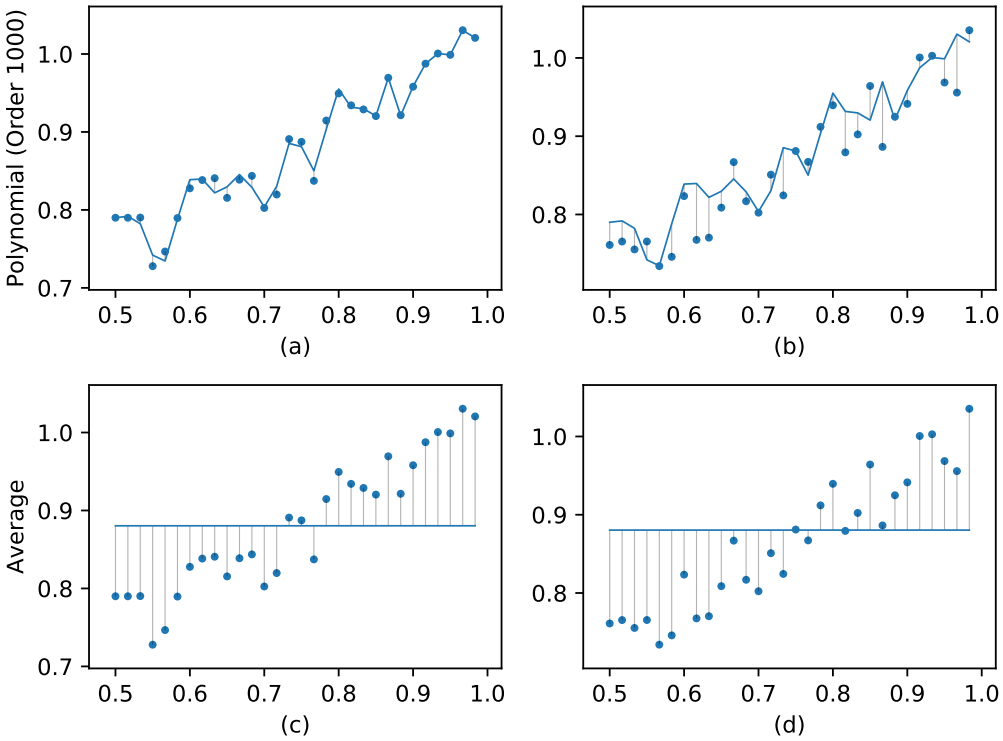


Figure 1 – Illustration of overfitting in: (a) points in the training set and the fitted curve (polynomial of order 1,000); (b) unseen points and its corresponding errors with respect to the trained model. Illustration of underfitting in: (c) points in the training set and their average value; (d) unseen points and the errors among them and such value (Adapted from Mello and Ponti (2018)).

$n \to \infty$.

Based on the ERM principle, Vapnik (1995) formalized additional concepts such as the Shattering Coefficient (SC) and the Vapnik-Chervonenkis (VC) dimension to quantify the complexity of some algorithm bias $\mathcal{F}$. SC is related to the growth of the number of distinct classifications as the sample size increases, and VC defines the greatest sample size for which all possible classifications are obtained. In that sense, the consistency of the ERM principle is only guaranteed if and only if SC is a polynomial function what consequently makes VC finite. This implies that some classification function $f \in \mathcal{F}$ is capable of producing all possible classification results for some sample size in general position (HAR-PELED; JONES, 2018), however that number of possibilities will be eventually bounded by the dimensionality composing the input space $\mathcal{X}$. If one adds up more hyperplanes to compose f, the complexity of SC also increases and its VC dimension, therefore there is a clear relation between SC and the number of the partitions formed to classify some dataset (HAR-PELED; JONES, 2018; MELLO; MANAPRAGADA; BIFET, 2019).

Complementary, AS defines convergence guarantees between the estimated error of an algorithm $\mathcal{A}$, in form $\overline{\ell}_{\mathcal{A}}(x,y)$, and its expected error $\mathbb{E}_{x\in\mathcal{X}}[\ell_{\mathcal{A}}(x,y)]$, taking into account bounded perturbations on input instances $X = \{x_1, x_2, \ldots, x_n\}$ that follow some fixed data distribution

$X \sim P(\mathcal{X})$, as follows:

1. Removal of the i-th element: $X^{/i} = \{x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n\}$;
2. Replacement of the i-th element: $X^{[i]} = \{x_1, x_2, \ldots, x'_i, \ldots, x_n\}$ with $x'_i \sim P(\mathcal{X})$;
3. Insertion of m new elements: $X^m = \{x_1, x_2, \ldots, x_n, x'_1, x'_2, \ldots, x'_m\}$ with $x'_1, \ldots, x'_m \sim P(\mathcal{X})$.

Under assumptions that depend on the type of stability, an algorithm is stable with respect to μ if $\mu(x, \phi(x)) < c$ given some perturbation $\phi : X \to X'$ and some divergence measure $\mu : X \times X' \to \mathbb{R}_+$. In this sense, some of the AS formalizations can guarantee probabilistic convergence between $f : X \to \mathbb{R}_+$ and $\mathbb{E}_{x \in \mathcal{X}}[f(x)]$ such as the Uniform Stability (BOUSQUET; ELISSEEFF, 2002), which holds whenever $\sup_{x \in X} |f(X) - f(X^{[i]})| < c_i$ for $i = 0, 1, \ldots, n$. From such stability and assuming that the underlying probability distribution is independent and identically sampled (i.i.d.), f converges to $\mathbb{E}[f]$ according to McDiarmid (1989)'s inequality:

$$P(|f(X) - \mathbb{E}_{X \subset \mathcal{X}}[f(X)]| > \varepsilon) \leq 2 \exp\left(\frac{-2\varepsilon}{\sum_{i=1}^{n} c_i^2}\right), \tag{1.1}$$

thus guaranteeing generalization.

1.2 Learning in the unsupervised scenario

Algorithm Stability relies only on $P(\mathcal{X})$ so that labels are not even necessary to proceed with guarantees, allowing its use to support a theoretical framework for Data Clustering (DC) and Hierarchical Clustering (HC) problems. However, the formalization of DC and HC problems requires some measure capable of comparing clustering models as studied in (CARLSSON; MÉMOLI, 2010), which considers models as ultrametric spaces in order to allow the use of the Gromov-Haursdorff distance (GROMOV; LAFONTAINE; PANSU, 1981) to evaluate their respective divergences.

Although the efforts by Carlsson and Mémoli (2010) in proposing stability and consistency results for HC problems and the definition of clustering properties by Kleinberg (2002), Ackerman, Ben-David and Loker (2010), there is no unified formalization connecting those scientific outcomes in order to provide a single and complete clustering framework from initial conditions to generalization bounds. This gap is the main motivation for this PhD thesis which adopts a topological perspective on DC once properties by Kleinberg (2002), Ackerman, Ben-David and Loker (2010) naturally rise from homology isomorphism which is guaranteed by the stability and consistency results proved along this manuscript.

Specifically, the DC properties (KLEINBERG, 2002; ACKERMAN; BEN-DAVID; LOKER, 2010) considered along this PhD thesis are: (i) scale-invariance – partitions should be

maintained after scaling the distance matrix (such as if data are zoomed in or out); (ii) partitions consistency – partitions should be maintained after reducing intracluster and/or increasing intercluster distances; (iii) locality – subpartitions must be preserved whenever clustering is performed over the corresponding subset of data samples; and, finally, (iv) richness – the distance function must allow the production of any desired partition.

In this sense, Kleinberg (2002) proves, for a finite and fixed dataset and disregarding the ambient space, that scale-invariance, partitions consistency, and richness are not simultaneously satisfied. However, we here question the need for the property of richness given it allows the obtained partitions to be non-consistent and non-stable. Instead, we consider the need for relaxing such property and only assume that topologies should maintain scale-invariance, partitions consistency and locality (ACKERMAN; BEN-DAVID; LOKER, 2010). For example, in order to prove their consistency theorem for single-linkage algorithm, Carlsson and Mémoli (2010) avoid over-refined partitions considering only measurable metric spaces coarser than a specific measurable set.

Complementary, Carlsson and Mémoli (2010) formulate HC using dendrogram structures and show they are analogous to ultrametric spaces. From that, they define properties to characterize HC for linkage-class algorithms, which they claim to be closely related to Kleinberg's, proving that the single linkage criterion is the only one satisfying them. One of such properties (to mention, the third of Theorem 18, pg. 1451 (CARLSSON; MÉMOLI, 2010)) states that dendrograms produced from Hierarchical Clustering must disregard subdendrograms corresponding to singleton elements. Our theoretical framework strengthens such property as it proves that over-refined partitions (singleton and quasi-singleton partitions) are prone to topological variances whenever unseen data are included. From the corresponding topological features, we also prove that it is possible to bound the consistency result proposed by Carlsson and Mémoli (2010), thus generalizing the DC and HC problems from a topological perspective.

1.3 Our approach

Topological spaces are mathematical structures described with respect to contiguity among elements, i.e., in terms of their neighborhoods. Therefore, we consider that, in order to perform some clustering, an algorithm must adopt open sets to define similarities among elements so the clustering procedure identifies contiguous groups. The collection of open sets (typically open balls (MUNKRES, 2000)) endowed in clustering models forms a topological space from points sampled from some unknown but fixed probability distribution $P(Z)$ of an unknown (also fixed) measurable topological space (Z, τ_Z). As $P(Z)$ and (Z, τ_Z) are unknown, we can only compare topological spaces produced from $X \sim P(Z)$ with another $X' \sim P(Z)$ given some perturbation $\phi : X \to X'$, as previously discussed in this section on Algorithmic Stability.

The most important relation considered between such spaces is the homeomorphism

which is a continuity bijection (MUNKRES, 2000) responsible for transforming some topological space without "cutting" or "gluing" its elements, thus being, at a first glance, an adequate candidate to compare clustering models. However, such relation requires the explicit definition of a map from one topological space to another and, as $P(Z)$ is unknown, such a map $\phi : X \to X'$ cannot be explicitly defined.

Homology is capable of algebraically and numerically characterizing topological spaces though, and even if not precisely in the sense that invariance in homology does not imply homeomorphism, it is still an adequate candidate for clustering comparison (HATCHER, 2000; CARLSSON, 2009). Homology is referred to as the association of algebraic structures to other mathematical objects (such as topological spaces) and, in this context, homology groups are abelian groups that identify features of such spaces. These groups are associated with the dimension of such features, depicting, for example, the connected components at the 0-th dimension, holes at the 1-th dimension and voids at greater dimensions (HATCHER, 2000).

For instance, a tridimensional spherical surface S^2 has homology groups $H_0(S^2) = \mathbb{Z}^1$, $H_1(S^2) = 0$ and $H_2(S^2) = \mathbb{Z}^1$, as it only presents one connected component, no holes and one void, respectively. The ranks associated with a p-th group (defined by its superscript) are named as p-th Betti numbers (HATCHER, 2000), which we prove in this PhD thesis to be a measure over the σ-algebra formed by the combination of all possible singular homology groups, and then renamed it as p-th Betti measure. Therefore, p-th Betti measures can be adopted to compare underlying structures, i.e., the endowed open sets of clustering models, allowing the assessment of features beyond connected components in data clustering. Two issues rise in terms of homology though: (i) the representation of clustering hierarchies, and (ii) the definition of data perturbations.

Persistent homology was then adopted in order to solve those issues as it allows the homology over filtrations corresponding to a sequence of inclusions of topological spaces. In other words, persistent homology depicts the persistence of homology groups along some ordered variable, such as time or radius. Take the torus data illustrated in Figure 2 as an example, where points are connected to each other whenever they are enclosed by the same open ball in order to reconstruct some topological space. In the context of persistent homology, note that: (i) partitions of the torus will last until they are merged, defining their persistence, and (ii) its hole will be represented with some radius $r > 0$ until being vanished when $r' > r$ (for some great value of r).

Hence, to employ such technique we assume that the insertion of new data in DC and HC models produces filtrations along which the persistence of homology groups can be studied. An HC model is then represented as a bifiltration (CARLSSON; ZOMORODIAN, 2009), with one dimension corresponding to hierarchies and another to data insertions, such as illustrated in Figure 3, from which some bidimensional persistent homology can be devised in order to evaluate changes in p-th Betti measures. Although there are two studies proving stability for persistent homology, none of them considers variations over topological spaces as data are

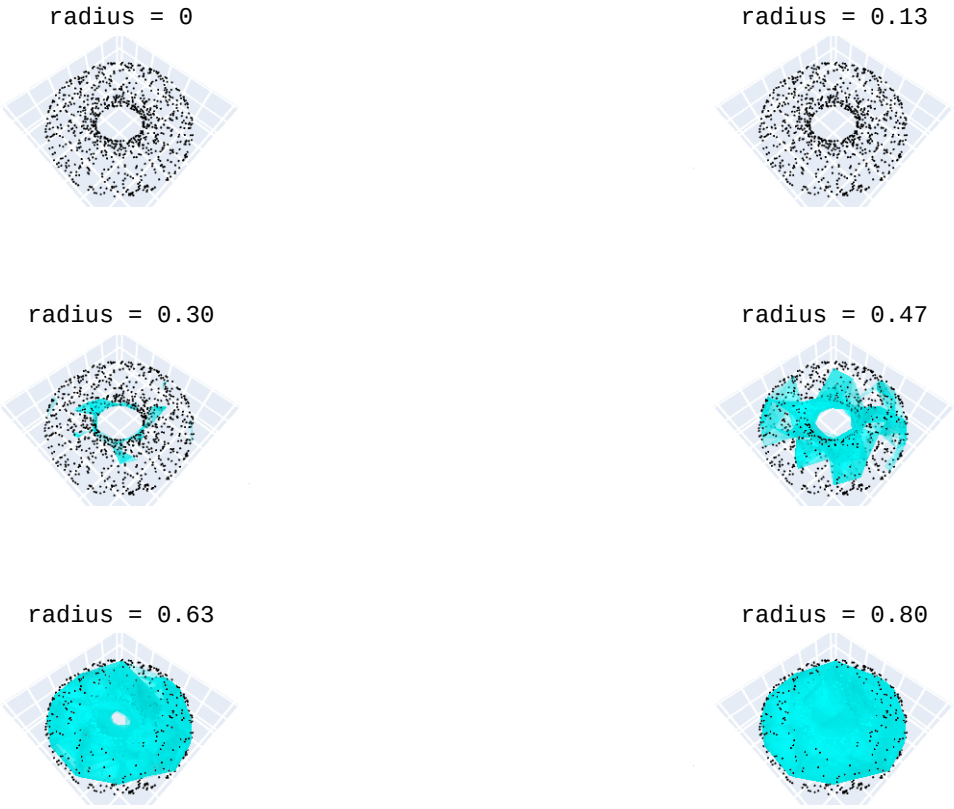


Figure 2 – Illustration of the simplexes built on top of a data set acquired uniformly from a toroidal topology and along the filtration associated with open balls of radius $0, 0.13, 0.30, 0.47, 0.63$ and 0.8 (Adapted from Khasawneh and Munch (2016)).

inserted (COHEN-STEINER; EDELSBRUNNER; HARER, 2007; CHAZAL *et al.*, 2009).

By defining the HC bifiltration, we employ Azuma (1967)'s inequality to prove that over-refined topological spaces produce divergent p-th Betti measures whenever data is subject to perturbations relative to data insertions. Therefore, considering some clustering model, if its corresponding Betti measures present bounded changes following a martingale sequence (VILLE, 1939), then the clustering algorithm is capable of generalizing from data.

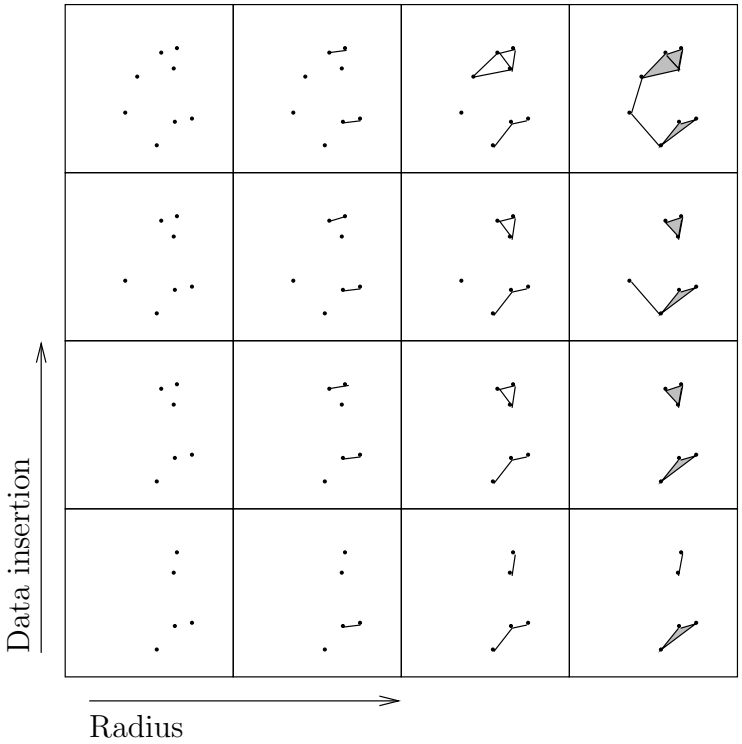


Figure 3 – Simplexes created in a bifiltration in which the associated radius increases along the horizontal frames and new data is collected along the vertical one (Adapted from Carlsson and Zomorodian (2009)).

1.4 Our results

The Coarse-Refinement Dilemma (CRD) is, then, the trade-off given as follows: whenever topological spaces are over-refined, corresponding p-th Betti measures are not consistent, thus the clustering algorithm is unable to generalize from data; on the other hand, if they are over-coarsed, then p-th Betti measures will be always the same within all dimensions, thus leading to a trivial scenario that fails to represent data structures. Then, it is possible to choose the appropriate set of clusters by assessing their stability/consistency and their representability. Moreover, we prove for the 0-th homology group that if some topological space is stable in terms of Betti measures, all spaces enclosing it will be also stable, something invalid for homology groups at greater degrees.

Considering the CRD, we define Topological Concept Drifts (TCD) as changes in topological features which are not associated with the inherent problem of over-refinement, i.e., they are the direct result of relevant modifications in topologies. We then performed experiments demonstrating the Coarse-Refinement Dilemma on multiple datasets and also studied the effects of the Topological Concept Drifts. Results allowed us to confirm the existence of CRD and identify drifts in the considered topological spaces. Although this PhD thesis models HC regarding only its hierarchy, Coarse-Refinement Dilemma can be extended to study density-based clustering algorithms (ESTER *et al.*, 1996), as density variations also devise filtrations, and to its hierarchical model proposed in (CAMPELLO; MOULAVI; SANDER, 2013). In fact, Carlsson and Zomorodian (2009) work allows a multiparameter study for clustering as long as these

parameters are ordinal and correspondent with topological inclusions, i.e., given a parameter r, if $r_i \leq r_j$, then two topological spaces (X, τ_i) and (X, τ_j) must respect $(X, \tau_i) \subseteq (X, \tau_j)$.

1.5 Contributions

With the goal of design a novel theoretical framework defining learning and generalization bounds for DC and HC in terms of topological variances, this PhD thesis resulted in the following contributions:

- A topological definition for both DC and HC problems;
- The characterization of the CRD;
- The definition and demonstration of generalization bounds for DC and HC;
- The presentation of upper and lower bounds for Carlsson and Mémoli (2010) consistency result for single-linkage algorithm, based on topological features;
- The definition of Topological Concept Drifts.

1.6 Claims

Also, based on the results of this thesis we claim that:

- Kleinberg (2002)'s richness axioms must be relaxed or disregarded when adopting a consistency/stability study of DC and HC;
- Over-refined clustering leads to non-consistency/non-stability and over-coarse ones lead to unrepresentability of the data;
- Homology groups invariance along data sampling are enough to guarantee scale-invariance, partitions consistency and locality;
- Topological variations which are not associated with the intrinsic instability of over-refined clustering comprise concept drifts occurred by changes in the underlying topological space from which data are acquired.

1.7 Thesis organization

This PhD thesis is organized as follows: Chapter 2 introduces the theoretical foundations adopted in this PhD thesis; Chapter 3 discusses the developments on theoretical frameworks for Data Clustering and Hierarchical Clustering; Chapter 4 defines and demonstrates the Coarse-Refinement Dilemma, formalizing generalization bounds for DC and HC algorithms; Chapter 5

details the experimental results performed to show CRD and TCD; Chapter 6 draws concluding remarks and future directions; references are listed in Bibliography; and Appendices include (i) the notations of this thesis in Appendix A, (ii) the dictionary of terms in Appendix B, (iii) the proofs of proposed theorems in Appendix C, (iv) the datasets images in Appendix D, (v) the barcode plots in Appendix E, (vi) charts with the measurements of TCD in Appendix F, and, finally, (vii) word clusters related with the experiment of document semantic changes G.

CHAPTER

2

THEORETICAL FOUNDATION

2.1 Initial considerations

In the data clustering context, metric spaces and topological spaces can adequately represent data similarities, defining partitions (clusters) in the absence of data labels. As topological spaces can be induced by metric ones, a topology-based framework for data clustering can count on more general space features which are present in spaces of lower hierarchy (e.g., vector, Hilbert and metric ones). In fact, the Coarse-Refinement Dilemma (CRD) claims that if the topological representation of a clustering model is overly refined, then there is no guarantee that the topological features will be stable or consistent. Therefore, given a clustering model built from a space of lower hierarchy, if its induced topology is overly refined, such stability and consistency are not guaranteed either.

2.2 Spaces

A space is a mathematical structure formed by a set of elements whose relations should respect a set of properties defining its type. In other words, such structures allow to study how the elements in their sets are related with one another taking their properties into account. For instance, the Euclidean space formed by an n-tuple $(x_1, x_2, \ldots, x_n)$ of real numbers, corresponding to the vector space $\mathbb{R}^n$, called points. Euclidean spaces have an affine structure such that, given a point p and a vector $\vec{v}$, they allow translations $A \times \vec{A} \to A : (a, \vec{v}) \mapsto a + \vec{v}$ of their points which, since the nature of the field $\mathbb{R}^n$, are invariant in terms of the permutations of operands (i.e., are commutative) and operations (i.e., they are associative), still allowing the identity and the inverse element. Therefore, as illustrated in Figure 4, it is possible to factorize a translation $a + \vec{v}$ onto $a + \vec{v}_1 + \vec{v}_2 + \cdots + \vec{v}_k$, which is invariant given the permutation of its operands and operations and it is also possible to define a kernel set.

Another example is the Hamming space which is typically composed of 2^N binary

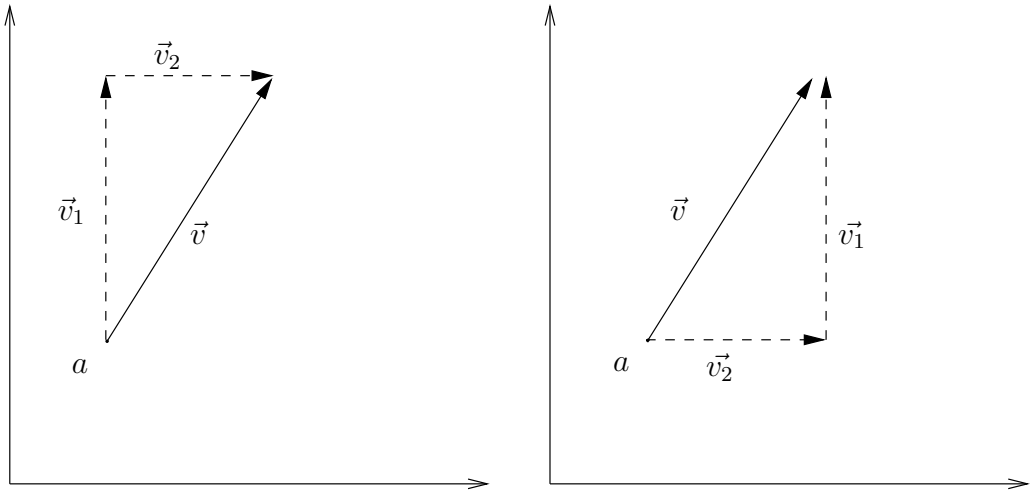


Figure 4 – Invariance of vectors in relation to the permutation of operands and decomposition of the vector $\vec{v}$ onto $\vec{v}_1$ and $\vec{v}_2$ (Adapted from Steinbruch and Winterle (1987)).

strings $\{(000\ldots0),(100\ldots0),(010\ldots0),(110\ldots0),\ldots,(111\ldots1)\}$ with length N, structured as illustrated in Figure 5. Formally, such space is a vector space over a finite binary field $\mathbf{GF}(2)$ (or Galois Field of Order 2) which means that, in practice, the addition and multiplication operators correspond to, respectively, the logical **OR** and **AND**, holding: the addition and multiplication associativeness, commutativeness, identity, inverse and the distributiveness of the multiplication over an addition. Moreover, a Hamming space can be embedded into a metric one when considering Hamming distances, i.e., the number of corresponding positions that elements are pairwise different in binary strings. Such embedding allows a summarized similarity measure between binary strings, however it loses the vector structure of the space, implying the lack of information about vector directions. It will also characterize, as illustrated in Figure 5, the notion of neighborhood as, given binary strings $\vec{b}$, it defines closed balls $B(\vec{b},r)$ around $\vec{b}$ containing the binary strings whose distances from $\vec{b}$ are less or equal to an integer r which is known as radius.

In this sense, spaces can be represented using a hierarchical scheme such that the lower level spaces can be endowed with properties of higher level ones, such as in the aforementioned example of the Hamming distance. Note that, although this endowment produces new representations for the original space structures, it loses other spatial information. For instance, as illustrated in Figure 6, the embedding of a vector space in a metric one results in the definition of neighborhoods called open balls, but with the cost of disregarding directions, angles and, consequently, linear dependencies of vector spaces. As it is also illustrated in Figure 6, a metric embedding into a topological space maps the open balls into a collection of open sets named as topology, losing any distance information. As consequence, the relevant space feature comes from the arrangement of its elements in the topology, not distances among them anymore.

Such space hierarchy, as illustrated in Figure 7, motivates the employment of topological spaces in the context of this PhD thesis in order to provide a more abstract definition for data clustering, allowing the study of such problem disregarding any distance information and

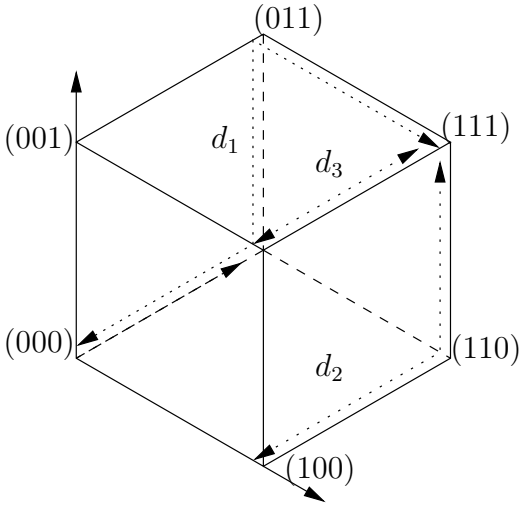


Figure 5 – Tridimensional Hamming space. Note that (000) belongs to the closed ball $B_3((111))$ centered in (111) with radius 3, (100) belongs to the closed ball $B_2((111))$ and (101) to $B_1((111))$ (Adapted from Klein (2013)).

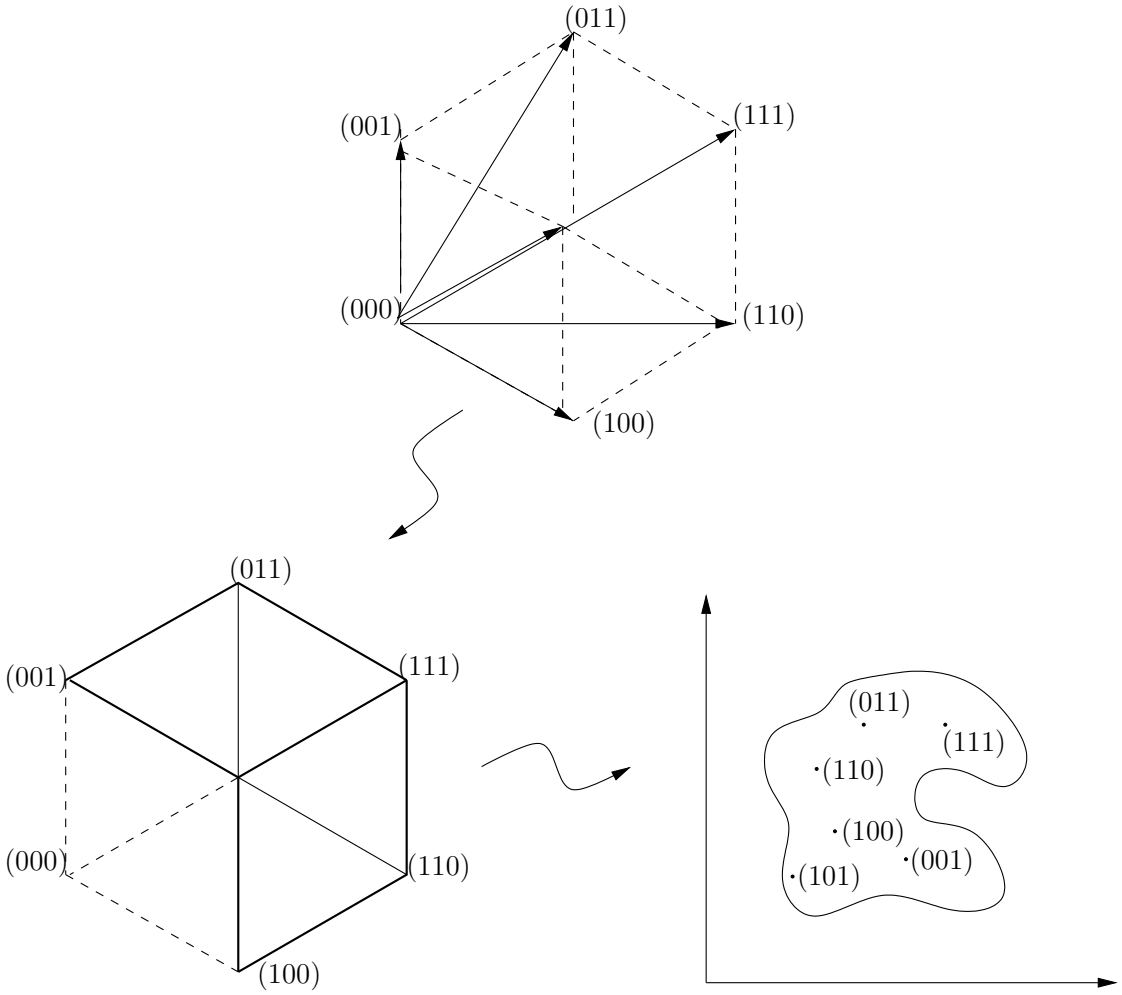


Figure 6 – Embedding of the Hamming vector space onto a Hamming metric space composed of the closed ball $B_2((111))$, and, sequentially, onto an arbitrary topological space (Adapted from Matousek (2002)).

considering only the neighborhoods produced for a particular dataset. It is worth to mention that, in practice, a distance function will be typically employed by a clustering algorithm exactly to define such neighborhoods. This restriction does not impact the topology-based framework considered in this thesis whatsoever as, in fact, the properties of topological spaces are present along all embeddings resultant from lower level spaces when addressing data clustering.

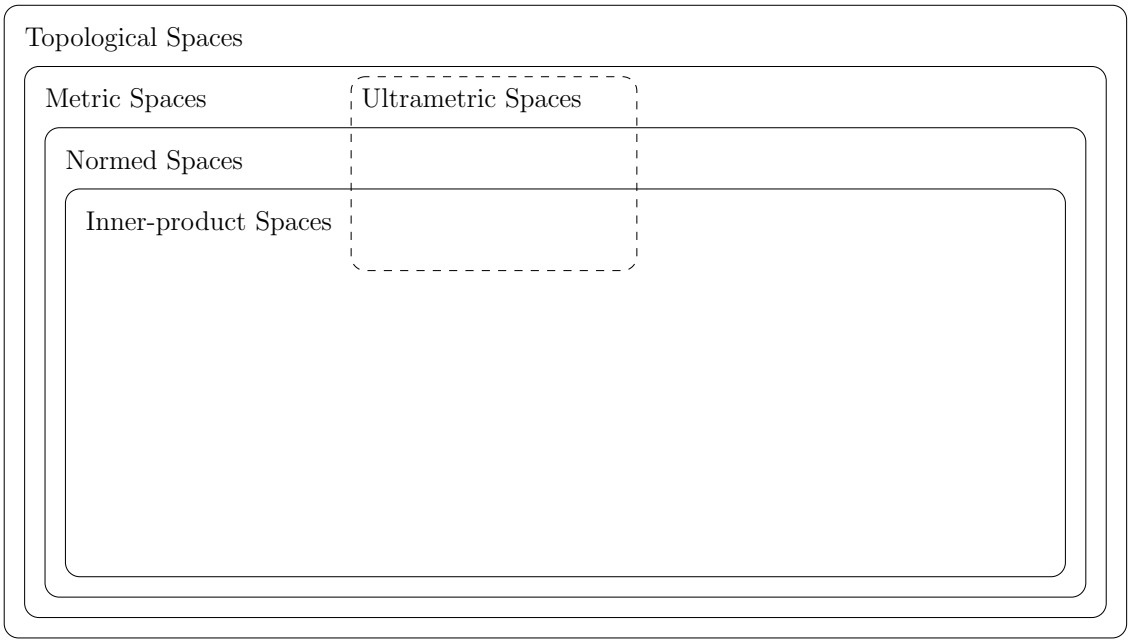


Figure 7 – Hierarchy of spaces, from topological to inner-product ones (Adapted from Formaggia, Saleri and Veneziani (2012)).

2.2.1 Vector and Hilbert spaces

Vector spaces are structures which count on a set V of n-tuples $\vec{v} = (v_1, v_2, \ldots, v_n)$, called vectors, endowed with the addition $V + V \to V : (\vec{v}, \vec{u}) \mapsto \vec{v} + \vec{u}$ and with the scalar multiplication $F \times V \to V : (\alpha, \vec{v}) \mapsto \alpha\vec{v}$. Such structures are built so that the set V forms an abelian group $(V, +)$ defined as follows:

Definition 1 (Abelian group)**.** Given a set V, limited or not, and elements $\vec{v}, \vec{u}, \vec{w} \in V$, an abelian group is defined by the tuple $(V, +)$ respecting the following properties:

- Closure under $+$: $\vec{v} + \vec{u} \in V$;
- Associativity: $\vec{v} + (\vec{u} + \vec{w}) = (\vec{v} + \vec{u}) + \vec{w}$;
- Commutativity: $\vec{v} + \vec{u} = \vec{u} + \vec{v}$;
- Identity existence: $\exists 0$ s.t. $\vec{v} + 0 = \vec{v}$, $\forall\, \vec{v} \in V$;
- Inverse existence: $\forall \vec{v} \in V$, $\exists\, (-v)$ s.t. $\vec{v} + (-\vec{v}) = 0$.

In this sense, translations of vectors maintain their results even if such operations and/or operators are permutated, besides, the existence of inverse operations allows the definition of opposed directions in the vector space.

In addition, the structure of scalar multiplication must assure field properties defined as:

Definition 2 (Field)**.** Given a set F of scalars and $\alpha, \beta, \gamma \in F$, a field is formed by the triple $(F, +, \cdot)$ respecting the following properties:

- Closure under $+$ and $\cdot$: $\alpha + \beta \in F$ and $\alpha \cdot \beta \in F$;
- Associativity: $\alpha + (\beta + \gamma) = (\alpha + \beta) + \gamma$, and $\alpha \cdot (\beta \cdot \gamma) = (\alpha \cdot \beta) \cdot \gamma$;
- Commutativity: $\alpha + \beta = \beta + \alpha$, and $\alpha \cdot \beta = \beta \cdot \alpha$;
- Identities existence: $\exists 0$ s.t. $\alpha + 0 = \alpha$, $\forall\, \alpha \in F$, and $\exists 1$ s.t. $1 \cdot \alpha = \alpha$, $\forall\, \alpha \in F$;
- Inverses existence: $\forall\, \alpha \in F$, $\exists -\alpha$ s.t. $\alpha + (-\alpha) = 0$, and $\forall\, \alpha \in F - \{0\}$, $\exists \alpha^{-1}$ s.t. $\alpha \cdot \alpha^{-1} = 1$;
- Distributivity: $\alpha \cdot (\beta + \gamma) = \alpha \cdot \beta + \alpha \cdot \gamma$.

As the scalar multiplication of vector spaces must respect the aforementioned properties of fields, it is typically said that a vector space is built over a field F. A vector space is then defined as:

Definition 3 (Vector space)**.** A set V of vectors endowed with a vector addition $V + V \to V : (\vec{v}, \vec{u}) \mapsto \vec{w}$ and with a scalar multiplication $F \times V : (\alpha, \vec{v}) \mapsto \vec{u}$ forms a vector space over the field F if and only if the following properties hold:

- $(V, +)$ is an abelian group;
- $\forall\, \alpha \in F$ and $\forall\, \vec{v}, \vec{u} \in V$, $\alpha(\vec{v} + \vec{u}) = \alpha\vec{v} + \alpha\vec{u}$;
- $\forall\, \alpha, \beta \in F$ and $\forall\, \vec{v} \in V$, $(\alpha + \beta)\vec{v} = \alpha\vec{v} + \beta\vec{v}$;
- $\forall\, \alpha, \beta \in F$ and $\forall\, \vec{v} \in V$, $\alpha(\beta\vec{v}) = (\alpha\beta)\vec{v}$;
- $\forall\, \vec{v} \in V$, $1\vec{v} = \vec{v}$ where 1 is the multiplicative identity of the field F.

The most important result of such properties is that they define the linear transformation $F \times V \times V \to V : (\alpha, \vec{v}, \vec{u}) \mapsto \alpha\vec{v} + \vec{u}$, which is composed of translations and scaling over vectors.

Vector spaces V can be also endowed with the inner product $\langle \cdot, \cdot \rangle : V \times V \to V$ such that $\langle \vec{v}, \vec{u} \rangle = \sum_{i \in I} v_i u_i$, in which I is a countable set of indices. Such operation must respect the following properties:

- Conjugate symmetry: $\langle \vec{v}, \vec{u} \rangle = \langle \vec{u}, \vec{v} \rangle^*$ such that * is the conjugate operator;

- First argument linearity: $\langle \alpha\vec{v}+\vec{u},\vec{w}\rangle = \alpha\langle\vec{v},\vec{w}\rangle + \langle\vec{u},\vec{w}\rangle$;
- Positive definiteness: $\langle\vec{v},\vec{v}\rangle > 0\ \forall\vec{v}\in V-0$.

If such space is complete, then it is named as a **Hilbert space** (SCHOLKOPF; SMOLA, 2001). With the inner product operation, the notion of angles arises in the vector space, and, consequently, it characterizes the linear dependency as it also defines the rotation transformation over vectors.

In the context of Supervised Machine Learning (SML), the inner product allows the definition of linear classifiers, for example: considering an input space X with $\vec{x}\in X$ and a binary output space $Y=\{+1,-1\}$, a Perceptron model (ROSENBLATT, 1957) with neuron weights $\vec{w}$ and bias θ applies a decision function $f: X\to Y$ over a transformation in form $(\vec{w},\vec{x},\theta)\mapsto\langle\vec{w},\vec{x}\rangle+\theta$ in order to perform a classification. Note that, the inner product allows rotation operations, $\vec{w}$ defines the angle relative to the canonical base of the Perceptron hyperplane, and θ corresponds to its translation along the input space. The decision function f then classifies the two half-spaces formed by such hyperplane as either $+1$ or -1.

Remark 1. It is well known that the Perceptron algorithm is unable to solve nonlinear problems (e.g. XOR), motivating the development of the Multilayer Perceptron (MLP) (ROSENBLATT; BUFFALO, 1961; WERBOS, 1974; RUMELHART; HINTON; WILLIAMS, 1986), which models multiple hyperplanes transversing the input space. As formalized in the context of the Statistical Learning Theory (VAPNIK, 1995; LUXBURG; SCHÖLKOPF, 2011; MELLO; PONTI, 2018), extra hyperplanes can be used to classify more complex input spaces in exchange of an increase in the shattering coefficient, thus relaxing generalization guarantees (Bias-Variance Dilemma (VAPNIK, 1995; LUXBURG; SCHÖLKOPF, 2011; MELLO; PONTI, 2018)). Note that, for the sake of curiosity, the hyperplanes will define partitions on the input space so that the addition of new hyperplanes will refine such partitions. Therefore, the embedded topological space will be over-refined, as illustrated in Figure 8, and its consistency is not guaranteed for associated homology groups, as proved in Chapter 4. Although this apparent relationship between the Bias-Variance and the Coarse-Refinement dilemmas is not completely addressed in this thesis, it is an interesting topic to approach in future studies.

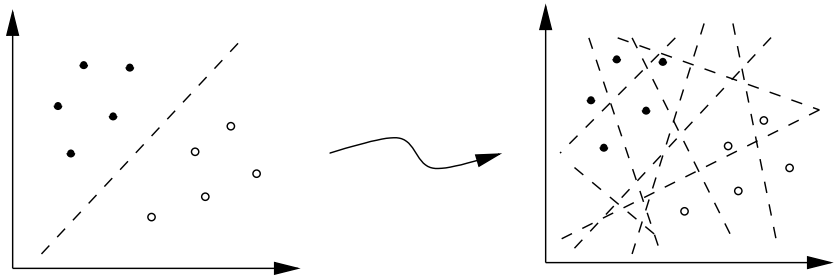


Figure 8 – Data space shattered by hyperplanes forming multiple partitions (Adapted from Guarascio, Manco and Ritacco (2019)).

In addition to the inner product, norms can also be endowed in vector spaces characterizing the notion of length. Given a vector space V over a field F, for all $\vec{v},\vec{u} \in V$ and all $\alpha \in F$, a norm $\|\cdot\| : V \to \mathbb{R}$ must respect the following properties:

- Triangle inequality: $\|\vec{v}+\vec{u}\| \leq \|\vec{v}\| + \|\vec{u}\|$;
- Absolute scalability: $\|\alpha\vec{v}\| = |\alpha|\|\vec{v}\|$;
- Positive definiteness: $\|\vec{v}\| = 0$ iff $\vec{v} = \vec{0}$, otherwise, $\|\vec{v}\| > 0$.

A norm can naturally rise from the inner product space given $\|\vec{v}\| := \sqrt{\langle \vec{v},\vec{v}\rangle}$ and, therefore, all inner product spaces are also normed vector spaces (such as illustrated in Figure 7). Nonetheless, the norm of a vector that begins at the tip of vector $\vec{v}$ and ends at $\vec{u}$ calculates the distance between the point $(v_1, v_2, \ldots, v_n)$ and $(u_1, u_2, \ldots, u_n)$ and, in this sense, a normed space can be endowed with a metric (Figure 7).

2.2.2 Metric and ultrametric spaces

Metric spaces are structures formed from relations among a set of elements which disregard the use of algebras and coordinate systems and, consequently, the characterization of directions, angles, and linear dependencies. In other words, such relations, known as distances, are enough to construct a metric. Formally, the metric spaces are defined as:

Definition 4 (Metric space)**.** Given a set X of elements and a function $d_X : X \times X \to \mathbb{R}^+$, known as the distance function, a metric space is the tuple (X, d_X) such that, for all $x, x' \in X$, d_X respects the following properties:

- Identity: $d_X(x,x') = 0$ iff $x = x'$;
- Symmetry: $d_X(x,x') = d_X(x',x)$;
- Triangular inequality: $d_X(x,x'') \leq d_X(x,x') + d_X(x',x'')$.

Note that a normed vector space V can naturally endow a metric d_V by the application of a norm over the subtraction of the vectors $\vec{v} \in V$. More precisely, given $\vec{v},\vec{u},\vec{w} \in V$ and $d_V := \|\vec{v}-\vec{u}\|$:

- $\|\vec{v}-\vec{u}\| = 0$ iff $\vec{v} = \vec{u} \iff d_V(\vec{v},\vec{u}) = 0$ iff $\vec{v} = \vec{u}$;
- The absolute scalability of normed vector spaces implies in $\|(-1)(\vec{v}-\vec{u})\| = \|\vec{u}-\vec{v}\| \iff d_V(\vec{v},\vec{u}) = d_V(\vec{u},\vec{v})$;
- $\|\vec{v}-\vec{w}\| = \|(\vec{v}-\vec{u})+(\vec{u}-\vec{w})\| \leq \|\vec{v}-\vec{u}\| + \|\vec{u}-\vec{w}\| \iff d_V(\vec{v},\vec{w}) \leq d_V(\vec{v},\vec{u}) + d_V(\vec{u},\vec{w})$,

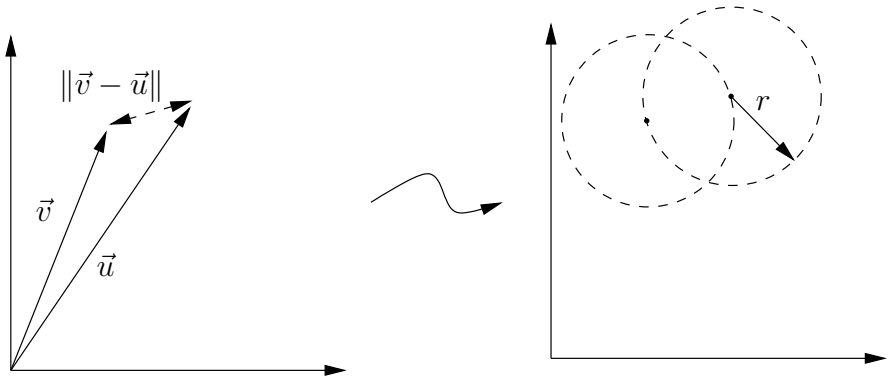


Figure 9 – Open balls of a metric space (on the right) induced by a normed space (on the left).

and, therefore, metric spaces can be induced from any normed vector space, as illustrated in Figure 9.

For instance, in the *k-nn* algorithm (ALTMAN, 1992), a metric space $(\mathbb{X}, d_{\mathbb{X}})$ is induced from a dataset $\mathbb{X}$ which is assumed to form a normed vector space. Then, taking $X \subset \mathbb{X}$ as the training set, $X' \subset \mathbb{X}$ as the test set and a constant r, for every point $x' \in X'$, there is a set $B(x', r) = \{x \in X | d_{\mathbb{X}}(x', x) < r\}$ known as open ball, which is defined as follows:

Definition 5 (Open ball)**.** Given a metric space (X, d_X), an open ball around a point $x \in X$ is a set which contains all points $x' \in X$ whose distance from x is less then a radius r, i.e., $B(x, r) = \{x' \in X \mid d_X(x, x') < r\}$.

Remark 2. Complementary, a **closed ball** is formed by the set centered at some $x \in X$ with radius r such that $D(x, r) = \{x' \in X \mid d_X(x, x') \leq r\}$.

Therefore, the *k-nn* algorithm finds, for every training point $x' \in X'$, an open (or closed) ball centered at x' containing only its k-nearest neighbors $x \in B_k(x', r)$ and, in order to classify x', *k-nn* assigns the majority class of $B_k(x', r)$ to it, as illustrated in Figure 10.

Remark 3. Typically, when considering metric spaces, machine learning techniques and theoretical frameworks assume that a dataset $X \subset \mathbb{X}$ forms a **finite metric space** (X, d_X), i.e., a metric space built on top of a finite set, as, in practice, there are only a limited number of data points available.

In Data Clustering (DC), metric spaces are usually assumed to obtain distance functions to measure similarities among data points. For instance, the DBSCAN algorithm (ESTER *et al.*, 1996) finds clusters based on the density of each data point neighborhood. More precisely, given a radius r, a density parameter ρ, and a dataset $Z \subset \mathbb{Z}$, the DBSCAN analyzes all open balls $B(z, r)$ around each data point $z \in Z$ and if $|B(z, r)| \geq \rho$, z is attached to every other point in its neighborhood. In this context, not only the neighborhood of z configures a similarity criterion to form clusters, as also its density.

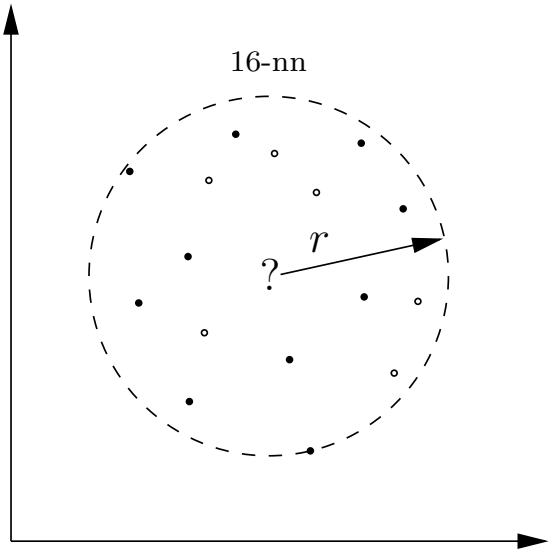


Figure 10 – Illustration of the procedure of a k-nn (with $k = 16$) algorithm in assigning a class for an unknown data point (Adapted from Wang *et al.* (2017)).

Remark 4. Some studies adopt a distance function disregarding the ambient space, i.e., the space surrounding a mathematical object, in which lies data points. For instance, let a distance function be defined as $d : I \times I \to \mathbb{R}$, with I being the set of indices associated with the data points in X, instead of $d : X \times X \to \mathbb{R}$. Therefore, no information about the surrounding of points is given.

Fixing a single radius in order to produce a clustering can disregard other relevant cluster structures, what motivates the development of Hierarchical Clustering (HC) techinques. In HC, a sequence of radii $r_0 < r_1 < \cdots < r_f$, determining the level of model hierarchy, is adopted to built a collection of open balls $\mathcal{B} = \{B(Z,z_1), B(Z,r_2), \ldots, B(Z,r_f)\}$ from which a collection of clusters $\mathcal{C} = \{C_{r_1}, C_{r_2}, \ldots, C_{r_f}\}$ is defined, such that $B(Z,z_1) \subseteq B(Z,z_2) \subseteq \cdots \subseteq B(Z,z_f)$ and $C_{z_1} \subseteq C_{z_2} \subseteq \cdots \subseteq C_{z_f}$. For instance, the HDBSCAN (CAMPELLO; MOULAVI; SANDER, 2013), a hierarchical variation of DBSCAN, considers multiple density-based clustering associated with radii $r_0, r_1, \ldots, r_f$ thus leading to a hierarchical model. In addition, linkage-based algorithms (SIBSON, 1973; DEFAYS, 1977) also adopt such HC paradigm but disregarding the density of open balls.

Such hierarchical model can be characterized by an ultrametric space as follows:

Definition 6 (Ultrametric space)**.** Given a set of elements X and a distance function $u_X : X \times X \to \mathbb{R}^+$, an ultrametric space is the tuple (X, u_X) such that, given $x, x' \in X$, u_X must respect the following properties:

- Identity: $u_X(x,x') = 0$ iff $x = x'$;
- Symmetry: $u_X(x,x') = u_X(x',x)$;

- Strong triangular inequality: $u_X(x,x'') \leq \max\{u_X(x,x'), u_X(x',x'')\}$.

In fact, the ultrametric space structure naturally defines the hierarchical scheme, given open balls $U(x,r) = \{x' \in X \mid u_X(x,x') < r\}$, $U(x,r) \cap U(x',r') \neq \varnothing$ iff $U(x,r) \subseteq U(x',r')$ or $U(x',r') \subseteq U(x,r)$, proved as follows:

1. First, let us prove that if $x' \in U(x,r)$ then $U(x,r) = U(x',r)$: Given points $x'' \in U(x,r)$ such that $u_X(x,x'') < r$, as $x' \in U(x,r) \iff u_X(x,x') < r$, we have that $u_X(x'',x') \leq \max\{u_X(x'',x), u_X(x,x')\} < r$. Therefore, all points $x'' \in U(x,r)$ belong to $U(x',r)$. Conversely, assuming $x \in U(x',r)$, all points $x'' \in U(x',r)$ belong to $U(x,r)$, hence, $U(x,r) = U(x',r)$;
2. As $U(x,r) \cap U(x',r') \neq \varnothing$, there are points $x'' \in U(x,r) \cap U(x',r')$, and, therefore, $u_X(x,x') < r$ and $u_X(x',x'') < r''$. According to Property 1, $U(x',r') = U(x'',r')$ and $U(x,r) = U(x'',r)$, implying that, if $r \leq r'$, $U(x'',r) \subseteq U(x'',r')$, hence, $U(x,r) \subseteq U(x',r')$, conversely, if $r' \leq r$, $U(x,r) \subseteq U(x',r')$.

Such hierarchical structures of ultrametric spaces can be organized, as presented in (CARLSSON; MÉMOLI, 2010), in tree-like structures called **dendrograms**. Formally, given a set X of elements, dendrograms are defined as the pair (X,ξ) such that $\xi : [0,r) \to \Gamma_X$ with Γ_X being a partition built up on top of X. Therefore, in the hierarchical clustering scenario, as illustrated in Figure 11, (X,ξ) associates the partitions of each hierarchical level, i.e., the open balls of the corresponding ultrametric space, to the radius considered in their construction.

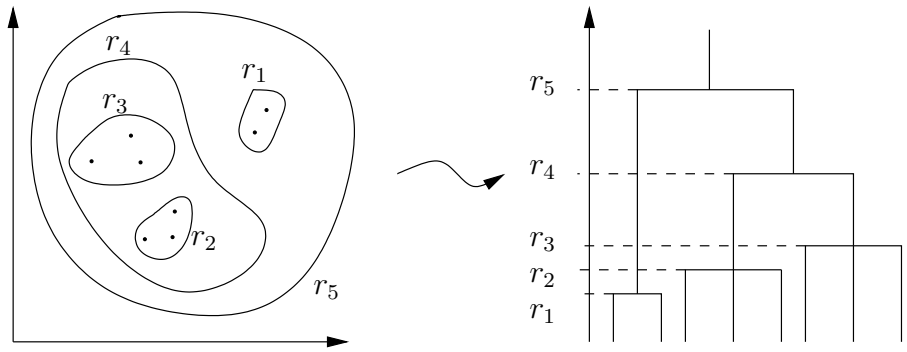


Figure 11 – Illustration of a transformation from an ultrametric space to a dendrogram (Based in Carlsson and Mémoli (2010)).

In addition, a metric space (X,d_X) (and, consequently, an ultrametric space) can induce a topological one. In this sense, its topology, i.e., its collection of open sets, is given by all subsets formed by the union of the open balls of (X,d_X).

2.2.3 Topological spaces

Topological spaces are mathematical structures that describe the relations among elements relying on a collection of **open** or **closed sets**. For instance, the set of elements $X =$

$\{a,b,c\}$ associated with a collection $\tau_X = \{\varnothing, X, \{a\}, \{b,c\}\}$ is organized in τ_X, such that b is related to c, and a is related to itself only. Note that, those arrangements are enough to represent correspondences, hence, if such topology is induced from a metric space, distance relations are disregarded. Formally, a topological space is defined as:

Definition 7 (Topological space)**.** Given a set of elements X and a collection of open subsets of X, τ_X, named as topology, a topological space is a tuple (X, τ_X) that respects the following properties:

- Triviality: $X \in \tau_X$ and $\varnothing \in \tau_X$;
- Closure under arbitrary unions: For $U_i \in \tau_X$ with $i \in I$ and I a set of indices, $\bigcup_{i \in I} U_i \in \tau_X$;
- Closure under limited intersections: For $U_1, U_2 \in \tau_X$, $U_1 \cap U_2 \in \tau_X$.

Remark 5. A topological space can be also defined considering closed sets by assuming the complements of those properties, which are then redefined as triviality, closure under arbitrary intersection and closure under limited unions.

Remark 6. Some sets, called **clopen sets** are closed and open simultaneously, as for example the set $A = (0,1)$ of a topological space (X, τ_X) with $X = (0,1) \cup (2,3)$ and τ_X a induced topology in $\mathbb{R}$. Note that $A^c = (2,3)$, which is also an open set, and therefore, also clopen.

In this sense, a **subspace** of a topological space is the subset $S \in X$ endowed with a topology induced by τ_X, such that $\tau_S = \{S \cap U \mid U \in \tau_X\}$. For instance, suppose $X = \{a,b,c,d,e\}$ with $\tau_X = \{\varnothing, X, \{a,b\}, \{c\}, \{d,e\}\}$, assuming $S = \{a,d,e\}$, (S, τ_S) forms a subspace of (X, τ_X) such that $\tau_S = \{\varnothing, S, \{a\}, \{d,e\}\}$, then there exists an **inclusion map** $\iota : (S, \tau_S) \hookrightarrow (X, \tau_X)$, i.e., $(S, \tau_S) \subseteq (X, \tau_X)$, so that (S, τ_S) is a refinement of (X, τ_X) which is a coarsening of (S, τ_S).

In addition, the relations represented by a topology are usually characterized in terms of neighborhoods:

Definition 8 (Neighborhood)**.** Given a point $x \in X$, a neighborhood is a subset $V \subset X$ such that there exists another open subset $U \subset V$ with $V \supseteq U \ni x$.

For instance, consider $X = \{(0,1), (2,3]\}$ and $\tau_X = \{\varnothing, X, (0,1), (2,3]\}$. Given $x = 0.5$, there is an open set $(a,b) \subseteq (0,1)$ which contains x. Thus, $(0,1)$ is a neighborhood of $x = 1$. Conversely, given $x = 3$, there is no open set (a,b) containing x, so $(a,3]$, with $a > 2$, is not a neighborhood of $x = 3$. In addition, it is also possible to characterize τ_X by means of its **connected components**.

The definition of such structures rely on connected spaces which are topological spaces that cannot be represented by a union of two or more disjoint non-empty subsets. For instance, the former example is not a connected space as its topology is formed by the disjoint non-empty subsets $(0,1)$ and $(2,3]$. Conversely, given $X' = \{((0,1], (1,2))\}$ and $\tau_{X'} = \{\varnothing, X', (0,1], (1,2)\}$,

$(X', \tau_{X'})$ is connected as no disjoint non-empty subset is capable of forming such topology. A connected component is, then, the coarsest subspace topology formed from a connected subset, which is a connected subspace topology of a topological space. In the former example, (X, τ_X) can have many connected subsets such as $(0.5, 0.7)$,$(2.4, 2.9)$ and $(0.1, 0.11)$, although there are only two connected components, to mention, $(0, 1)$ and $(2, 3)$.

Another important property of topological spaces is compactness, which generalizes the concept of closed sets in Euclidean spaces. Intuitively, compact topological spaces are closed and bounded, for example, if $X = [0, \infty)$ with a trivial topology $\tau_X = \{\varnothing, X\}$, the topological space (X, τ_X) is closed but not bounded, however, if $X = [0, 10]$ the topological space endowed with the trivial topology is compact. Formally, compactness is defined as:

Definition 9 (Compactness)**.** A **cover** $C = \{U_i \mid i \in I\}$ of some topological space (X, τ_X) is an indexed family of subsets U_i of (X, τ_X) with $i \in I$, such that $\bigcup_{i \in I} U_i = X$, then a subcover is a subset of C which also covers (X, τ_X). A topological space is compact if every of cover C has a finite subcover.

For instance, the trivial topology of $X = [0, 1)$ produces a non-compact topological space as there are open subcovers that can arbitrarily approximate to 1 on the rightmost bound, and if $X = [0, 1]$, its trivial topology forms a compact topological space given all covers have a finite subcover. Note that, without losing generality, DC and HC models can rely on compact topological spaces, which support a measure if they are Hausdorff separable. In that sense, a Hausdorff space is a topological space defined as follows:

Definition 10 (Hausdorff space)**.** A topological space (X, τ_X) is a Hausdorff space if and only if, for every pair of points $x, x' \in X$, there are neighborhoods $U \ni x$ and $V \ni x'$, such that $U \cap V = \varnothing$.

Then, we assume in this PhD thesis, without losing generality, that DC and HC models endow Hausdorff compact spaces.

Still in the context of DC and HC, the open balls typically adopted to build up a model induce a collection of neighborhoods around each data point. In order to compare the clusters formed from neighborhoods associated with different sets of points, which are assumed to be acquired from the same i.i.d. principle, there is a function $f : (X, \tau_X) \to (X', \tau_{X'})$ mapping a topological representation of some clustering partition of dataset X into another produced from a second dataset X'. As a consequence, f can be studied to evaluate the changes from (X, τ_X) to $(X', \tau_{X'})$.

In this sense, continuity is a function property described as follows:

Definition 11 (Continuous function)**.** A function $f : (X, \tau_X) \to (X', \tau_{X'}))$ is continuous, if for all open sets $V \subset (X', \tau_{X'})$, the pre-image $f^{-1}(V)$ is an open set U, such that $U \subset (X, \tau_X)$.

For example, let (X, τ_X) be a topological space in the usual topology with $X = \{(0,1),(1,2)\}$, and $(X', \tau_{X'})$ be another topological space with $X' = \{(0,2)\}$, if $f((X, \tau_X)) = (X', \tau_{X'})$, then there is a continuous f such that for every point in $(X', \tau_{X'})$, exists an open set $U \subset X$ whose pre-image is an open set $f^{-1}(V \subset X')$. Now, suppose $g((X', \tau_{X'})) = (X, \tau_X)$, given the neighborhood $N_{X'}(1)$ of $(X', \tau_{X'})$, there is no open set associated with a pre-image $g(U)$, e.g., the open set $(1-\varepsilon, 1+\varepsilon) \subset X'$. Intuitively, a continuous map is capable of merging neighborhoods as illustrated in Figure 12.

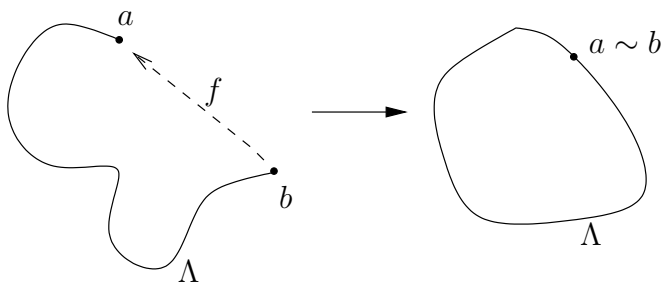


Figure 12 – Continuous mapping f applied over the curve Λ "gluing" the point a to the point b (Adapted from Hatcher (2000)).

An important continuous map studied in Topology is the **quotient map**, which counts on an **equivalence relation** $\sim$, on set X, respecting the following properties:

- Reflexivity: $x \sim x \; \forall x \in X$;
- Symmetry: $x \sim x' \iff x' \sim x \; \forall x, x' \in X$;
- Transitivity: $x \sim x'$ and $x' \sim x'' \implies x \sim x'' \; \forall x, x', x'' \in X$.

An equivalence relation induces a set named **equivalence class** that when defined on an element x is given by $[x] = \{x' \in X \mid x' \sim x\}$, i.e., a set with all elements equivalent to x. Then, a quotient map $\chi : (X, \tau_X) \to (X, \tau_X)/\sim$ transforms a topological space in a **quotient space** $(X, \tau_X)/\sim = \{[x] \mid x \in X\}$, i.e., the set of all equivalence classes in (X, τ_X).

Intuitively, the quotient map is a continuous map which "glues" all elements of the same equivalence class into a single one. For example, suppose an undirected graph G, illustrated in Figure 13, composed of vertices $W = \{a, b, c, d, e, f, g\}$ and edges $E = \{\{a,b\},\{a,c\},\{a,d\},\{f,e\}\}$. Note that such graph induces a topological space (W, τ_W), so its topology τ_W contains all of its combinations of connected components. Thus, given a quotient map $\chi : (W, \tau_W) \to (W, \tau_W)/\sim$, with the relation $x \sim x'$ satisfied if and only if x and x' belong to the same connected component, the quotient space is $(W, \tau_W)/\sim = \{[a],[f],[g]\}$.

In addition, given a continuous function $f : (X, \tau_X) \to (X', \tau_{X'})$, if f^{-1} is also continuous, f is referred to as **homeomorphic**, thus corresponding to transformations that do not merge nor divide any neighborhood in the topological space, as shown in Figure 14. For instance, still

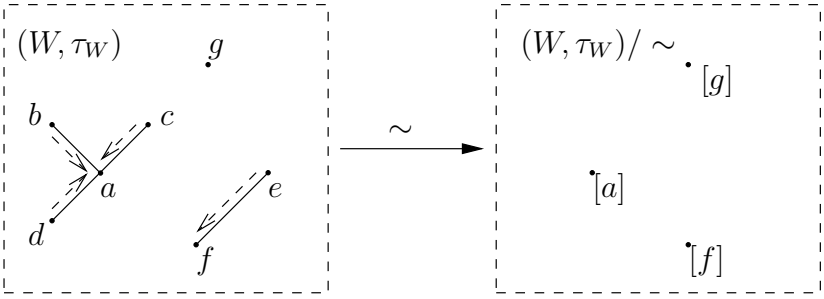


Figure 13 – Quotient map χ applied over the topological spaces produced by the graph G, which the quotient space $(W, \tau_W)/\sim = \{[a], [f], [g]\}$(Adapted from Hatcher (2000)).

considering the former example, if $X' = \{(1,10),(20,40)\}$ then f would be homeomorphic. In this sense, homeomorphisms could be considered in order to compare the topological spaces induced by different clustering partitions. However, in DC and HC problems, the probability distribution of the dataset is unknown and, therefore, f cannot be explicitly defined.

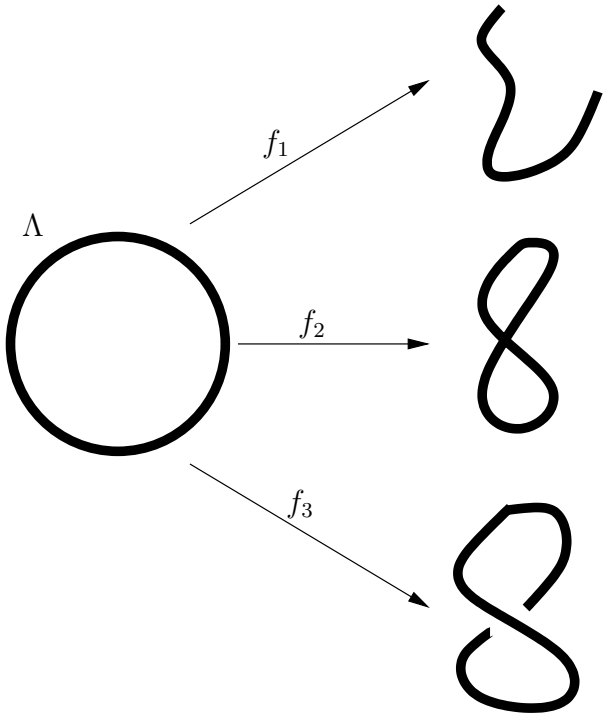


Figure 14 – Examples of, respectively: (i) f_1 – non-homeomorphic function, cutting the curve Λ, (ii) f_2 – non-homeomorphic functions "gluing" the curve Λ, and, (iii) f_3 – homeomorphic function "twisting" the curve Λ (Adapted from Hatcher (2000)).

This motivates the adoption of a framework which does not require the definition of such a map in order to allow the comparison among clustering partitions. In this sense, homology provides such framework as it relies on algebraic and graph representations of topological spaces, as well as on the quantification of their topological features.

2.3 Homology

Suppose a disk D and a circle R on top of which a topological space is built up. As illustrated in Figure 15, it is possible to define a function $f : D \to R$ which maps every neighborhood in the interior of D onto an interval lying on the circle R. Such transformation is clearly not continuous nor homeomorphic. Although, consider that the function f is not defined such as, for instance, in the construction of a topological space from data points. Such scenario does not allow the explicit definition of a function f that transforms, for example, D in R.

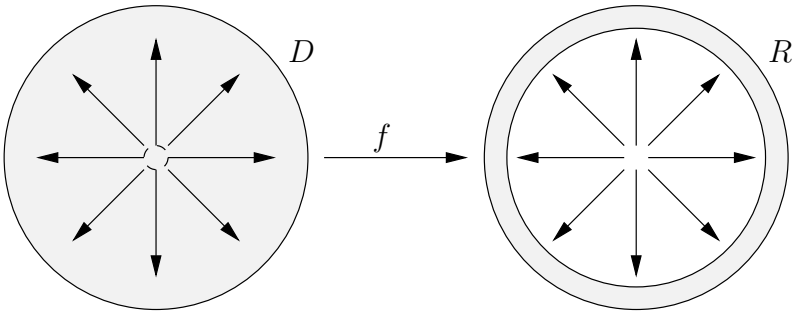


Figure 15 – Transformation of a disc D into a ring R by a non-homemorphic function f. (Adapted from Hatcher (2000)).

There are features that can be compared though, such as the number of connected components and the number of holes in D and R. Therefore, as D and R have the same number of connected components, often defined as 0-dimensional "holes", they are equivalent when such feature is considered. However, a circle R contains a hole and a disk D does not, hence, they are not equivalent with respect to 1-dimensional holes. Complementary, if a spherical surface S and a solid sphere B are studied, there is only one connected component and no holes for both objects. However, differently from B, S presents a 2-dimensional hole, also known as void, i.e., holes of dimension greater than or equal two.

In this sense, regardless the properties of continuity and homeomorphisms, topological spaces can be compared by using such topological features. Homology is a branch in mathematics (HATCHER, 2000) whose goal is to describe topological spaces using algebraic objects, such as groups and modules, in order to provide complementary structural information on topological spaces. For instance, a topological space can be described as **CW-cells** which are a generalization of oriented graph elements, such as vertices and edges. CW-cells does not only rely on vertices and oriented edges but also on oriented areas and n-dimensional volumes, such as illustrated in Figure 16.

A set of CW-cells, as illustrated in Figure 16, forms a **CW-complex** C_p whenever elements are attached to their $(i-1)$-dimensional boundaries, being $i < p$ the dimension of a CW-cell, forming the inclusions $C_0 \subset C_1 \subset \cdots \subset C_p$, such that each C_i contains CW-cells with dimensions of, at most, i. In this sense, the CW-complex is capable of building up the corresponding topological space by means of a quotient map χ that attaches its CW-cells. For

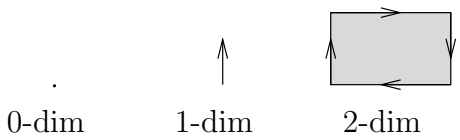


Figure 16 – Examples of 0,1 and 2-dimensional CW-cells. (Adapted from Hatcher (2000)).

instance, suppose the following CW-complex:

$$\begin{array}{ccc} A & \xrightarrow{a} & B \\ \downarrow{\scriptstyle c} & & \downarrow{\scriptstyle d} \\ C & \xrightarrow{b} & D. \end{array}$$

If a quotient map χ appends the edge a on b and c on d following their orientations, it results in a 3-dimensional torus, as illustrated in Figure 17. Now suppose the following CW-complex:

$$\begin{array}{ccc} A & \longrightarrow & B \\ \downarrow{\scriptstyle a} & & \uparrow{\scriptstyle b} \\ C & \longrightarrow & D, \end{array}$$

the quotient map which only appends a on b, respecting their orientations so that $D = A$ and $C = B$ (Figure 17), forms a Möebius strip. Note that CW-complexes are also often represented as topological spaces.

Topological features can then be defined along the inclusion $C_0 \subset C_1 \subset \cdots \subset C_p$, considering abelian groups that correspond to each complex C_i, known as the homology group $H_i(X)$ of a topological space (X, τ_X). In this sense, there exists a sequence of homomorphisms ∂_p, denominated boundary operators, between complexes, such that:

$$\ldots \xrightarrow{\partial_{p+1}} C_p \xrightarrow{\partial_p} C_{p-1} \xrightarrow{\partial_{p-1}} \ldots \xrightarrow{\partial_2} C_1 \xrightarrow{\partial_1} C_0 \xrightarrow{\partial_0} 0$$

is defined as **chain complex**. A boundary operator is a map $\partial_i : C_i \to C_{i-1}$, such that it transforms a complex C_i into a lower dimensional complex C_{i-1} containing the boundaries of C_i. For example, an oriented open triangle $\Delta^1 = \{\{A \to B\}, \{B \to C\}, \{C \to A\}\}$ forming the following CW-complex:

$$\begin{array}{ccc} A & & \\ \downarrow & \nwarrow & \\ B & \longrightarrow & C \end{array}$$

has boundaries $\mathrm{im}(\partial_1) = \{\{A\}, \{B\}, \{C\}\}$.

Remark 7. More precisely, CW-complexes are topological spaces whose constructions are given by maps $\phi_\alpha : S^i \to C_i$ attaching the topological space of C_i with a collection of disks D_α^{i+1}, thus producing the quotient space of the disjoint union $C_i \sqcup D_\alpha^{i+1}$. Such quotient space is defined as $C_i \sqcup D_\alpha^{i+1}$ and, given $x \in C_i$, such that $x \sim \phi_\alpha(x)$, for $x \in \partial D_\alpha^p$, as illustrated in Figure 18. For

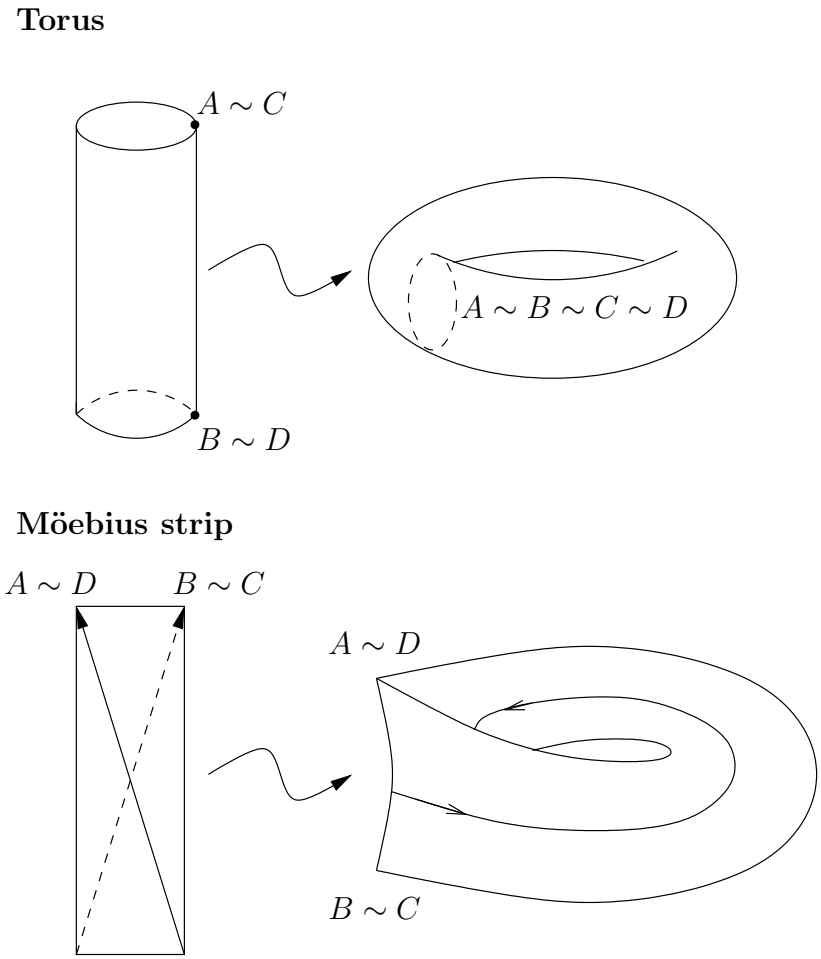


Figure 17 – Illustration of the: (i) Torus formed from the Diagram 2.3, represented above, and, (ii) Möebius strip formed from the Diagram 2.3, represented below. (Adapted from Hatcher (2000)).

example, given a CW-complex C_0 defined in terms of three vertices $\{\{0\},\{1\},\{2\}\}$, there is a collection of three disks $D^1_\alpha = \{D_1, D_2, D_3\}$ with $D_j = [0,1]$, for $j = 1,2,3$, such that $\phi_\alpha(D_j)$ provides an edge formed by D_j that connects two vertices of C_0, thus generating the open triangle $\Delta = \{\{1,2\},\{2,3\},\{3,1\}\}$ from $C_0 \sqcup_\alpha D^1_\alpha$.

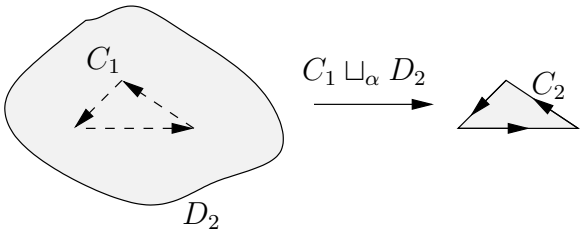
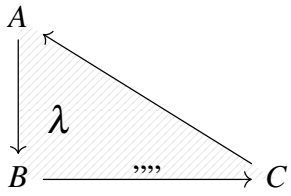


Figure 18 – Illustration of a C_2 CW-complex being formed by the quotient of the disjoint union $C_1 \cup_\alpha D_2$ between a 1-dimensional CW-complex and a 2-dimensional disk. (Adapted from Hatcher (2000)).

Now suppose the oriented and closed triangle $\Delta^2 = \{\{A \to B \to C \to A\}\}$ which forms

the following CW-complex:



(2.1)

The boundaries $\text{im}(\partial_3)$ and $\text{im}(\partial_2)$ of Δ^2 are given by the boundaries of the oriented area λ in Diagram 2.1 which is equal to Δ^1, i.e., $\text{im}(\partial_2) = \Delta^1 = \{\{A \to B\}, \{B \to C\}, \{C \to A\}\}$. However Δ^1 and Δ^2 result in different topological features, as Δ^1 has a hole and Δ^2 does not.

Algebraically, the direct sum of groups of the oriented edges $\{\{A \to B\}, \{B \to C\}, \{C \to A\}\}$ is $a_1(B-A) + a_2(C-B) + a_3(A-C) = 0$, which leads to the following system of equations:

$$\begin{cases} a_1 = a_2 \\ a_3 = a_1 \\ a_2 = a_1 \end{cases} \tag{2.2}$$

The System of Equations 2.2 has as solution the space $a_1 = a_2 = a_3$ with rank equals one, then, as $\text{im}(\partial_2) = 0$, the cycle represented in Δ^1 corresponds to the kernel $\ker(\partial_1) = \mathbb{Z}$. However, topologically, Δ^2 does not have such cycle as it presents a 2-dimensional cell attaching its edges, i.e., $\text{im}(\partial_2) = \ker(\partial_1) = \mathbb{Z}$. In this sense, the cycles in Δ^2 are defined by $\ker(\partial_1)/\text{im}(\partial_2)$, i.e., by using ∂_1 except for those attached to 2-dimensional CW-cells.

In this sense, a homology group is defined as:

Definition 12. Given a topological space (X, τ_X) and a chain complex $C(X)$, a homology group is a quotient abelian group defined as:

$$H_i(X) := \ker(\partial_i)/\text{im}(\partial_{i+1}),$$

and whose elements are called **homology classes**.

Hence, a homology group finds the cycles in an i-dimensional complex which are not attached by $(i+1)$-dimensional cells. For instance, suppose the CW-complex present in Figure 19, designed such that each vertex is associated with a data point. When the 0-dimensional homology group $H_0(X)$ is considered, it is possible to evaluate how many connected components are present in such a complex. In addition, $H_1(X)$ represents how many holes are present in the respective topological space. DC and HC can adopt such groups in order to quantify the topological features which rise from a specific model. Such quantification is often defined by the rank, $\text{rank}(H_i(X))$, of the homology group of interest, named ***i*-th Betti number**.

However, DC and HC problems do not have orientable surfaces and, therefore, they can rely on the employment of simpler structures to form complexes, instead of CW-cells. Such structures are named simplices, being the p-dimensional generalization of triangles defined as:

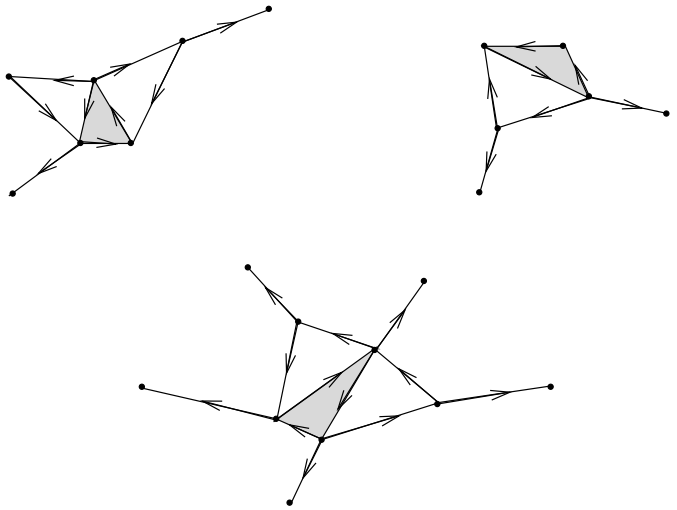


Figure 19 – Illustration of the complex associated with a dataset X, in which can be identified the homology groups $H_0(X) = \mathbb{Z}^3$, $H_1(X) = \mathbb{Z}^5$ and $H_2(X) = 0$ and their respective Betti numbers $\beta_0(X) = 3$, $\beta_1(X) = 5$ and $\beta_2(X) = 0$ (Adapted from Hatcher (2000)).

Definition 13. A p-simplex is a p-dimensional convex hull built up on top of non-collinear p vertices $v_1, v_2, \ldots, v_p \in \mathbb{R}^p$ such that:

$$\sigma = \left\{ \sum_{i=1}^{p} a_i v_i \mid \sum_{i=1}^{p} a_i = 1 \text{ with } a_i \geq 0 \, \forall \, i \right\}.$$

Hence, a simplicial complex is defined as:

Definition 14. A simplicial complex K is a set of simplices $\sigma = \{\sigma_1, \sigma_2, \ldots, \sigma_n\}$ such that:

- Every face is a simplex in K;
- Every non-empty intersection of any two simplices $\sigma_j, \sigma_k \in K$ is a face of σ_j and σ_k.

Let us consider the homology group $H_0(X)$ associated with the simplicial complex representing models built up from the single linkage algorithm, such as illustrated in Figure 20. As the radius increases, more simplices are created through the connection of data points and, also, the number of connected components decreases as clusters are attached to each other. Hence, there is a variation on the topological features (i.e., Betti number) along the topological inclusions formed by such a process. This PhD thesis demonstrates that such features are not stable nor consistent for the most refined topologies. Such demonstration relies on persistent homology to allow the study of the birth and death of homology classes (EDELSBRUNNER; HARER, 2008).

2.3.1 Persistent homology

The consequences of an over-refinement upon the stability and consistency of an HC model can be studied through persistent homology, which identifies the birth and death of

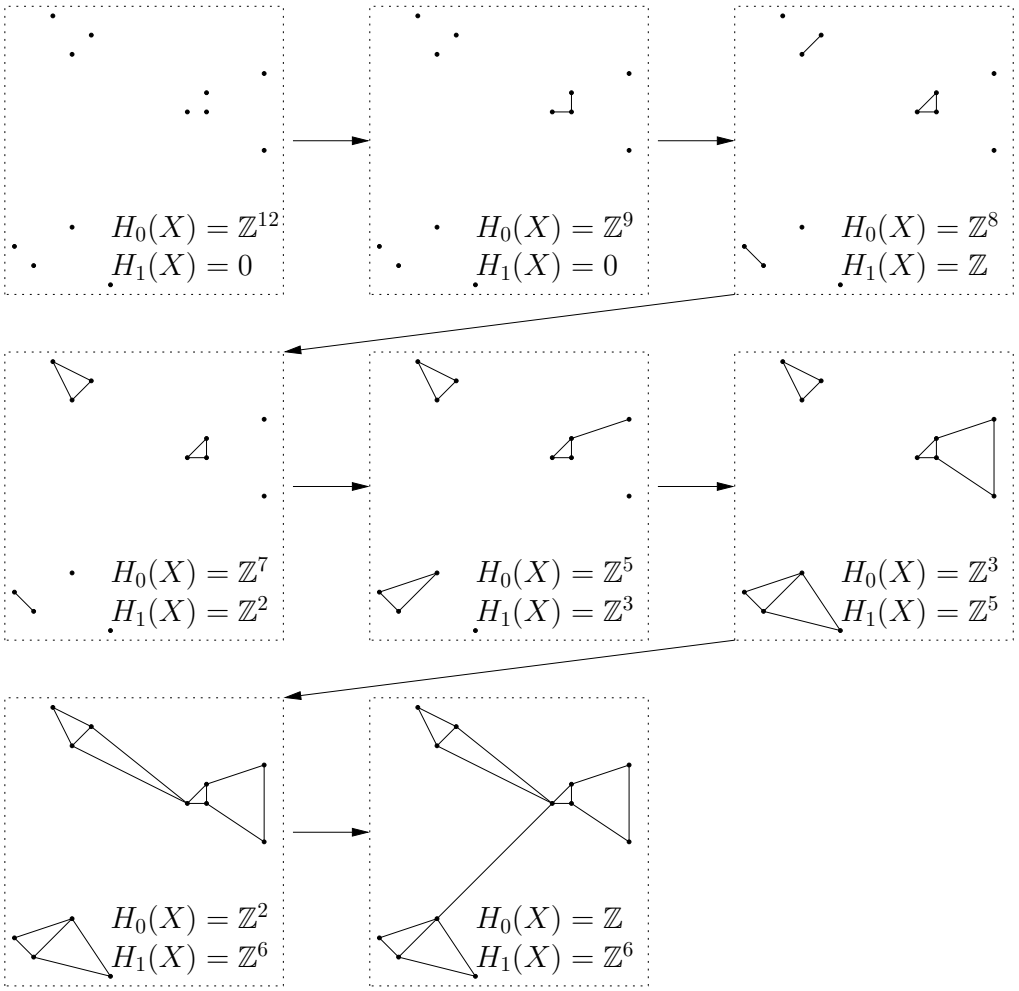


Figure 20 – Illustration of clusterings produced by the single linkage algorithm and their respective associated 0 and 1-dimensional homology groups (Adapted from Carlsson and Mémoli (2010)).

homology classes along a sequence of topological inclusions $\mathcal{F}(X) := (X_0, \tau_{X_0}) \subseteq (X_1, \tau_{X_1}) \subseteq \cdots \subseteq (X_n, \tau_{X_n})$, known as **filtration** of topological spaces. Such inclusions correspond to each hierarchical level of an HC clustering model and, therefore, the persistent homology, in the context of HC, represents how the associated topological space changes along HC levels. For instance, a dendrogram (X, ξ) built up on top of the set $X = \{-1, 0, 1, 2.9, 3, 3.1, 9, 10\}$ such that:

- $\xi(0.1) = \{\{-1\}, \{0\}, \{1\}, \{2.9, 3, 3.1\}, \{9\}, \{10\}\}$;
- $\xi(1) = \{\{-1, 0, 1\}, \{2.9, 3, 3.1\}, \{9, 10\}\}$;
- $\xi(1.9) = \{\{-1, 0, 1, 2.9, 3, 3.1\}, \{9, 10\}\}$;
- $\xi(5.9) = \{\{-1, 0, 1, 2.9, 3, 3.1, 9, 10\}\}$,

devises a filtration $\mathcal{F}(X) := (X, \tau_{\xi(0.1)}) \subset (X, \tau_{\xi(1)}) \subset (X, \tau_{\xi(1.9)}) \subset (X, \tau_{\xi(5.9)})$ in which $(X, \tau_{\xi(r)})$ is equipped with a topology induced from $\xi(r)$ on the usual topology $\mathbb{R}$, such that for

every pair $x_i, x_j \in X$ there is an open set $(x_i, x_j) \in \tau_{\xi(r)}$.

Consider the 0-dimensional homology classes $a_i \in A$ associated with $(X, \tau_{\xi(0.1)})$, such that for all $x \in X$ there is a map $h : X \mapsto A$ in which $x \mapsto a_i$, i.e., $h(-1) = a_1, h(0) = a_2, h(1) = a_3, h(2.9) = a_4, h(3) = a_5, h(3.1) = a_6, h(9) = a_7$, and $h(10) = a_8$. In this sense, the homology classes a_4, a_5 and a_6 were born at $\xi(0)$ and persist until they are merged at $\xi(0.1)$, therefore their persistence is given by 0.1. In addition, the function $f : (X, \tau_{\xi(r)}) \to \mathbb{R}$ mapping the induced topological space onto the corresponding radius is an example of a tame function.

Precisely, tame functions are employed to map, along the filtration, subspace topologies into the real line:

Definition 15 (Tame functions)**.** Let (X, τ_X) be a topological space. A tame function is a continuous map $f : (X, \tau_X) \to \mathbb{R}$ such that $(X, \tau_X)_{\alpha_i} \subseteq (X, \tau_X)_{\alpha_j}$ whenever $\alpha_i < \alpha_j$, where $(X, \tau_X)_\alpha := f^{-1}(-\infty, \alpha]$ is taken with the subspace topology. Moreover, a tame function f must satisfy the following properties:

- The homology groups $H_p((X, \tau_X)_\alpha)$ are of finite rank for every p;
- There are finitely many $\alpha_i \in \mathbb{R}$, such that $H[(X, \tau_X)_{\alpha_i}]$ and $H[(X, \tau_X)_{\alpha_i + \varepsilon}]$ are not isomorphic; in which α_i's are called the **critical values** of f.

Remark 8. Whenever a homology class emerges at $H_p[(X, \tau_{\alpha_j}(X))]$ and vanishes at $H_p[(X, \tau_{\alpha_j}(X))]$, given $i < j$, its **persistence** is defined as $\alpha_j - \alpha_i$.

Considering the aforementioned example, the pre-image of the tame function f is given by $\xi(\cdot)$ such that $f^{-1}(r) = \xi(r)$.

In this sense, a persistent homology group identifies the homology classes that persists along some interval $[\alpha_{i-1}, \alpha_i)$ in the co-domain of the tame function. Formally, the persistent homology groups are defined as:

Definition 16 (Persistent Homology Group)**.** Let (X, τ_X) be a topological space equipped with the filtration that arises from a tame function f, as in Definition 15. Given $\alpha_{i-1} < \alpha_i$, we have the inclusion $f^{i,j} : (X, \tau_X)_{\alpha_i} \subseteq (X, \tau_X)_{\alpha_j}$. The persistent homology group of degree p is the image of the induced homomorphism:

$$\mathbf{f}_p^{i,j} : H_p[(X, \tau_X)_{\alpha_i}] \to H_p[(X, \tau_X)_{\alpha_j}].$$

Given the aforementioned example, let $\alpha_1 = 1$ and $\alpha_2 = 1.9 + \varepsilon$, then the persistent homology group $\mathrm{im}(\mathbf{f}_0^{1,2})$ is $\mathbb{Z}^2$ as only two connected components persist in the interval $[\alpha_1, \alpha_2)$. In addition, the rank of the image of $\mathbf{f}_p^{i,j}$ is called the (i, j)**-persistent Betti number**, i.e., $\beta_p^{i,j} = \mathrm{rank}(\mathrm{im}(\mathbf{f}_p^{i,j}))$, which allows to compute the number of homology classes persisting inside an interval $[i, j)$.

Considering that data insertion produces topological inclusions, if the topological properties of clusters persist along such operations, which must respect the same i.i.d. distribution of the adopted dataset (i.e., if $\mathbf{f}_P^{0,m}$ is an isomorphism given datasets X and $X^m = \{X, x'_1, \ldots, x_m\}$ from which some filtration $(X, \tau_X) \subseteq (X_1, \tau_{X_1}) \subseteq \cdots \subseteq (X_m, \tau_{X_m})$ is defined by a clustering algorithm), the clustering model is stable in terms of p-dimensional homology groups. We claim in this PhD thesis that over-refined topologies do not provide persistent homology classes whenever data is subject to perturbations due to data insertions.

However, the filtration associated with the data insertion does not identify changes in topological spaces obtained along the HC levels, thus only supporting the study of non-hierarchical clustering models, one can take advantage of the **multidimensional persistent homology**, proposed by Carlsson and Zomorodian (2009), to assess the persistence of homology groups from a multiparameter perspective. Let $v, u \in \mathbb{N}$ and a reflexive and transitive relation $\lesssim$ such that $u \lesssim v$ iff $u_i \leq v_i \; \forall i = 1, \ldots, n$, a **multifiltration** of a topological space (X, τ_X) is defined as the topological inclusions $(X_v, \tau_{X_v}) \subseteq (X_u, \tau_{X_u})$ with $v \lesssim u$ and $(X_w, \tau_{X_w}) \subseteq (X, \tau_X)$, for all $w \in \mathbb{N}^n$. In addition, as illustrated in Figure 21, there is a function $F : \mathbb{R}^n \to K$ that maps a n-tuple into a subcomplex K_v of a finite complex K, such that there is a finite set $C = \{v_i \in \mathbb{R}^n\}$ of critical coordinates that form the n-tuples v in which homology classes die or were born.

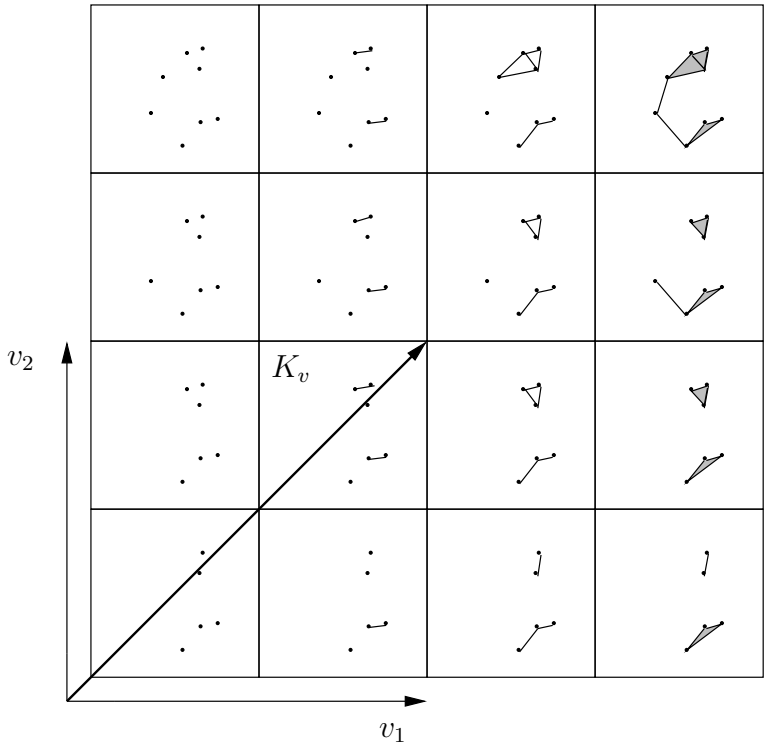


Figure 21 – Vector representation for K_v (Adapted from Carlsson and Zomorodian (2009)).

For instance, consider the study of persistent homology groups in HC. Whenever the adopted radius increases from r_i to r_j, with $i < j$, and new m data points are inserted to build up a

next HC model, there is an inclusion $K_{(r_i,0)} \subseteq K_{(r_j,m)}$ such that the following diagram commutes:

$$\begin{array}{ccc} K_{(r_i,m)} & \xrightarrow{\iota_r} & K_{(r_j,m)} \\ {\scriptstyle \iota_m}\uparrow & & \uparrow{\scriptstyle \iota_m} \\ K_{(r_i,0)} & \xrightarrow{\iota_r} & K_{(r_j,0)}, \end{array}$$

with ι_r and ι_m being homomorphisms. Therefore, it is possible to study the morphism $K_{(r_i,m)} \to K_{(r_f,m)}$ such that if $\iota_m(K_{(r_i,m)})$ and $\iota_m(K_{(r_f,0)})$ are isomorphisms, then the mapping of $\iota_r(K_{(r_i,0)})$ is the same of $\iota_r(K_{(r_i,m)})$, i.e., they map the same domain into the same image. The behavior of such maps are equivalent and, if this equivalence is respected with the insertion of new data, then we can state that a clustering model is consistent. Assuming DC and HC as processes based on i.i.d. sampling, a probability measure can be devised to assess their associated complexes as random variables.

2.4 Measures and Borel sets

Measures are employed to map subsets of elements from X into the positive real line in order to represent their relative order (in terms of the order theory (BURRIS; SANKAPPANAVAR, 1981)), such as length, area, or volume. For instance, suppose a finite countable set $X = \{\varnothing, x_1, x_2, \ldots, x_n\}$, a counting function $\mu : X \to \mathbb{R}^+$ is a measure such that $\mu(X' \subseteq X) = |X|$. Note that one of the requirements of a measure is that there must be an empty set $\varnothing$ belonging to X with measure equals to zero. A measure requires the set X to be equipped with a σ-algebra defined as follows:

Definition 17 (σ-algebra)**.** A σ-algebra on X is a collection Σ of subsets of X such that:

- $X \in \Sigma$ (implying in $\varnothing \in \Sigma$);
- Σ is closed under countable unions;
- Σ is closed under complement.

For instance, take the open set $X = (a,b)$, the collection $\Sigma = \{\varnothing, (a,b), (a,c), [c,b)\}$ forms a σ-algebra for all $a \leq c \leq b$. Therefore, it is possible to build up a measure on X such that $\mu(X' \subseteq X) = d - c$ for all $d, c \in (a,b)$, and $d \geq c$ with a σ-algebra $\Sigma = \{(a,c], (c,d), [d,b) \mid \forall\, a < c < d < b\}$. Considering a **measurable space** (X, Σ), a measure is defined as:

Definition 18 (Measure)**.** A measure is a function $\mu : \Sigma \to \mathbb{R}^+$ which, given $E_i \in \Sigma$, respects the following properties:

- Non-negativity: $\mu(E_i) \geq 0$, for all $E_i \in \Sigma$;

- Null empty set: $\mu(\varnothing) = 0$;
- Countable additivity: $\mu(\bigcup_{i=1}^{n} E_i) = \sum_{i=1}^{n} \mu(E_i)$, with E_i being composed of disjoint subsets of Σ,

from which a **measurable space** is defined as the triple (X, Σ, μ).

Suppose now a topological space (X, τ_X) on the usual topology $\mathbb{R}$ such that $X = (a, b)$, it is possible to take sets $\mathcal{B}(X)$ of open sets or closed sets in (X, τ_X) in order to form a σ-algebra as, for example, $\mathcal{B}(X) = \{\varnothing, (c-\varepsilon, c), (d, d+\varepsilon), \{(c-\varepsilon, c), (d, d+\varepsilon)\}\}$ with $a < c-\varepsilon < c$ and $d < d+\varepsilon < b$. Such open sets, named Borel sets, are defined as:

Definition 19 (Borel sets)**.** A Borel set $\mathcal{B}(X)$ over X is any set containing open (or closed) sets U_i of a topological space (X, τ_X) such that they are:

- Closed under countable union: Given a countable set of indices I, $\bigcup_{i \in I} U_i \in \mathcal{B}(X)$;
- Closed under countable intersections: Given a countable set of indices I, $\bigcap_{i \in I} U_i \in \mathcal{B}(X)$;
- Closed under relative complement: $U_i - U_j \in \mathcal{B}(X)$, for all $U_i, U_j \in \mathcal{B}(X)$.

The collection of all Borel sets forms a **Borel algebra** which, therefore, allows the definition of a measure. Then, this PhD thesis considers that there is an unknown Hausdorff compact topological space (Z, τ_Z) from which data points are sampled. As (Z, τ_Z) is Hausdorff compact, then it endows a Borel algebra from which it is possible to define a measure, thus allowing the probabilistic analysis of how topologies change along the data sampling.

In this context, a probability space is a measurable space $(\Omega, \mathcal{E}, P)$ such that Ω is the set of all possible elements to be acquired, $\mathcal{E}$ is the collection of possible events, i.e., a sequence of elements acquired by some probability distribution, and $P : \mathcal{E} \to [0, 1]$ corresponds to the probability function that measures the frequency that events occur.

In this sense, P must assure the properties of a measure and its image must lie in the interval $[0, 1]$ such that $P(\Omega) = 1$. So P measures how frequent is some event $e \in \mathcal{E}$ inside the universe set Ω. For instance, in the former example, the probability that an element belongs to some interval (c, d) is given by $(d-c)/(b-a)$. Moreover, a probability measure P allows the definition of the expected value of the random variables $x \in X \subset \Omega$, thus leading to a "center of mass" for the probability distribution as follows:

$$\mathbb{E}_X[x] = \int_{x \in X} x dP(x),$$

such that the random variables diverge from the expected value according with its variance $\mathbb{E}_X[(x - \mathbb{E}_X[x])^2]$.

Furthermore, concentration inequalities formalize the probability of random variables diverging from their expected values according to some factor ε, i.e., $P(|x - \mathbb{E}_X[x]| > \varepsilon)$. Considering a measure μ_X defined on top of a topological space (X, τ_X), corresponding to topological features (e.g. number of connected components), this PhD thesis employs those inequalities in order to study $P(|\mu_X(x) - \mathbb{E}_{\mu_X}[\mu_X(x)]| > \varepsilon)$.

2.5 Concentration inequalities and stability

The main goal of concentration inequalities is the definition of probabilistic bounds for the divergence between a random variable (or some measure of it) and its expected value. Such divergence can be used to assess an estimator, e.g., the generalization measure defined in the Statistical Learning Theory (VAPNIK, 1995; LUXBURG; SCHÖLKOPF, 2011; MELLO; PONTI, 2018) as $|R_{\text{emp}}(f) - R(f)|$, with $R_{\text{emp}}, R : \mathcal{F} \to \mathbb{R}^+$ being, respectively, the empirical and its expected risk, $\mathcal{F}$ corresponds to the space of admissible functions, and $f \in \mathcal{F}$ is some classifier or regression function (LUXBURG; SCHÖLKOPF, 2011; MELLO; PONTI, 2018). In that supervised context, Vapnik (1995) employed Chernoff's and Hoeffding's bounds (CHERNOFF, 1952; HOEFFDING, 1963) in order to prove the Empirical Risk Minimization, which is a direct result of evaluating how the estimator $R_{\text{emp}}(f)$ approaches $R(f)$ as the sample size $n \to \infty$.

Although such proof requires the assumption of labeled data examples, concentration inequalities and algorithmic stability do not require such restriction, thus permitting us to study the unsupervised machine learning scenarios, as performed in this thesis. In this sense, this chapter covers the concentration inequalities proposed by Markov (GHOSH, 2002), Chernoff (1952), Hoeffding (1963) and, finally, Azuma (1967). Inequalities such as Azuma's assume restrictions which, when held, naturally impose consistency whenever stability is assured. Therefore, taking a measure $\mu_X : (X, \tau_X) \to \mathbb{R}^+$ over some topological space associated with a clustering model, given the proper restrictions are respected and the stability is demonstrated, then the topological features analyzed in terms of μ_X are consistent with respect to the clustering model.

In order to demonstrate Azuma's inequality, we must introduce Markov's first, which is defined as:

Definition 20 (Markov's inequality)**.** Given a non-negative random variable $x \in \Omega$ sampled from some probability distribution function $P : \Omega \to [0, 1]$ and some scalar $\varepsilon \in \mathbb{R}^+$:

$$P(x > \varepsilon) \leq \frac{\mathbb{E}[x]}{\varepsilon}.$$

Such inequality provides a loose bound that proportionally decays along ε, as illustrated in Figure 22. Therefore, taking x as $|x' - \mathbb{E}[x']|$, the following inequality holds:

$$P(|x' - \mathbb{E}[x']| > \varepsilon) \leq \frac{\sigma^2}{\varepsilon^2}, \tag{2.3}$$

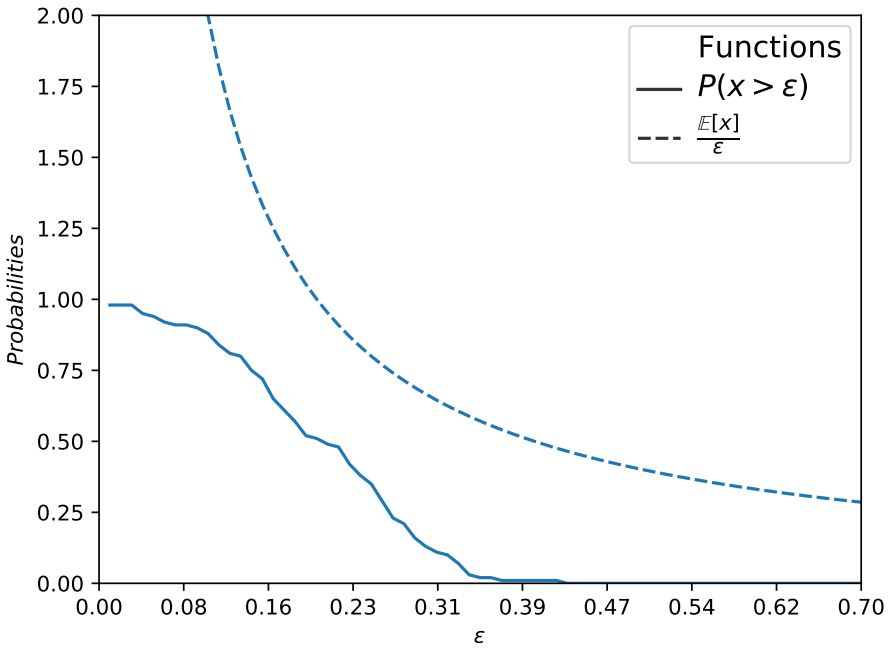


Figure 22 – The proportional decay of $\frac{\mathbb{E}[x]}{\varepsilon}$ in comparison with $P(x > \varepsilon)$ given 100 samples acquired from a normal distribution $\mathcal{N}(0.2, 0.1^2)$. Note that $\frac{\mathbb{E}[x]}{\varepsilon} \geq 1$ for the initial values of ε, in this sense, for such values, $x > \varepsilon$ always holds (Adapted from Mello *et al.* (2019)).

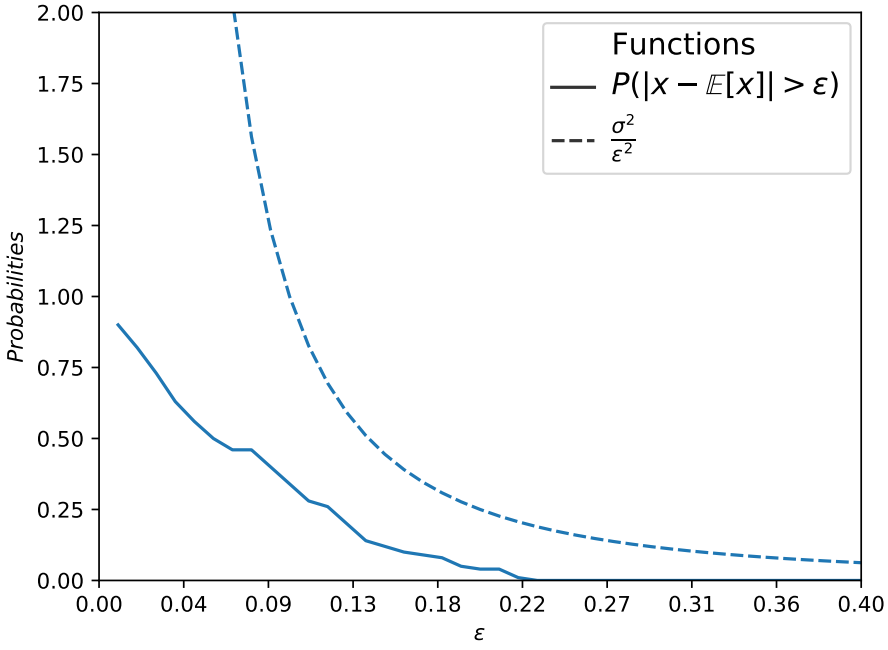


Figure 23 – Chart comparing $\frac{\sigma^2}{\varepsilon^2}$ with $P(|x - \mathbb{E}[x]| > \varepsilon)$, it was produced taking 100 samples acquired from a normal distribution $\mathcal{N}(0.2, 0.1^2)$.

with σ^2 being the variance of the random variables x.

The Inequality 2.3 is known as **Chebyshev's inequality** (FERENTINOS, 1982) and it implies that ε must be sufficiently greater than the variance of x, as shown in Figure 23, in order to guarantee consistency. Restricting such random variables in some interval $[a,b]$, Chernoff proposes a tighter bound than Markov's inequality when analyzing consistency (KOTZ; IBRAGIMOV; HAS'MINSKII, 2013), being defined as:

Definition 21 (Chernoff's inequality)**.** Given an i.i.d. random variable $x \in [a,b]$ sampled from some probability distribution function $P : [a,b] \to [0,1]$, Chernoff's inequality is:

$$P(x - \mathbb{E}[x] > \varepsilon) \leq e^{\frac{-2\varepsilon^2}{(b-a)^2}}.$$

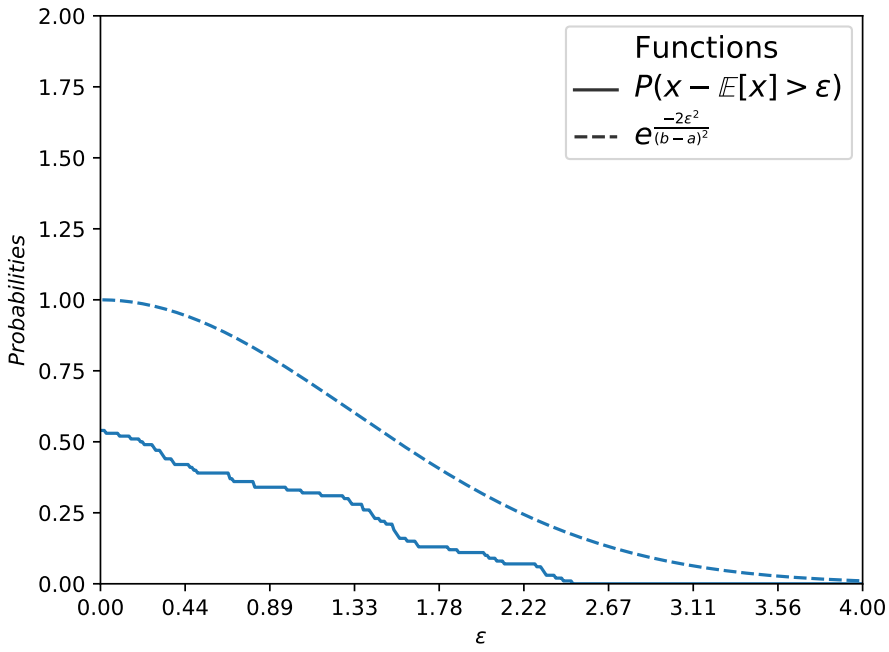


Figure 24 – Chart comparing $e^{\frac{-2\varepsilon^2}{(b-a)^2}}$ with $P(x - \mathbb{E}[x] > \varepsilon)$, it was produced taking 100 samples acquired from a uniform distribution $\mathcal{U}(4,9)$, i.e., with $a = 4$ and $b = 9$ (Adapted from Mello *et al.* (2019)).

In this sense, Chernoff (1952) presents an inequality that exponentially decays, contraposing the fractional decay of Markov's inequality as illustrated in Figure 24. Hoeffding (1963) further generalizes Chernoff's inequality by considering the sum S_n of n random variables x_i such that:

Definition 22 (Hoeffding's inequality)**.** Given $S_n = x_1 + x_2 + \cdots + x_n$ such that $x_i \in [a_i, b_i]$ with $1 \leq i \leq n$ is independent and identically sampled from some probability distribution function $P : \Omega \to [0,1]$:

$$P(S - \mathbb{E}[S] > \varepsilon) \leq e^{\frac{-2\varepsilon^2}{\Sigma_{i=1}^{n}(b_i - a_i)^2}}.$$

Hoeffding's Inequality allows the study of consistency over averages whenever the size of some dataset grows.

Azuma's Inequality considers (HOEFFDING, 1963) to study a specific class of stochastic processes referred to as martingales:

Definition 23 (Martingale)**.** A martingale is a stochastic process endowed on a sequence of random variables $x_1, x_2, \ldots, x_n$, in form:

$$\mathbb{E}[x] < \infty, \text{ and}$$
$$\mathbb{E}[x_n | x_1, x_2, \ldots, x_{n-1}] = x_{n-1}.$$

It is also possible to study a martingale with respect to another sequence with $y = f(x)$ such that $\mathbb{E}[y] < \infty$ and $\mathbb{E}[y_n | x_1, x_2, \ldots, x_{n-1}]$. In addition, a martingale difference sequence is defined as:

Definition 24 (Martingale difference sequence)**.** A martingale difference sequence is a stochastic process endowed on a sequence of random variables $x_1, x_2, \ldots, x_n$, given $V_i = X_i - X_{i-1}$, in the form:

$$\mathbb{E}[x] < \infty, \text{ and}$$
$$\mathbb{E}[V_n | x_1, x_2, \ldots, x_n] = 0.$$

By employing such differences, Azuma's inequality states that whenever $|x_i - x_{i-1}|$ is upper bounded by some constant c_i, the probability of x_1 diverging from x_n, by a factor of ε, converges to zero in form:

Definition 25 (Azuma's inequality)**.** Given a martingale sequence $x_1, x_2, \ldots, x_n$ sampled from some probability distribution function P, if $|x_i - x_{i-1}| < c_i$:

$$P(|x_n - x_1| > \varepsilon) \leq 2e^{\frac{-\varepsilon^2}{\Sigma_{i=1}^{n} c_i^2}}.$$

From that, suppose some measure $\mu_X : (X, \tau_X) \to \mathbb{R}^+$ such that (X, τ_X) corresponds to a clustering model. In an ideal scenario, when learning is guaranteed, it is expected that whenever such model is reconstructed from more data, the measure of μ_X should not change and, therefore, the filtration given by such data insertions must form a martingale with respect to μ_X. Therefore, whenever the measure on clustering models forms a martingale, learning is guaranteed if $|\mu_X(x_i | x_1, \ldots, x_n) - \mu_X(x_{i-1} | x_1, \ldots, x_n)| < c_i$, i.e, if μ_X is stable with respect to data insertions.

2.6 Final considerations

In the context of Machine Learning, spaces are of upmost importance in the development of ML theoretical frameworks as they formalize datasets from an axiomatic structure, thus supporting the development of more sophisticated properties and conclusions. Defined from

a particular space type (e.g., vectorial, metric or topological), such properties are inherited throughout the space hierarchy and, in this sense, topological attributes can be extrapolated along more restricted spaces (see Figure 7).

A theoretical framework can take advantage of such formalization while studying the invariances of ML models, as for instance: (i) by analyzing changes in input vector directions with respect to hyperplanes; (ii) by verifying if metric or even topological spaces obtained from clustering algorithms are somehow invariant to data bounded perturbations (e.g. isometry, homeomorphism or isomorphism of homology groups). Such spatial invariance leads to a novel notion of generalization which, instead of only relying on error functions, endows more qualitative information from clustering models.

CHAPTER

3

RELATED WORK

3.1 Initial considerations

Machine learning is divided into two main paradigms: Supervised Machine Learning (SML) and Unsupervised Machine Learning (UML). The first relies on the Statistical Learning Theory (SLT), which defines properties to ensure the consistency of the Empirical Risk Minimization (ERM) principle, thus leading to model generalization. UML is not supported by SLT as Vapnik's theory assumes labeled data examples, thus requiring a new theoretical approach to characterize generalization and devise learning bounds [1].

From this point of view, this chapter discusses the recent reports associated with the theoretical formalization of the Data Clustering (DC) problem as follows: (i) Section 3.2 introduces the axiomatic characterization of data clustering by Kleinberg (2002); (ii) Section 3.3 approaches the collaborative report by Ackerman, Ben-David and Loker (ACKERMAN; BEN-DAVID; LOKER, 2010), which extends Kleinberg's by defining a taxonomy for clustering properties. This same section includes a study on clustering quality measures (BEN-DAVID; ACKERMAN, 2009); (iii) Section 3.4 details Carlsson and Mémoli (2010) formalization on consistency and stability for hierarchical clusterings; and, finally, (iv) Section 3.5 features essays on persistent homology and its applications.

3.2 Kleinberg's clustering formalization

After criticizing the vagueness on clustering foundations, Kleinberg (2002) defines suitable partitions in terms of the relative similarity among points, given elements belonging to the same group should be more similar than points from other groups, what consists in the main

[1] Note that Ben-Hur *et al.* (2002) developed a clustering algorithm, named as Support Vector Clustering (SVC), which is based on Support Vector Machine (CORTES; VAPNIK, 1995) formulation, however, no clear relation between SVC and Vapnik's SLT is presented.

motivation for his axiomatic framework built up on top of three main assumptions: there is (i) a fixed dataset X with at least two data points; (ii) a distance function $d : X \times X \to \mathbb{R}^+$ that not necessarily satisfies the triangle inequality (hence d does not necessarily form a metric space); and (iii) a clustering function $f : \mathcal{D} \to \Gamma$ in which $\mathcal{D}$ represents the set of admissible distance functions and Γ the produced partitions.

Kleinberg disregards the ambient space of X such that the definition of any underlying space is impossible to be made. Therefore, his formulation relies only on the dataset X without any additional assumption, given the distance function $d \in \mathcal{D}$ is defined upon a countable set I, which is formed by the indices of the elements in X such that $d : I \times I \to \mathbb{R}^+$, $i, j \in I$ and $(i, j) \mapsto d(i, j)$. In this sense, only the distance matrix $\mathcal{D}$ is assumed for the application of the clustering function f. From that, Kleinberg defines the properties of scale-invariance, consistency (different from the concept of statistical consistency), and richness as mandatory to characterize and analyze clustering models:

Definition 26 (Scale-invariance)**.** Given a scalar α, a cluster Γ is scale-invariant if and only if $f(\mathcal{D}) = f(\alpha\mathcal{D})$, i.e., scalar multiplication over the matrix $\mathcal{D}$ does not change the resulting partitions;

Definition 27 (Consistency)**.** Let a Γ-transformation be defined such that, if d' is a Γ-transformation of d, then $d'(i, j) \leq d(i, j)$ for the elements i, j contained in the same cluster and $d'(i, j) \geq d(i, j)$ for the elements i, j belonging to different clusters. Then, given d and d' two distance functions, if d' is a Γ-transformation of d then Γ is consistent if and only if $f(d') = f(d)$, i.e., reducing the distance inside clusters while increasing the distance between them must preserve the partitions;

Definition 28 (Richness)**.** Let Range(f) be the set of all partitions $f(d) = \Gamma$ for some distance function d. Then, Γ is rich if and only if Range(f) is equal to the set of all partitions in X, i.e., "every partition of X is a possible output" (KLEINBERG, 2002).

Considering such properties, Kleinberg then proves that, in light of Arrow's theorem (ARROW, 1951), they are impossible to be simultaneously satisfied, thus implying that one of them must be relaxed somehow.

This PhD thesis presents an argumentation to justify the disregarding or relaxation of the richness axiom when the dataset X is considered to be acquired from some probability distribution, i.e., when X is not fixed. Briefly, there are topological spaces refined enough along the hierarchical clustering representation that produce inconsistent measure functions when new data are included. This topic is properly addressed in Chapter 4.

3.3 On clustering quality measures and additional properties

Kleinberg's axiomatic framework motivated Ben-David and Ackerman (BEN-DAVID; ACKERMAN, 2009) to propose clustering quality measures as they claim the insatisfatibility of axioms is an artifact of Kleinberg's formulation and not an inherent property of clustering. The authors define such Clustering Quality Measure (CQM) as a function μ mapping the triple (X,Γ,d), formed by a dataset, a partition and a distance function, respectively, into an ordered set, reflecting how adequate such model is.

Ben-David and Ackerman (2009) then redefine Kleinberg's axioms in terms of this CQM as follows:

Definition 29 (Scale-invariance)**.** A quality measure μ satisfies the scale-invariance if for every partition Γ built up on top of (X,d), and every positive α, the following holds $\mu(X,\Gamma,\alpha d) = \mu(X,\Gamma,d)$;

Definition 30 (Consistency)**.** A quality measure μ is consistent if for every partition Γ built up on top of (X,d) whenever d' is a Γ-transformation of d the following holds $\mu(X,\Gamma,d') \geq \mu(X,\Gamma,d)$;

Definition 31 (Richness)**.** A quality measure μ is rich if for each non-trivial partition Γ of X there is a distance function d over X such that $\Gamma = \arg\max_{\Gamma_i} \mu(\Gamma_i,X,d)$.

In this sense, the authors define a measurable function $\mu(X,\Gamma,d)$ on the partitions, but: (i) there is no information if (X,Γ,d) is measurable nor the necessary conditions were defined to allow such measurability; and (ii) Ackerman and Ben-David's axiom of richness imposes that there is a distance function d such that $\Gamma = \arg\max_{\Gamma_i} \mu(\Gamma_i,X,d)$, therefore, as there is always a Γ-transformation shrinking all partitions Γ it will always exist a distance function d' such that $\max_{\Gamma_i}\{\mu(\Gamma_i,X,d)\} = \infty$ so this new axiom of richness always holds.

Later on, Ackerman, Ben-David and Loker (ACKERMAN; BEN-DAVID; LOKER, 2010) propose a taxonomy for data clustering properties to support users while studying and selecting the most appropriate clustering methods, to mention: isomorphism invariance, scale-invariance, consistency, locality, inner and outer consistencies, k-richness, outer-richness and threshold richness. In the context of this PhD thesis, the property of locality is defined as:

Definition 32. A clustering function f is local if the behavior on a union of subsets of clusters only depends on distances among elements of that union (ACKERMAN; BEN-DAVID; LOKER, 2010). Thus, for $C' \subseteq C$:

$$f(\bigcup C', d) = C'.$$

Then, this thesis takes into account the isomorphism invariance with respect to topological properties devised from homology groups associated with clustering models, as it will naturally

imply scale-invariance, consistency, and locality. Therefore, richness must be relaxed so that a clustering algorithm provides such isomorphism invariance.

By relaxing richness, **this PhD thesis claims that a single property is enough to guarantee adequate clustering: the corresponding topological space has to be invariant when the related dataset is subject to new data acquisition**. In such context, Carlsson and Mémoli study the metric invariance for hierarchical clustering algorithms along data perturbations, then proving that a modification on the single linkage satisfies such condition (CARLSSON; MÉMOLI, 2010).

3.4 Carlsson and Mémoli's consistency for hierarchical clustering

In attempt to complement Kleinberg's axioms, Carlsson and Mémoli study and develop a theoretical framework to support the agglomerative hierarchical clustering, which was then characterized according to three conditions that were used to prove the uniqueness for a modified version of the single linkage algorithm. This algorithm version is stable and statistically consistent with respect to the Gromov-Hausdorff distance.

Their formulation considers that hierarchical clustering models are built on top of a finite metric space (X,d) by producing nested partitions represented in terms of dendrograms, which are equivalent to ultrametric spaces, as also shown. In order to proceed with their study, they reformulated permutation-invariant linkage functions using the average, complete and single criteria, by allowing the agglomeration of more then just two elements at each iteration. Then, a hierarchical clustering method is defined as a map $\mathfrak{L} : \mathbb{X} \to \mathbb{U}$ such that $(X,d) \in \mathbb{X} \mapsto (X,u) \in \mathbb{U}$, with (X,u) being an ultrametric space associated with the corresponding dendrogram. In this sense, Carlsson and Mémoli define three properties to characterize the hierarchical clustering (CARLSSON; MÉMOLI, 2010):

Definition 33 (Carlsson and Mémoli's properties for hierarchical clustering)**.** Let $\mathfrak{L}$ be a hierarchical clustering method:

- $\mathfrak{L}\left(\{p,q\}, \left(\begin{smallmatrix} 0 & \delta \\ \delta & 0 \end{smallmatrix}\right)\right) = \left(\{p,q\}, \left(\begin{smallmatrix} 0 & \delta \\ \delta & 0 \end{smallmatrix}\right)\right)$ for all $\delta > 0$, i.e., a two point metric space must be mapped into a dendrogram in which they are merged at radius δ, i.e., there exists a trivial clustering;

- Whenever $X, Y \in \mathbb{X}$ and $\phi : X \to Y$ are such that $d_X(x,x') \geq d_Y(\phi(x),\phi(x'))$ for all $x,x' \in X$, then $u_X(x,x') \geq u_Y(\phi(x),\phi(x'))$ also holds for all $x,x' \in X$ where $\mathfrak{L}(X,d_X) = (X,u_X)$ and $\mathfrak{L}(Y,d_Y) = (Y,u_Y)$, i.e., a shrinkage in metric spaces must proportionally impact the ultrametric, such that it preserves the corresponding dendrograms. For instance, if a metric

space with two clusters shrinks, and merges them, this must also occur in the corresponding dendrogram;

- Given $x \neq x' \in X$, where $\mathfrak{L}(X,d) = (X,u)$ with $(X,d) \in \mathbb{X}$ and $\text{sep}(x,x') := \min_{x \neq x'} d(x,x')$, then $u(x,x') \geq \text{sep}(X,d)$, i.e., there is a t such that the produced dendrogram represents singletons when $0 \leq t < \text{sep}(X,d)$, therefore exists an initial interval in which no points are merged.

Carlsson and Mémoli prove that such properties are satisfied only if the considered algorithm is the single linkage. Considering such an algorithm, they prove its stability in terms of the Hausdorff and Gromov-Hausdorff distances, defined as:

Definition 34 (Hausdorff distance)**.** Given two sets X, Y and a distance function d:

$$d_H(X,Y) := \max\{\sup_{x \in X} \inf_{y \in Y} d(x,y), \sup_{y \in Y} \inf_{x \in X} d(x,y)\},$$

and

Definition 35 (Gromov-Hausdorff distance)**.** Given two metric spaces $(X,d_X), (Y,d_Y)$ and two isometric embedding f,g, such that $f : (X,d_X) \to (M,d_M)$ and $g : (Y,d_Y) \to (M,d_M)$:

$$d_{GH}(X,Y) := \inf_{f,g} d_H(f(X,d_X), g(Y,d_Y)).$$

They respectively measure: (i) the maximal distance from a set to another; and (ii) the "effort" to transform two metric spaces into an isometric one. Therefore, given a compact metric space (Z,d_Z), two hierarchical clustering models are finite stable if $d_{GH}((X,u_X),(X',u_{X'})) \leq d_H^Z(X,Z) + d_H^Z(X',Z)$. In this sense, Carlsson and Mémoli (2010) prove that there is a compact metric space that bounds the hierarchical clustering for the single linkage algorithm, even when data is subject to perturbations given by the underlying probability distribution.

In order to formulate the probabilistic convergence for the single linkage, Carlsson and Mémoli consider a hierarchical clustering model on metric measure spaces. In this sense, the measurable space (X,d_X,μ_X), endowed with a metric space (X,d_X) and a Borel probability measure μ_X on X with compact support $\text{supp}(\mu_X)$, is defined as an mm-space (measure metric space) (GROMOV; LAFONTAINE; PANSU, 1999). The authors also define a function $f_X : \mathbb{R}^+ \to \mathbb{R}^+$ as $r \mapsto \min_{x \in \text{supp}(X)} \mu_X(B_X(x,r))$, which is non-decreasing and $f(X) > 0$ for all $r > 0$. Let $F_X : \mathbb{N} \times \mathbb{R}^+ \to \mathbb{R}^+$ be a function defined by $(n,\delta) \mapsto \frac{e^{-mf_X(\delta/4)}}{f_X(\delta/4)}$, Carlsson and Mémoli assume δ_0 to be fixed and, therefore, $F_X(\cdot,\delta_0)$ to be bounded and decreasing, in order to prove the following theorem:

Theorem 1. Let (Z,d_Z,μ_Z) be an mm-space, $\text{supp}(\mu_Z) = \bigcup_{\alpha \in A} U^{(\alpha)}$ for a finite index set A, and $\mathbf{U} = \{U^{(\alpha)}\}$ be a collection of disjoint, compact, path-connected subsets of Z. Let (A,d_A) be the metric space arising from $\mathbf{U}$ and $\delta_A := \text{sep}(A,d_A)/2$. Let $X = \{x_1,x_2,\ldots,x_n\}$ be a collection

of n independent random variables with distribution μ_Z and d_X be the restriction of d_Z to $X \times X$. Then, for $\zeta \geq 0$ and $n \in \mathbb{N}$:

$$P_{\mu_Z}\left(d_{GH}\left(\mathfrak{L}(X,d_X),\mathfrak{L}(A,d_A)\right) > \zeta\right) \leq F_Z(n,\min(\zeta,\delta_A/2)), \tag{3.1}$$

with $\mathfrak{L}$ being the ultrametric space produced by Carlsson and Mémoli (2010)'s modified single linkage algorithm.

Complementarily, this PhD thesis not only considers topological spaces for the proper formulation of the statistical consistency but also proves, in Chapter 4.3.1, upper and lower bounds for Carlsson and Mémoli's metric consistency based on the number of connected components produced by an hierarchical clustering algorithm.

3.5 Persistent homology

Carlsson (2009) points out four strong motivations to the employment of topological analysis on data: (i) the production of qualitative results; (ii) the absence of theoretical justification in adopting the metric space; (iii) they are not restricted to coordinate systems; and (iv) their hierarchical scheme provides complementary information for data clusterings. Although the filtration performed along the hierarchical clustering represents multiple categorical aspects of some dataset, hence producing richer information, there are over-refined topological spaces associated with such filtration which do not produce stable measures.

Persistent homology characterizes the variations of homology groups along such filtrations, thus allowing the study of the instability in hierarchical clustering models. Therefore, if such changes do not indefinitely diverge when data is subject to perturbations, i.e., it is stable, then this PhD thesis considers that such a hierarchical clustering model is capable of generalizing data, thus leading to learning. Consequently, the clusters belonging to a hierarchical model can be chosen whenever they are stable/consistent.

Although Cohen-Steiner, Edelsbrunner and Harer (2007) and Chazal *et al.* (2009) report a proof that persistence diagrams are stable in terms of the Bottleneck distance (EDELSBRUNNER; HARER, 2008), their consideration of a fixed topological space diverges from the problem formulated in this thesis as its main goal is the study of perturbations in topological spaces along the sampling of new data, thus devising stability and consistency results.

Several applications can take advantage of persistent homology and our consistency formulation by considering, along the filtration, only the topological spaces associated with stable homology groups, leading to consistent measures. To mention, persistent homology has gained prominence in other scientific areas, such as in medical and material sciences, given its capability of summarizing the geometrical forms present in complex datasets. Lawson *et al.* (2019), for example, employ it to improve a prognostic predictor for prostate cancer. Such

predictor, known as Gleason Score (FURIHATA; TAKEUCHI, 2011), takes into account the cytomorphological features of some biopsy from which cellular shapes devise five categories known as architectural types. Such geometrical characteristic of cells motivated the adoption of persistent homology to produce a richer set of architectural types, allowing a more detailed prognostic system (LAWSON *et al.*, 2019).

On a complementary point of view,Takiyama *et al.* (2017) apply persistent homology over invasive ductal carcinoma (the most common type of breast cancer) information to improve its prediction and prognostic. Data is collected by using the Imunnohistochemestry process, which marks the cells of a tissue whenever antibodies bind to antigens. Its evaluation is typically made from visual inspection, being inherently subjective and thus motivating the authors to model such data by using persistent homology, leading to the development of the Persistent Homology Index (PHI) which devises a similar result to the one produced by some pathologist's visual analysis.

Finally, Kimura *et al.* (2018) employ persistent homology to find trigger sites in a material that are specific regions of heterogeneity from which reactions such as fractures, corrosion, and degradation may initiate. A persistent diagram is used to describe the topological features of iron sinter (raw material of iron fabrication) X-ray-based measurements, in order to correlate those features with rupture formation.

3.6 Final considerations

Data clustering theoretical frameworks, such as the ones presented in Kleinberg (2002), Ackerman, Ben-David and Loker (2010), list a series of properties that support the study of clustering algorithms with respect to some mild assumptions. However, some of those properties, as described in Kleinberg (2002), are impossible to be satisfied when simultaneously taken into account. In this sense, this PhD thesis strongly considers Carlsson and Mémoli's stability and consistency developments and claims that a single property is enough to support a proper clustering model: the guarantee of the statistical consistency.

This work counts on topological spaces and persistent homology instead of metric and ultrametric spaces, as there is no theoretical guarantee that a clustering input space is essentially metric, and the evaluation of consistency over topological features can guarantee invariance of additional geometric structures regarding those of metric nature. In addition, persistent homology has gained special attention in other scientific areas given its capability to analyze complex datasets, hence the theoretical framework proposed in this PhD thesis can also support such studies.

CHAPTER

4

THE COARSE-REFINEMENT DILEMMA

4.1 Initial considerations

The Data Clustering (DC) problem is defined on top of multiple mathematical structures such as partitions and metric spaces, which are the most employed to characterize it. There is no evidence that data clustering is restricted to such structures, what motivates the study of topological features to complement the formulation of the DC problem. From this perspective, this chapter is organized as follows: (i) Section 4.2 introduces the DC problem and a brief demonstration of the proposed Coarse-Refinement Dilemma (CRD) for the homology group H_0; (ii) Section 4.3 presents the generalization for CRD in Hierarchical Clustering (HC) models and the resulted consistency and stability formulations, comprising the main contributions of this thesis; and, finally, (iii) Section 4.3.1 associates such consistency and stability results to the metric consistency by Carlsson and Mémoli (2010).

4.2 Introduction of the Coarse-Refinement Dilemma

Data clustering can be roughly defined as the process of organizing data into groups having some sort of "meaning" even in the absence of class labels, in spite of the difficulty in characterizing instances without any previous information. In this sense, such "meaning" is typically related with proximity, or similarity, among data points. Equivalence relations are suitable to formalize such proximity as it is endowed with just three reasonable properties: identity, symmetry and transitivity. Intuitively, such relations assure that: (i) a data point is similar to itself; (ii) if a data point is similar to another, the latter is also similar to the former; and, finally, (iii) if two data points are similar to third one, then they are also similar with each other.

In this sense, we specifically define data clustering as the process of organizing data into groups such that they are contiguous with respect to an equivalence relation. Then, the

first question that arises is if an equivalence relation guarantees contiguity when data points, or its underlying space, are subject to mild transformations. For instance, Kleinberg (2002) states that clustering partitions, which in our scenario are the structures defining the equivalence relations, must be invariant when the space is subject to scalar and Γ-transformations. On the other hand, Carlsson and Mémoli (2010) demonstrate the metric stability and consistency for a variation of the single linkage algorithm, which become invariant with respect to its associated ultrametric spaces when data is subject to bounded perturbations and, also, with respect to its corresponding equivalence relation.

In order to define the equivalence relations for data clustering, two issues must be considered: (i) the underlying structure on top of which such relations are constructed, e.g., partitions, metric spaces, topological spaces, among others; and (ii) the level of equivalence, e.g., the radius of open balls when metric spaces are taken into account to formulate the DC problem. Such levels of equivalence are defined as:

Definition 36 (Levels of equivalence)**.** Given a set X and an equivalence relation $\sim_\alpha$ associated with an element α of an ordered set A, α_i is a level of equivalence if there is a sequence:

$$X/\sim_{\alpha_0} \subseteq X/\sim_{\alpha_1} \subseteq \cdots \subseteq X/\sim_{\alpha_i} \subseteq \cdots \subseteq X/\sim_{\alpha_\infty},$$

being α_0 the equality equivalence relation ($x \sim x$ and $x \not\sim x'$ for all $x, x' \in X$), and α_∞ the trivial equivalence relation ($x \sim x'$ for all $x, x' \in X$).

For instance, partitional algorithms such as k-means and spectral clustering typically present an inherent mechanism to optimize the level of equivalence with respect to an objective function, e.g., in k-means algorithms, such level is given by the affine spaces determined by the centroids and, in case of spectral clustering, it is produced by the eigenspace maximizing the similarities among data points in average. Such spaces can induce topological spaces thus fitting in our framework, they will not present invariance in terms of homology groups though as the maximum number of connected components will be always fixed and no homology group of degree greater than one can be represented.

Conversely, density and hierarchical-based clusterings depend on the choice of one or multiple levels of equivalence, e.g., those associated with radii of open balls, which can negatively impact the suitability of a clustering model, as shown: (i) the first issue is that over-refinement and over-coarsening produces tautological relations, respectively, a point is only equivalent to itself (equality) or a point belongs to the universe set (triviality); and (ii) the second, and more specific issue, is whether continuously refined equivalence relations, i.e., relations which continuously "approximate" equality, are adequate and thus properly represent data.

Considering a probabilistic perspective, whenever over-refined relations are assumed, then no inference, except a guess, can be made as it will only rely on one element. For instance, in the supervised scenario, in light of the Statistical Learning Theory (SLT), the memory function

producing the most overfitted classification rises from such equality relation as it classifies x as y_i if and only if $x = x_i$. On the other hand, considering the DC problem, equality relations will produce clusters unable to represent additional data samples, thus constantly requiring the definition of new groups. In such context, the topological framework supports the notion of continuous approximation between refined and equality relations without imposing any metric restriction and with such relations being determined by the neighborhoods of a topological space, as illustrated in Figure 25.

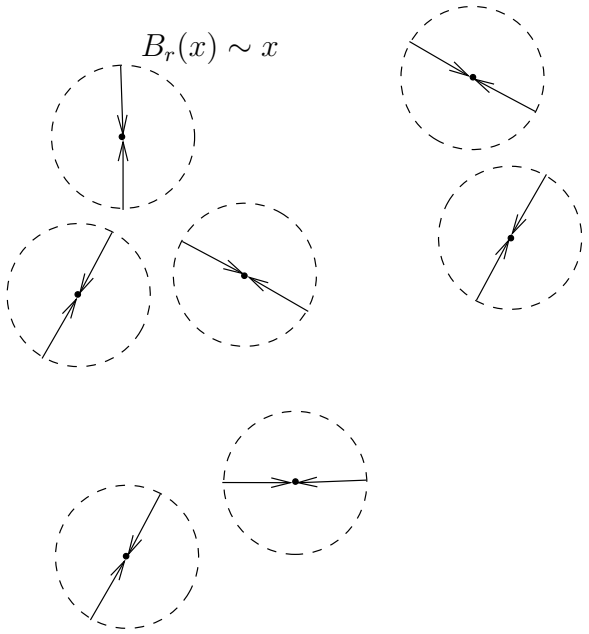


Figure 25 – Illustration of the continuous approximation from $B_r(x)$ to x itself, resembling the production of a memory function defined in Statistical Learning Theory (LUXBURG; SCHÖLKOPF, 2011).

For instance, suppose $X = \{-0.1, 0, 0.1, 0.9, 1, 1.1\}$ and consider (X, τ_X) as the topological space induced by closed balls of radius 0.1 with a distance function $d := |x - x'|$ for $x, x' \in X$, from which the following is derived $\tau_X = \{\{-0.1, 0, 0.1\}, \{0.9, 1, 1.1\}\}$. If such radius is reduced by an infinitesimal value ε, then $\tau_X = \{\{-0.1\}, \{1\}, \{0.1\}, \{0.9\}, \{1\}, \{1.1\}\}$ is obtained, thus resulting in an equality relation defined by the contiguity of the sets in the topology.

However, if points are acquired such that $x_1 = [-0.1, -0.1 + \varepsilon)$ or $x_1 = [0 - \varepsilon, 0)$, $x_2 = (0, 0 + \varepsilon]$ or $x_2 = [0.1 - \varepsilon, 0.1)$, $x_3 = (0.9, 0.9 + \varepsilon]$ or $x_3 = [1 - \varepsilon, 1)$ and $x_4 = (1, 1 + \varepsilon]$ or $x_4 = [1.1 - \varepsilon, 1.1)$, then $\tau_X = \{\{-0.1, 0, 0.1\}, \{0.9, 1, 1.1\}\}$ is produced, demonstrating again that there is a relation between the refinement of the topological space and the support of the associated probability measure from which samples are acquired. Then, it is assumed that there are data points in X, independent and identically acquired from a measurable topological space (Z, τ_Z) contained in a larger space (Ω, τ_Ω) (typically the usual topological space $\mathbb{R}^d$), with an unknown structure but having some associated probability distribution $P(Z)$ from which some data clustering algorithm attempts to unveil (Z, τ_Z). In this sense, the goal of a clustering algorithm is to fetch the inherent relations of neighborhoods belonging to (Z, τ_Z), which are

also of unknown nature. The produced clustering model endows a collection of neighborhoods, which is referred in this PhD thesis to as a **neighborhood topology** $\mathcal{N}(X)$ such that:

Definition 37 (Neighborhood topology)**.** Given a set X and a neighborhood map $\eta : x \mapsto \mathcal{N}(x)$, with $\mathcal{N}(x)$ being an open neighborhood of x, a neighborhood topology is defined as $\mathcal{N}(X) := \overline{\bigcup_{x \in X} \mathcal{N}(x)}$, being $\overline{A}$ the closure of the topology τ_A.

From such topology, it is expected that the associated topological space adequately represents (Z, τ_Z), i.e., it presents some invariance with respect to (Z, τ_Z), being homeomorphisms or isomorphisms among homology groups. Therefore, the Data Clustering problem is defined as:

Definition 38 (Data Clustering problem)**.** The Data Clustering problem consists in finding, given sampled instances $x \in X$ from an unknown topological space $(Z, \tau_Z) \subset (\Omega, \tau_\Omega)$ endowed with a Borel σ-algebra, and with $(X, \tau_X) \subset (Z, \tau_Z)$, a neighborhood topology $\mathcal{N}(X)$ of $(X, \mathcal{N}(X))$ which adequately represents (Z, τ_Z). In this sense, such neighborhood topology $\mathcal{N}(X)$ should approximate topological features of the unknown topology τ_Z. Random variables are independent and identically sampled from some unknown probability distribution which is supported on Z. Clusters are obtained from an equivalence relation derived from $\mathcal{N}(X)$.

For instance, suppose the topological space (Z, τ_Z) formed on the usual topology $\mathbb{R}$, such that $\tau_Z = \{(-1,1),(2,3)\}$. Such topological space is assumed to endow a Borel σ-algebra, thus being Borel-measurable with support $\text{supp}_Z(\mu) = \{z \in Z \mid z \in N_z \in \{(-1,1),(2,3)\} \implies \mu(N_z) \geq 0\}$, i.e., the measure lies on the neighborhoods around every $z \in Z$. Now suppose a uniform probability distribution $P(Z)$ defined over the topological space (Z, τ_Z), such that every open set $(a,b) \subset \tau_Z$ has probability $\frac{(b-a)}{(3-2)+(1-(-1)} = \frac{b-a}{3}$. Then take a set $X = \{-0.5, -0.4, 2.5\}$ sampled from Z and a neighborhood topology induced from open balls $B_{0.1}(x) = \{x' \in X \mid d(x,x') < 0.1\}$, such that $\mathcal{N}(X) = \{[-0.6,-0.3],[2.4,2.6]\}$. Clearly, (Z, τ_Z) and $(X, \mathcal{N}(X))$ are homeomorphic and the probability of a new sample producing a non-homeomorphic space is $\mu(\tau_Z / \mathcal{N}(X)) = \mu(\tau_Z) - \mu(\mathcal{N}(X)) = (1.3 + 0.4 + 0.4 + 0.4)/3 = 0.8\overline{3}$. In fact, as $\mathcal{N}(X)$ is too refined when comparing to (Z, τ_Z), a new sample has a high probability to form an non-homeomorphic topological space. Once the underlying topological space and the probability distribution are unknown, no such analysis can be made through data points. Even if another sample X' is acquired from Z, some analysis comparing $\mathcal{N}(X)$ and $\mathcal{N}(X')$ cannot be explicitly performed as $P(Z)$ is unknown. In spite of that, homology allows the study of topological structures, such as connected components and holes, disregarding the presence of (Z, τ_Z) and $P(Z)$, such as made in Topological Data Analysis (TDA). In this sense, a data clustering model can be evaluated by verifying how the associated homology behaves when data is subject to perturbations. Depending on the guarantees of stability, the consistency of the homology group can be assured.

Then, over-refined and over-coarsed topological spaces are defined as:

Definition 39 (Over-refined topological spaces)**.** Given some perturbation X' of a dataset X, the measurable topological space $(X,\mathcal{N}(X))$ is over-refined with respect to a topological equivalence relation $\sim_{\textbf{Top}}$ if $P(X) \subset P(Z) \implies (X,\tau_X) \not\sim_{\textbf{Top}} (X',\tau_{X'})$ almost certainly, i.e, if η produces topological spaces associated with a finer probability measure $P(X)$ than $P(Z)$, such that (X,τ_X) is unlike to preserve its topological features when data is subject to perturbations.

Remark 9. Topological equivalence relations refer to any relation that represents an invariance over a topological structure such as: homeomorphisms, homotopy equivalences (HATCHER, 2000), and isomorphisms between homology groups.

Definition 40 (Over-coarsed topological spaces)**.** Given some perturbation X' of a dataset X with d dimensions and the S^d sphere, the measurable topological space $(X,\mathcal{N}(X))$ is over-coarsed with respect to a topological equivalence relation $\sim_{\textbf{Top}}$ if $P(Z) \subset P(X) \implies (X,\tau_X) \sim_{\textbf{Top}} (X',\tau_{X'}) \sim_{\textbf{Top}} S^d$ almost certainly, i.e, if the neighborhood map η produces topological spaces associated with a probability measure $P(X)$ coarser than $P(Z)$, such that (X,τ_X) presents the same topological features to the sphere S^d.

In this sense, it is possible, for instance, to study the impact of over-refinements in the 0-dimensional homology group H_0, associated with the number of connected components. Whether H_0 is not isomorphic for both $\mathcal{N}(X)$ and $\mathcal{N}(X')$, then such neighborhood topologies will not form homeomorphic topological spaces. Therefore, if the neighborhood topology $\mathcal{N}(X)$ is refined enough to produce singletons, then, depending on how perturbations are bounded, the 0-dimensional homology group H_0 of $\mathcal{N}(X)$ will be only isomorphic to H_0 of $\mathcal{N}(X')$ if $|X| = |X'|$.

Theorem 2. Two over-refined topologies $\mathcal{N}(X)$ and $\mathcal{N}(X')$, being X' perturbed by substituting x_i with x_i' but following the same data distribution, produce isomorphic 0-dimensional homology groups if and only if $|X| = |X'|$.

Proof 1. Assume that, given a neighborhood topology $\mathcal{N}(X)$, the space (Ω,τ_Ω) (we remind the reader that, typically, (Ω,τ_Ω) is a closed cube in $\mathbb{R}^p$) is endowed with a probability measure P (in the Borel σ-algebra of (Ω,τ_Ω)) supported in $(Z,\tau_Z) \subset (\Omega,\tau_\Omega)$ and such that each corresponding cluster has measure $\leq \varepsilon$. Consider the presence of some sampled perturbation $x' = x+\delta$, in which δ is an element of a measurable set D endowed with a probability function. One issue that motivated the dilemma on over-coarsed versus over-refined topologies (in the same sense of the Bias-Variance dilemma) in this thesis is that whenever an element in any $C \in \mathcal{N}(X)$ becomes unrelated with any neighborhood in $\mathcal{N}(X)$, the topology changes as well as its homology group H_0. Then, we formulate the probability of the topology being "cut" (or divided) as:

$$\begin{aligned} P((x+D) - \textstyle\bigcup_{C\in\mathcal{N}(X)} C) &= P(x+D) \\ &\quad - P\left(\textstyle\bigcup_{C\in\mathcal{N}(X)} (C \cap (x+D))\right) \\ &= 1 - \textstyle\sum_{C\in\mathcal{N}(X)} P(C|x+D). \end{aligned} \tag{4.1}$$

Note that, if $\varepsilon \to 0$, i.e., the model tends to over-refinement, and some measure $\mu_Z(D)$ does not depend on ε, then $P(C|x+D) \to 0$ and $P(x \notin C) \to 1$ for every $C \in \mathcal{N}(X)$. In this case, the topological space determined by $\mathcal{N}(X')$ is not homeomorphic to the one determined by $\mathcal{N}(X)$ nor its homology group $H_0[(X', \mathcal{N}(X'))]$ is isomorphic to $H_0[(X, \mathcal{N}(X))]$. On the other hand, if the elements $C \in \mathcal{N}(X)$ cover $x+D$, then $P(x \notin C) = 0$ for every $C \in \mathcal{N}(X)$ and, therefore, $H_0[(X', \mathcal{N}(X'))] \cong H_0[(X, \mathcal{N}(X))]$.

Conversely, an excessive coarsening on the neighborhood topology leads to a representation problem as some details will vanish. For example, as the radius adopted in the data clustering increases, the hole of the 2-dimensional torus disappears (Figure 2 in Chapter 1). Therefore, there is a dilemma in choosing the adequate refinement for a neighborhood topology in the data clustering problem in order to guarantee the model stability while also ensuring data representability. As in Carlsson (2009), we consider that data clustering presents richer information whenever a hierarchical structure is considered, besides we claim that the lower hierarchies are unable to generalize data. As a consequence, this PhD thesis takes persistent homology to support the formulation of a generalization bound for the DC and HC problems.

4.3 Coarse-Refinement Dilemma and generalization bounds for data clustering

From a topological perspective, a Hierarchical Clustering algorithm performs a sequence of operations to find, from a pre-defined neighborhood map, the corresponding filtration of topological spaces such that:

$$\mathcal{F}(X, \eta) := (X, \mathcal{N}_{\alpha_0}(X)) \subseteq (X, \mathcal{N}_{\alpha_1}(X) \subseteq \cdots \subseteq (X, \mathcal{N}_{\alpha_k}(X))$$

with $\mathcal{N}_{\alpha_i}(X) := \overline{\bigcup_{x \in X} \eta_{\alpha_i}(x)}$, and then it associates $\mathcal{F}(X, \eta)$ with a pre-defined equivalence relation.

For instance, a single linkage algorithm adopts a distance function as the pre-defined neighborhood map to produce a metric space. A filtration is recursively defined such that clusters are merged whenever they present the minimum distance among their points, i.e., given a distance function d and some collection C of clusters, if two clusters $C_1, C_2 \in C$ are such that $(C_1, C_2) := \arg\min_{C_1, C_2 \in C} d(c, c')$ for all $c \in C_1$ and $c' \in C_2$, they are merged, forming another hierarchical level. In this context, the equivalence relation arises from the contiguity of the elements belonging to clusters.

However, among all spaces in the filtration, there are those which will not guarantee consistency for topological features of a model, as they form over-refined partitions. In this sense, we claim that it is reasonable to define as an adequate HC model the one whose topological features do not change even when data is subject to perturbations. So, the Hierarchical Clustering problem is defined as:

Definition 41 (Hierarchical Clustering problem)**.** The Hierarchical Clustering problem consists in finding, for each $i \in \{1,\dots,k\}$ in which k is a refinement index (e.g. such as the radius index for a metric space), some filtration of topological spaces that preserves their topological features when data is subject to perturbations.

Assuming that such perturbed dataset X' is produced by the substitution of an element in X, such that $X' = \{x_1,\dots,x_{i-1},x'_i,x_{i+1},\dots,x_n\}$ and there is some neighborhood topology $\mathcal{N}(X)$, it is not possible to identify whether $(X,\mathcal{N}(X)) \subseteq (X,\mathcal{N}(X))$ or $(X,\mathcal{N}(X)) \subseteq (X,\mathcal{N}(X))$ such that their relations are not explicitly defined. On the other hand, if the inclusion of new samples is considered as a perturbation such that $X^m = \{x_1,\dots,x_n,x'_1,\dots,x'_m\}$, then $(X,\mathcal{N}(X)) \subseteq (X^q,\mathcal{N}(X^q)) \subseteq (X^r,\mathcal{N}(X^r)) \subseteq (X^m,\mathcal{N}(X^m))$ for all $0 < q < r < m$.

Therefore, with $\mathcal{X}_i$ being a topological space associated with a neighborhood topology $\mathcal{N}_{\alpha_i}(X)$, and X^m being the inclusion of m new data samples, it is possible to construct a bifiltration $\mathcal{F}(X,X/X^m,\eta)$ as follows:

$$\begin{array}{ccccccc}
\mathcal{X}_1 & \xrightarrow{\iota_{\mathcal{N}}} & \mathcal{X}_2 & \xrightarrow{\iota_{\mathcal{N}}} & \dots & \xrightarrow{\iota_{\mathcal{N}}} & \mathcal{X}_k \\
\downarrow{\scriptstyle \iota_X} & & \downarrow{\scriptstyle \iota_X} & & \downarrow{\scriptstyle \iota_X} & & \downarrow{\scriptstyle \iota_X} \\
\mathcal{X}_1^1 & \xrightarrow{\iota_{\mathcal{N}}} & \mathcal{X}_2^1 & \xrightarrow{\iota_{\mathcal{N}}} & \dots & \xrightarrow{\iota_{\mathcal{N}}} & \mathcal{X}_k^1 \\
\downarrow{\scriptstyle \iota_X} & & \downarrow{\scriptstyle \iota_X} & & \downarrow{\scriptstyle \iota_X} & & \downarrow{\scriptstyle \iota_X} \\
\vdots & \xrightarrow{\iota_{\mathcal{N}}} & \vdots & \xrightarrow{\iota_{\mathcal{N}}} & \vdots & \xrightarrow{\iota_{\mathcal{N}}} & \vdots \\
\downarrow{\scriptstyle \iota_X} & & \downarrow{\scriptstyle \iota_X} & & \downarrow{\scriptstyle \iota_X} & & \downarrow{\scriptstyle \iota_X} \\
\mathcal{X}_1^m & \xrightarrow{\iota_{\mathcal{N}}} & \mathcal{X}_2^m & \xrightarrow{\iota_{\mathcal{N}}} & \dots & \xrightarrow{\iota_{\mathcal{N}}} & \mathcal{X}_k^m
\end{array}$$

with inclusions (homomorphisms) $\iota_{\mathcal{N}} := \mathcal{X}_i^l \subseteq \mathcal{X}_j^q$ and $\iota_X := \mathcal{X}^{l-1} \subseteq \mathcal{X}^l$, and such that, for every i,j,X^l and X^q with $q > l$, the following diagram is commutative:

$$\begin{array}{ccc}
\mathcal{X}_i^q & \xrightarrow{\iota_{\mathcal{N}}} & \mathcal{X}_j^q \\
\downarrow{\scriptstyle \iota_X} & & \downarrow{\scriptstyle \iota_X} \\
\mathcal{X}_i^l & \xrightarrow{\iota_{\mathcal{N}}} & \mathcal{X}_j^l.
\end{array}$$

Persistent homology groups $\mathbf{f}_p^{i,j} : H_p(\mathcal{X}_i^l) \to H_p(\mathcal{X}_j^l)$ and $\mathbf{f}_p^{l,q} : H_p(\mathcal{X}_i^l) \to H_p(\mathcal{X}_i^q)$ can be then defined considering the tame functions $f : (\mathcal{X}_i^l) \mapsto (-\infty,i]$ and $g : (\mathcal{X}_i^l) \mapsto (-\infty,l]$ as well as the study of consistency and stability of HC models can be made from the critical coordinates of the bifiltration $\mathcal{F}(X,X/X^m,\eta)$.

For instance, consider the 0-dimensional homology group when a neighborhood topology $\mathcal{N}_i(X)$ is formed from open balls and let $\mathcal{N}_i(X^m)$ be a neighborhood topology built up after m new data, such that the radius associated to each collection of open balls i is r_i. If the diameter of $\mathcal{N}_i(X^m)$ is sufficiently greater than that of $\mathcal{N}_i(X)$, a new instance x'_q is likely to form another

cluster (with probability close to one) given $P(\mathcal{N}_i(X^m)/\mathcal{N}_i(X)) = \frac{\mu_{X^m}(\mathcal{N}_i(X))}{\text{supp}_{X^m}(\mathcal{N}_i(X)^m)}$. This implies that $\text{rank}\{\text{im}(\mathbf{f}_0^{0,i} \circ \mathbf{g}_0^{0,m})\} = \text{rank}\{\text{im}(\mathbf{f}_0^{0,i})\} + m$, where $\mathbf{g}$ is the map induced on the homology after the data inclusions, being related to the tame function g.

Hence, $\mathbf{f}_0^{0,i}(H_p(\mathcal{X}_i))$ will never adequately represent $\mathbf{f}_p^{0,i}(H_p(\mathcal{X}_i^m))$ and the analyzed topological features will vary along data insertions. In order to $\mathbf{f}_0^{0,i}(H_p(\mathcal{X}_i))$ properly represent $\mathbf{f}_p^{0,i}(H_p(\mathcal{X}_i^m))$, $\mathbf{g}_p^{0,m}$ must likely be an isomorphism. Therefore, a DC model is consistent from the following definition:

Definition 42 (DC p-homology consistence)**.** A DC model, associated with the neighborhood topology $\mathcal{N}_i(X)$, is p-homology consistent if $\mathbf{g}_p^{0,q}(H_p(\mathcal{X}_i))$ is most likely an isomorphism, for all $q = 1,\ldots,m$ and $m \to \infty$.

In addition, Hierarchical Clustering requires the study of a filtration along the domain of the tame function f, then a set of morphisms $\mathbf{g}_p^{0,m}$ is applied over $\mathbf{f}_p^{0,k}(H_p(\mathcal{X}_k))$, for $k = i,\ldots,j$, to verify isomorphisms among them. In the ideal scenario, all $\mathbf{g}_p^{0,m}$ are isomorphisms for $k = i,\ldots,j$, but, in Hierarchical Clustering, the indices of filtrations are previously chosen thus restricting the set of morphisms. Therefore, critical points lying on any interval $[k-1,k)$ will be disregarded causing a resolution problem (in terms of $\text{dom}(f)$) that is inherent of such analysis. Assuming this limited set of morphisms $\mathbf{g}_p^{0,m}$, the behavior of $\mathbf{f}_p^{i,j}(H_p(\mathcal{X}_k))$ can be study as follows:

Lemma 3. Given the persistences $\mathbf{f}_p^{i,j}$ and $\mathbf{g}_p^{0,m}$ of a bifiltration, if $\mathbf{g}_p^{0,m}(H_p(\mathcal{X}_i))$ and $\mathbf{g}_p^{0,m}(H_p(\mathcal{X}_i^m))$ are isomorphisms then $\mathbf{f}_p^{0,i}(H_p(\mathcal{X}_0^m)) \cong \mathbf{f}_p^{0,i}(H_p(\mathcal{X}_0))$ and $\mathbf{f}_p^{i,j}(H_p(\mathcal{X}_i^m)) \cong \mathbf{f}_p^{i,j}(H_p(\mathcal{X}_i))$.

Proof 2. The proof follows from the commutativeness of the bifiltration diagram.

Considering Lemma 3, if the isomorphism is almost certain for $\mathbf{g}_p^{0,q}(H_p(\mathcal{X}_i))$ and $\mathbf{g}_p^{0,q}(H_p(\mathcal{X}_j))$ given $q = 1,\ldots,m$ and $m \to \infty$, then, $\mathbf{f}_p^{i,j}(H_p(\mathcal{X}_i^m)) \cong \mathbf{f}_p^{i,j}(H_p(\mathcal{X}_i))$ is likely to occur for any inserted data, i.e., $\mathbf{f}_p^{i,j}(H_p(\mathcal{X}_i))$ consistently represents $\mathbf{f}_p^{i,j}(H_p(\mathcal{X}_i^m))$ along the values of i, j. In sequence, we define a Hierarchical Cluster p-homology consistence as:

Definition 43 (HC p-homology consistence)**.** A HC model, associated with the filtration $\mathcal{F}(X,\eta)$, is p-homology consistent if for all $i \leq k \leq j$, $\mathcal{N}_k$ is $p-$homology consistent (Definition 42).

In this context, by verifying if $\mathbf{g}_p^{0,m}(H_p(\mathcal{X}_i))$ is non-isomorphic not only allows to locally study $\mathcal{X}_i$, but also to represent $\mathbf{f}_p^{i,t}$ as $t \to \infty$ such as shown next:

Theorem 4. Given the associated simplicial complexes built up from $\mathcal{X}_i$, if $\mathbf{g}_p^{0,m}(H_p(\mathcal{X}_i))$ is not an isomorphism then $\mathbf{f}_p^{i,t}(H_p(\mathcal{X}_i)) \ncong \mathbf{f}_p^{i,t}(H_p(\mathcal{X}_i^m))$ for $t \to \infty$.

Proof 3. Given simplicial complexes are considered, then, for an arbitrary unknown compact topological space (Z,τ_Z), $H_p[(Z,\mathcal{N}_t(Z))] = \mathbb{Z}$ for $p = 0$ and $H_p[(Z,\mathcal{N}_t(Z))] = 0$ for $p > 0$ and $t \to \infty$. Therefore, if $\mathbf{g}_p^{0,m}(H_p(\mathcal{X}_i))$ is not an isomorphism, as

$\text{dom}(\mathbf{f}_p^{i,t}(H_p(\mathcal{X}_i))) \ncong \text{dom}(\mathbf{f}_p^{i,t}(H_p(\mathcal{X}_i^m)))$, i.e., the homology classes of $\text{dom}(\mathbf{f}_p^{i,t}(H_p(\mathcal{X}_i)))$ which die, or are created after i, are different when compared to $\mathbf{f}_p^{i,t}(H_p(\mathcal{X}_i^m))$, then $\mathbf{f}_p^{i,t}(H_p(\mathcal{X}_i)) \ncong \mathbf{f}_p^{i,t}(H_p(\mathcal{X}_i^m))$ (as illustrated in Figure 26).

Corollary 1. If $\mathbf{f}_p^{i,t}(H_p(\mathcal{X}_i)) \ncong \mathbf{f}_p^{i,t}(H_p(\mathcal{X}_i^m))$, there is at least one critical point k in interval $[i,t)$ for which $\text{dom}(\mathbf{f}_p^{k,t}(H_p(\mathcal{X}_k)) \cong \text{dom}(\mathbf{f}_p^{k,t}(H_p(\mathcal{X}_k^m))$ as $t \to \infty$ (as illustrated in Figure 26).

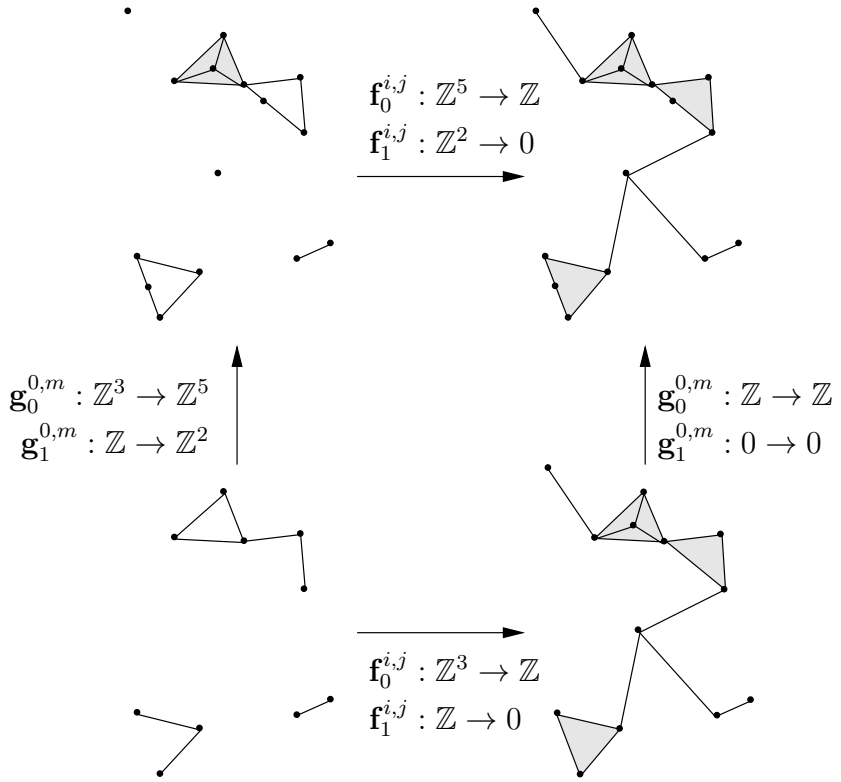


Figure 26 – Illustration of the changes in $\mathbf{f}_p^{i,t}$ domain caused by an non-isomorphic $\mathbf{g}_p^{0,m}$ applied over $H_p(\mathcal{X}_i)$ (Adapted from Carlsson and Zomorodian (2009)).

Then, as $t \to \infty$, the p-th persistence homology $\mathbf{f}_p^{i,t}(H_p(\mathcal{X}_i))$ and $\mathbf{f}_p^{i,t}(H_p(\mathcal{X}_i^m))$ will be equivalent if $\mathbf{g}_p^{0,m}(H_p(\mathcal{X}_i))$ is an isomorphism. Although, even if $\mathbf{g}_p^{0,m}(H_p(\mathcal{X}_i))$ is isomorphic, there is no guarantee that a critical point $k \in [i,t)$ will produce an isomorphism $\mathbf{g}_p^{0,m}(H_p(\mathcal{X}_k))$. An adequate representation is then assured when enough points in $[i,j)$ are chosen to characterize all possible critical points in the filtration.

The morphism $\mathbf{g}_p^{0,m}$ devises the changes along the levels of some hierarchical clustering model (given by the p-th persistent homology $\mathbf{f}_p^{i,t}(H_p(\mathcal{X}_i))$), when the HC model is subject to data insertions. Then, a probability measure P_μ endowed with a measure μ, built up from the produced p-th homology groups, is required to study the probability for which $\mathbf{g}_p^{0,m}(H_p(\mathcal{X}_i))$ is isomorphic. The p-th Betti number can be considered such measure:

Theorem 5. The p-th Betti number is a measure over a σ-algebra formed by a collection of all p-th homology groups associated with simplicial complexes.

Proof 4. Let (Z,τ_Z) be a topological space and recover the definition of the p-th Betti number, we have $\beta_p := \text{rank}\{H_p[(Z,\tau_Z)]\} = \text{rank}\{\text{im}(\partial_{p+1})/\ker(\partial_p)\}$. Considering p-dimensional simplicial complexes, for a limited natural number b, $\text{im}(H_p[(Z,\tau_Z)])$ is a free abelian group $\mathbb{Z}^b$ and, therefore, $\text{rank}\{H_p[(Z,\tau_Z)]\} = b$. Hence, there is a collection $\mathcal{H} = \{0,\mathbb{Z}^{b_1},\mathbb{Z}^{b_2},\dots\}$, such that β_p is a group transformation $\beta_p : (\mathcal{H},\oplus) \to (\mathbb{N},+)$ where $\oplus$ and $+$ are the direct and the usual sum respectively, and β_p is a bijective map. We can construct a $\sigma-$algebra closed under a countable disjoint union $(\phi(\mathcal{H}),\Sigma)$ from the one-by-one mapping $\phi : \mathcal{H} \to \mathcal{A}$, with $\mathcal{A} = \{A_{b_1},A_{b_2},A_{b_3},\dots | A_{b_i} \cap A_{b_j} = \varnothing \; \forall \; i \neq j\}$ such that, for a set of indices $I \subset \mathbb{N}$, there is an isomorphism between $\bigoplus_{i\in I}\mathbb{Z}^{b_i}$ and $\bigsqcup_{e\in E}A_{b_i}$ (hence a natural isomorphism for $F : (\mathcal{H},\oplus) \mapsto (\mathcal{A},\sqcup)$). From Σ, we have the following measure function:

$$\mu(A_{b_i} \in \Sigma) = b_i,$$

which respects:

- Non-negativity as $b_i \geq 0$;
- Null-empty set with $b_i = 0$ (induced by $\text{im}(\partial_{p+1}C_{p+1})/\ker(\partial_p C_p) = \varnothing$);
- Countable-additivity as $\mu(\bigsqcup_{i\in I}A_{b_i}) = \sum_{i\in I} b_i$.

It is worth to mention that these properties are associated one-by-one with the rank of the abelian group $(\mathcal{H},\oplus)$ (p-th Betti number) since:

- $\mu(A_{b_i}) = \text{rank}\{\mathbb{Z}^{b_i}\}$;
- $\mu(\bigsqcup_{i\in I}A_{b_i}) = \text{rank}\{\bigoplus_{i\in I}\mathbb{Z}^{b_i}\} = \sum_{i\in I} b_i$.

Therefore, $\beta_p[(X,\tau_X)]$ is a measure endowed of a disjoint set σ-algebra from which a probability measure $P_{\beta_p}(X)$ can be defined with expected value $\mathbb{E}_X(\beta_p(X)) = \sum_{k=1,\dots,\infty}\beta_p(X_k)P_{\beta_p}(X_k)$.

For instance, suppose four neighborhood topologies $\mathcal{X}_i^l$, with $l = 0,1,\dots,3$, forming a sequence $(H_p(\mathcal{X}_i^l))_{l=0,\dots,3} = (\mathbb{Z}^3,\mathbb{Z}^4,\mathbb{Z}^5,\mathbb{Z}^5)$ associated with the filtration $\mathcal{F}(X,X^m,\eta)$ and the collection $\mathcal{H} = \{H_p(\mathcal{X}_i^l)\}$, from which it is possible to define a measurable space $S_\mu := (\mathcal{A},2^{\mathcal{A}})$ endowed with a measure μ given $\mathcal{A} = \{A_3,A_4,A_5,A_5\}$. Such space supports, then, a probability measure of μ which boundly diverges within an open ball with radius ε whenever new data are sampled, i.e., $P(|\mu(\mathcal{X}_i^l) - \mu(\mathcal{X}_i^0)| < \varepsilon)$. Therefore, in this example, the probability of μ diverging (i) by one is $1/3$, and (ii) by two is $2/3$.

In this sense, topological features can be measured by p-th Betti numbers, employed to compare topological spaces $\mathcal{X}_i$ built up from a Data Clustering or Hierarchical Clustering algorithm. Let us define the ***p*-th Betti measure** as a map from some topological filtration, induced by the tame function g (corresponding to data insertions), to a filtration of σ-algebras

that devises a stochastic process $\beta_p(\mathcal{X}_i^m|\mathcal{X}_i, x'_1, x'_2, \ldots, x'_m)$ with x'_i representing data insertions. As $\mathbb{E}[\beta_p(\mathcal{X}_i^m)] = \beta_p(\mathcal{X}_i)$ is required to guarantee that a DC or HC model is consistent, $\beta_p(\mathcal{X}_i^m)$ should behave as a martingale (VILLE, 1939), from which its difference is defined as:

$$\mathbb{E}\left[\beta_p(\mathcal{X}_i^m) - \beta_p(\mathcal{X}_i^{m-1})\right] = 0.$$

From Equation 4.2, the expectation of arising and vanishing p-th degree homology classes produced by the inclusion of m new data points is:

$$\mathbb{E}\left[\beta_p(\mathcal{X}_i^m) - \beta_p(\mathcal{X}_i^{m-1}) + \cdots + \beta_p(\mathcal{X}_i^1) - \beta_p(\mathcal{X}_i)\right] = 0, \tag{4.2}$$

implying:

$$\mathbb{E}\left[\beta_p(\mathcal{X}_i^m) - \beta_p(\mathcal{X}_i)\right] = 0. \tag{4.3}$$

Hence, the convergence of a generalization measure is defined as:

Definition 44 (p-homology generalization)**.** Given neighborhood topologies $\mathcal{N}(X)$ and $\mathcal{N}(X^m)$, a p-homology generalization is given whenever:

$$G_p(\mathcal{X}_i) := |\beta_p(\mathcal{X}_i^m) - \beta_p(\mathcal{X}_i)| \tag{4.4}$$

tends to zero as $m \to \infty$.

In this sense, adopting Azuma (1967) Inequality (Inequality C.1 in C), the lemma of the DC p-homology convergence (the proof is in C) is given by:

Lemma 6 (DC p-homology convergence)**.** Let X and X^m be two datasets with i.i.d. samples such that $X^m = X \cup \{x'_1, \ldots, x'_m\}$, $\mathcal{X}_i$ be the neighborhood topological space built up from a points cloud and $\beta_p(\cdot)$ be the p-th Betti number calculated upon the neighborhood topologies, the probability that the average absolute difference $\sum_{q=1}^m |\beta_p(\mathcal{X}_i^q) - \beta_p(\mathcal{X}_i)|$ is bounded by ε exponentially decays with respect to:

$$P\left(\sum_{q=1}^{m} |\beta_p(\mathcal{X}_i^q) - \beta_p(\mathcal{X}_i)| > m\varepsilon\right) \leq 2\exp\left(\frac{-m^2\varepsilon^2}{2\sum_{q=1}^m c_{q,p}^2}\right), \tag{4.5}$$

given $c_{q,p} = |\beta_p(\mathcal{X}_i^q) - \beta_p(\mathcal{X}_i)|$ is bounded.

Then, whenever the difference between the number of p-th homology classes of $\mathcal{X}_i^q$ and $\mathcal{X}_i$ asymptotically increases faster than m, there is no guarantee that, in average, $\beta_p(\mathcal{X}_i^q)$ approximates $\beta_p(\mathcal{X}_i)$. Conversely, if $|\beta_p(\mathcal{X}_i^q) - \beta_p(\mathcal{X}_i)|$ is asymptotically smaller than m, $|\beta_p(\mathcal{X}_i^m) - \beta_p(\mathcal{X}_i)| \to 0$, and, therefore, $\mathbf{g}_{0,m}(\mathcal{X}_i)$ is likely to produce an isomorphism.

Remark 10. For instance, considering the case of H_0, if each new sample increases the number of connected components by one, i.e., if an over-refinement occurs, the terms $c_{q,p}$ grow as follows:

$$\sum_{q=1}^{m} c_{q,0}^2 = \sum_{q=1}^{m} q^2 = \frac{1}{6}m(m+1)(2m+1),$$

consequently, in this case, Inequality 4.5 diverges as follows:

$$P\left(\sum_{q=1}^{m}|\beta_p(\mathcal{X}_i^q)-\beta_p(\mathcal{X}_i)|>m\varepsilon\right)\leq 2\exp\left(\mathcal{O}(m^{-1})\right),$$

hence, there is no guarantee that $\mathbf{g}_{0,m}(\mathcal{X}_i)$ is isomorphic.

Corollary 2. The generalization can not be assured whenever:

$$\Omega\left(c_{q,p}\right)=m,$$

and convergence of the right-hand term in Inequality 4.5 occurs whenever:

$$\mathcal{O}\left(c_{q,p}\right)<\sqrt{m}.$$

Therefore, whenever $\mathcal{O}\left(c_{q,p}\right)<\sqrt{m}$, $\mathbf{g}_{0,m}(\mathcal{X}_i)$ is likely to produce an isomorphism. In this sense, a maximum value $\overline{c}_{p,q}$ bounded as described in Corollary 2 also guarantees consistency [1], then:

Corollary 3. Let $\overline{c}_{q,p}=\max_{q=1,\ldots,m}(c_{q,p})$ and m be the number of new samples added to the dataset X:

$$\overline{c}_p<\varepsilon\sqrt{m/2\ln 2}. \tag{4.6}$$

Such criterion can be used to select which subspace in a filtration will be considered for a clustering model, avoiding non-stable and non-consistent partitions.

The result of Lemma 6 can, therefore, be extended for the HC problem. Remember that every simplicial complex of $\mathcal{X}_i^r$ and $\mathcal{X}_i^q$ in the filtrations, with $r,q=1,\ldots,m$, must be associated with the same pre-image of the tame function f. In this sense, the convergence of HC models in terms of p-th Betti numbers is held by respecting the following theorem:

Theorem 7 (HC p-homology convergence)**.** Let $\mathrm{im}(f)^{-1}$ be the pre-image of the tame function f associated with the filtrations $\mathcal{X}_i^q\subseteq\cdots\subseteq\mathcal{X}_j^q$, and a bijective function $h:\mathrm{im}(f)^{-1}\to\mathbb{N}$ which maps the points of $\mathrm{im}(f)^{-1}$ onto the index of such filtrations. The DC p-homology convergence for each one of those points is given by:

$$P\left(\sup_{t\in\mathrm{im}(f)^{-1}}\sum_{q=1}^{m}|\beta_p(\mathcal{X}_{h(t)}^q)-\beta_p(\mathcal{X}_{h(t)})|>m\varepsilon\right)\to 0 \text{ as } m\to\infty,$$

[1] For the sake of curiosity, note that as $c_{q,p}$ only assumes natural values, the sequence $\mathcal{C}=(c_{p,l})_{l=1,\ldots,m}$ is sparse whenever DC p-th consistence is guaranteed, something that comprises the future work.

with:

$$P\left(\sup_{t\in \mathrm{im}(f)^{-1}} \sum_{q=1}^{m} |\beta_p(\mathcal{X}^q_{h(t)}) - \beta_p(\mathcal{X}_{h(t)})| > m\varepsilon\right) \leq 2M \exp\left(\frac{-m\varepsilon^2}{2\tilde{c}_p^2}\right),$$

where M is the cardinality of the pre-image $\mathrm{im}(f)^{-1}$ of the tame function f, and $\tilde{c}_p = \max_{i\in\mathbb{R}} \bar{c}_{p,i}$.

Corollary 4. Considering Theorem 7, let us define $\overline{\Delta\beta}_{p,t} = \sum_{q=1}^{m} |\beta_p(\mathcal{X}^q_{h(t)}) - \beta_p(\mathcal{X}_{h(t)})|$ and take $\delta = P\left(\sup_{t\in \mathrm{im}(f)^{-1}} \overline{\Delta\beta}_{p,t} \leq m\varepsilon\right)$, then the HC p-homology convergence is guaranteed whenever:

$$\frac{\tilde{c}_p^2 \ln M}{m} \to 0 \ \text{ as } \ m \to \infty. \tag{4.7}$$

Corollary 5. If $\tilde{c}_p$ is constant, the convergence of Equation 4.7 is equivalent to:

$$\frac{\ln M}{m} \to 0 \ \text{ as } \ m \to \infty.$$

Remark 11. The parameter $\tilde{c}_p$ can be stated as constant, for example, such as in partitional clustering when using K-means (STEINHAUS, 1956; MACQUEEN, 1967), K-medoids (KAUFMAN; ROUSSEEUW, 1987) and Self-Organizing Maps (KOHONEN, 1982) which rely on optimization procedures, typically based on some distance from data points and pre-determined centroids. Considering an i.i.d. data distribution and enough samples, those centroids will never sufficiently vary as new data are acquired, hence producing the same number of connected components. Although, as the number of centroids increases w.r.t. the new samples acquired, it is likely that those algorithms identify every point as a single cluster, and the DC 0-homology consistence is not guaranteed.

Considering the H_0 homology group, the generalization is guaranteed for every critical point greater than the first that ensures the DC 0-homology consistence, as proved in the following lemma:

Lemma 8. Assuming the H_0 homology group, we have that for every critical point $i < j$ which implies in $\beta_0(\mathcal{X}^q_i) > \beta_0(\mathcal{X}^q_j)$, if $\mathcal{X}_i$ is DC 0-homology consistent then $\mathcal{X}_j$ is also DC 0-homology consistent.

Proof 5. As $\mathcal{X}^q_i$ is DC 0-homology consistent, following Lemma 6, we have that $\max_{q=1,\ldots,m}(|\beta_p(\mathcal{X}^q_i) - \beta_p(\mathcal{X}_i)|)$ must be bounded and, therefore, $\beta_p(\mathcal{X}^q_i)$ and $\beta_p(\mathcal{X}_i)$ are also bounded. In this sense, as $\beta_p(\mathcal{X}^q_i), \beta_p(\mathcal{X}_i) \geq 0$, $\beta_p(\mathcal{X}^q_j) \leq \beta_p(\mathcal{X}^q_i)$ and $\beta_p(\mathcal{X}_j) \leq \beta_p(\mathcal{X}_i)$, then $\max_{q=1,\ldots,m}(|\beta_p(\mathcal{X}^q_j) - \beta_p(\mathcal{X}_j)|) < \infty$ implies that $\mathcal{X}_j$ is stable, hence it is DC 0-homology consistent.

Following these consistency results, we conclude that Kleinberg (2002) richness axiom must be relaxed as it considers partitions which do not guarantee the p-homology consistence. Section 3.4 discusses how our results are related with Carlsson and Mémoli (2010)'s consistency for their adapted single linkage method.

4.3.0.1 The Topological Concept Drift

Considering the scenario in which the over-refined partitions are disregarded, i.e., all clusters will be consistent/stable as long as my probability distribution is i.i.d., we can assume a variation occurring in the topology associated with the clusters is generated by an anomaly in the system which is being measured, i.e, in the unknown underlying topological space or probability distribution. We define these anomalies as Topological Concept Drifts (TCD) such that:

Definition 45 (Topological Concept Drifts)**.** Given a filtration $\mathcal{F}(X,\eta)$ of p-th homology consistent topological spaces $\mathcal{X}_i$ built on top of some dataset X acquired by an i.i.d. probability distribution, an Topological Concept Drift occurs whenever for a new data acquisition X', $H_p(\mathcal{X}_i) \ncong H_p(\mathcal{X}'_i)$ for one or more index i.

In this sense, it is possible to measure when a topological change occurs given anomalous variations in the underlying and unknown topology. For instance, the filtration parameters (such as radius) which lead to inconsistency in $\mathcal{X}_i$ and $\mathcal{X}'_i$ are disregarded in the clustering construction on top of both X and X'. Then, $\mathcal{X}_i$ and $\mathcal{X}'_i$ are compared one by one for all i in order to detect the TCD. Note that the filtration parameters can be removed by the criterion presented in Equation 4.6 for obtaining the p-th homology consistent clusters.

4.3.1 An ultrametric analysis for the Coarse-Refinement Dilemma

Recording the consistency proof by Carlsson and Mémoli (2010), discussed in Section 3.4, Carlsson and Mémoli assume a function $f_X : \mathbb{R}^+ \to \mathbb{R}^+$ as $r \mapsto \min_{x\in \text{supp}(X)} \mu_X(B_X(x,r))$, which is non-decreasing and $f(X) > 0$ for all $r > 0$, and another function $F_X : \mathbb{N}\times\mathbb{R}^+ \to \mathbb{R}^+$ be defined by $(m,\delta) \mapsto \frac{e^{-mf_X(\delta/4)}}{f_X(\delta/4)}$ in order to prove the consistency for their modified version of the single linkage algorithm.

In this sense, f_X defines the finest measure for a given radius r. For instance, suppose $r_i \leq r_j$ such that r_i and r_j are associated with the same partition that is originally formed by r_i and, therefore, $f_X(r_i) = f_X(r_j) = \mu_X(B_X(x_i,r))$. On the other hand, F_X characterizes the exponential decay in terms of the number of elements and the radius adopted to build up some clustering. In order to prove consistency, Carlsson and Mémoli fix a radius δ to bound F_X and avoid its increasing. However, as $B_X(x,r)$ reduces, $\mu_X(B_X(x,r))$ tends to be uniform as $B_X(x,r)$ induces the trivial topology and, therefore, $f_X(\varepsilon) = 1/m$ for an infinitesimal epsilon.

Given $\beta_0(\mathcal{X}_\delta)$, i.e., the number of connected components in the clustering associated with the radius δ, we prove that if $\beta_0(\mathcal{X}_\delta)\ln\beta_0(\mathcal{X}_\delta)$ reduces w.r.t. m, then Carlsson and Mémoli (2010)'s consistency is guaranteed. On the other hand, if the 0-dimensional homology group increases, thus Inequality 3.1 does not converge to zero.

Theorem 9. If $\frac{\beta_0(\mathcal{X}_{\delta/4})\ln\beta_0(\mathcal{X}_{\delta/4})}{m}$ diverges as $m \to \infty$, Carlsson and Mémoli (2010)'s consistency is not guaranteed and, given a measurable topological space $\overline{\mathcal{X}}_r$ with $\mathcal{B}(\overline{\mathcal{X}}_r) \subseteq B^*(X,r)$, being

$B^*(X,r) = \arg\min_{B_X(x,r)|x\in \text{supp}(X)} \mu_X(B_X(x,r))$, if $\frac{\beta_0(\overline{\mathcal{X}}_r)^2}{m}$ converges then Carlsson and Mémoli (2010)'s consistency holds.

Proof 6. At the first step of the proof for Theorem 1, Carlsson and Mémoli (2010) study the probabilistic convergence, in terms of the Hausdorff distance, between a restricted mm-space $\mathcal{S}_{mm}(X) = (X, d_X, \mu_X)$ and the support of μ_X in form:

$$P_{\mu_Z}\left(d_H^X(X, \text{supp}(\mu_X)) > \delta\right) \leq F_X(n,\delta).$$

On a next formulation, they find that:

$$P_{\mu_Z}\left(d_H^X(X, \text{supp}(\mu_X)) > \delta\right) \leq Ne^{-mf_X(\delta/4)},$$

with N being equals to the cardinality of the maximal $\delta/4$-packing of $\text{supp}(\mu_X)$. A δ-packing is defined as:

Definition 46. Let $(\mathfrak{M}, d_{\mathfrak{M}})$ be a metric space and $\mathcal{M} \subset \mathfrak{M}$. $\{\mathcal{M}_1, \ldots, \mathcal{M}_n\}$ is an δ-packing of $\mathcal{M}$ if $\bigcap_{i=1}^n B(\mathcal{M}_i, \delta) = \varnothing$.

Therefore, the topological space induced by the space $(\mathfrak{M}, d_{\mathfrak{M}})$ is a Hausdorff space and its cardinality is given by its number of connected components $\beta_0(\mathcal{X}_{\delta/4})$, implying that Carlsson and Mémoli (2010)'s consistency implicitly considers the number of 0-dimensional homology classes in its formulation. Then, let $\mathfrak{L}(\cdot,\cdot)$ be defined as the single linkage function, whenever the induced neighborhood topology is over-refined, there is no guarantee that Inequality 3.1 will converge since:

$$P_{\mu_Z}\left(d_{GH}\left(\mathfrak{L}(X,d_X), \mathfrak{L}(A,d_A)\right) > \zeta\right) \leq P_{\mu_Z}\left(d_H^X(X, \text{supp}(\mu_X)) > \zeta)\right),$$

as proved in Carlsson and Mémoli (2010). More precisely, as $1/f_X(\delta/4)$ is a superior bound for N, then:

$$\frac{\beta_0(\mathcal{X}_{\delta/4}) \ln \beta_0(\mathcal{X}_{\delta/4})}{m} \leq \frac{\ln N}{mf_X(\delta/4)},$$

so the divergence of Inequality 6 occurs whenever:

$$\frac{\beta_0(\mathcal{X}_{\delta/4}) \ln \beta_0(\mathcal{X}_{\delta/4})}{m} \to \infty \text{ as } m \to \infty.$$

Hence, if the number of connected components (0-dimensional homology classes) increases w.r.t. the number of data points m, there is no convergence guarantee for Carlsson and Mémoli (2010)'s consistency, i.e., if the clustering model is not DC 0-homology consistent (Definition 42), then Carlsson and Mémoli (2010)'s consistency is not guaranteed. In order to prove the upper bound for Carlsson and Mémoli (2010)'s consistency, we apply Taylor's Theorem in $\ln N$ resulting in:

$$\frac{\ln N}{mf_X(\delta/4)} \leq \frac{N}{mf_X(\delta/4)} - \frac{1}{mf_X(\delta/4)}.$$

As there is a measurable topological space $\overline{\mathcal{X}}_r$ with $\mathcal{B}(\overline{\mathcal{X}}_r) \subseteq B^*(X,r)$, being $B^*(X,r) = \arg\min_{B_X(x,r)|x\in \text{supp}(X)} \mu_X(B_X(x,r))$, then:

$$\frac{\beta_0(\overline{\mathcal{X}}_r)^2}{m} \geq \frac{N}{mf_X(\delta/4)}$$

implies

$$\frac{\ln N}{mf_X(\delta/4)} \leq \frac{\beta_0(\overline{\mathcal{X}}_r^2)}{m} - \frac{1}{mf_X(\delta/4)} \leq \frac{\beta_0(\overline{\mathcal{X}}_r)^2 - \beta_0(\overline{\mathcal{X}}_r)}{m}.$$

Therefore, if there is some $\mathcal{B}(\overline{\mathcal{X}}_r) \subseteq B^*(X,r)$ assuring the convergence of $\frac{\beta_0(\overline{\mathcal{X}}_r^2)}{m}$, Carlsson and Mémoli (2010)'s consistency is guaranteed, i.e., if the DC 0-homology convergence ($\frac{\beta_0(\overline{\mathcal{X}}_r)^2}{m} \leq \frac{\bar{c}_p^2}{m} \to 0$ as $m \to \infty$ in Inequality 4.6) is guaranteed, then Carlsson and Mémoli (2010)'s consistency holds. Conversely, if Carlsson and Mémoli (2010)'s consistency diverges, DC 0-homology convergence does not hold. Following the Lemma 8, as the considered mathematical object is the ultrametric space, those consistencies are ensured for all radius $r > \delta/4$ that were determined by the single linkage algorithm. In addition, as Carlsson and Mémoli (2010)'s consistency does not hold in over-refined metric spaces, then the H_0 homology group of its induced topology will not be DC 0-homology consistent.

4.4 Final considerations

The generalization property in the context of data clustering is here defined as the invariance in terms of topological features, what guarantees the scale-invariance and consistency formulated by (KLEINBERG, 2002). For 0-dimensional homology groups, such property is guaranteed whenever the associated topological spaces are coarse enough, w.r.t. the support of the unknown underlying topological space (Z, τ_Z), to not produce new clusters as new data are included. On the other hand, if such topological spaces are excessively coarse, then the clustering model will simply correspond to the universe set.

As the underlying topological space (Z, τ_Z) is unknown, it is only possible to evaluate such clustering model with respect to its reconfiguration after new data samplings. Taking this into account, the next chapter will provide experimental studies about how those models change whenever new data are sampled from the same distribution, as well as a concept drift analysis based on our consistency results.

CHAPTER

5

EXPERIMENTS AND RESULTS

5.1 Initial Considerations

The choice of an adequate clustering model depends on a trade-off between p-homology consistence and the detailing of topological features devised from the refinement (or coarsening) of the neighborhood topologies in the Data Clustering (DC) or Hierarchical Clustering (HC) model, i.e., the Coarse-Refinement Dilemma (CRD) discussed in the previous chapter. If p-th homology classes lie on a stable filtration interval with respect to data insertions, then they represent well the topology of data points as it persists along the data sampling. On the other hand, whenever such classes change with respect to those perturbations, they are not p-homology consistent so they should be disregarded.

Supposing the p-homology consistence of topological spaces built on top of i.i.d. sampled data points (with guaranteed statistical representability), as new data are sampled further changes in its associated p-th homology groups indicate the presence of some new topological feature, i.e., a Topological Concept-Drift (TCD). In this scenario, we describe and present, along this chapter, experimental results to corroborate the CRD and TCD. This chapter is organized as follows: (i) experimental setup and methods discussing the tools and frameworks employed in our experiments; (ii) experiments performed on toy scenarios that include synthetic datasets to study the CRD and the TCD; (iii) dynamical systems attractors were considered in an attempt to analyze the clustering of data points sampled from Lorenz, Rössler and Mackey-Glass systems; and, lastly, (iv) we considered the TCD in a real-world problem featuring document semantics along time, in terms of 0-th homology groups. Such datasets were chosen given they present a variety of topological characteristics regarding the 0 and 1-dimensional homology groups, except for the document-related one from which we did not have any previous information about the underlying topological space.

5.2 Experimental setup and methods

The experimental study of this PhD thesis mainly relies on two aspects: the analysis of CRD and the study of TCD with respect to 0-th homology groups. Given the distinct nature of the datasets employed in such aspects, different experimental approaches are required. The synthetic datasets corresponding to the Torus, Crescent Moon, and the attractors are evaluated in terms of their 0-th and 1-th homology groups, requiring techniques for the simplicial complex construction. The 0-th and 1-th homology groups were selected as follows: (i) the required allocated memory increases exponentially with the degree of the homology group, and, (ii) the considered groups already contain relevant information about such datasets. Computational limitations also determined two different strategies adopted to build up the simplicial complexes along each dataset given its corresponding sample size and dimension. For instance, the Vietoris-Rips (VIETORIS, 1927) complex were employed on the Torus while the Lazy Witness (SILVA; CARLSSON, 2004) on the remaining datasets.

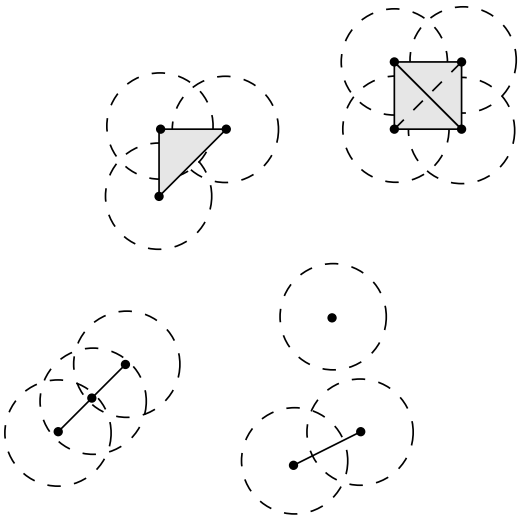


Figure 27 – Construction of a simplicial complex using the Vietori-Rips technique (Adapted from Carlsson (2009)).

The Vietoris-Rips complex is an abstract simplicial complex defined from a metric space given open balls with radius r such that the simplices are formed whenever the distance of its vertices is less then $2r$, i.e., open balls around them intersect themselves as illustrated in Figure 27. Such algorithm presents exponential asymptotic complexity and requires extensive computational resources depending on the dataset size. In this sense, Lazy Witness reduces the asymptotic complexity in the construction of simplicial complexes by setting up a subset of points, named as **landmarks**, over which the witness complex is formed. Given a set X of data samples, a set L of landmarks, and some radius r, as illustrated in Figure 28, a simplex $\Delta = [v_1 v_2 \dots v_k]$ is defined in the witness complex whenever there is a point $w \in X/L$ such that $\Delta \subseteq B(w,r)$, i.e., the simplex is contained in the open ball with radius r around w. Hence, the

asymptotic complexity of the Lazy Witness algorithm is derived from the number of landmarks and not from the dataset size.

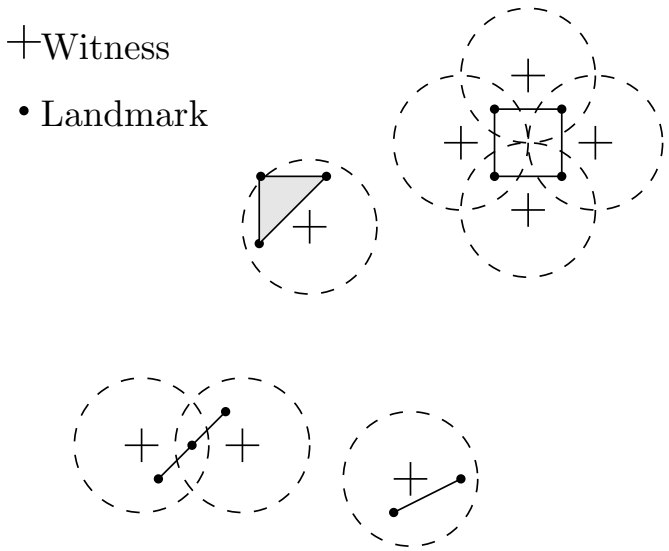


Figure 28 – Construction of a simplicial complex using the witness complex technique (Adapted from Silva and Carlsson (2004))

.

In order to provide a visualization of the persistence among such built-in simplicial complexes, Barcode plots (CARLSSON, 2009) were adopted. As illustrated in Figure 46 of Appendix E, a Barcode plot is constructed on top of a bidimensional space in which the vertical axis represents each homology class produced along a filtration whose indices are mapped into the horizontal axis. Such filtration indices correspond to the radius of open balls employed by either the Vietoris-Rips or the Lazy Witness algorithm (VIETORIS, 1927; SILVA; CARLSSON, 2004). From such space, a barcode is horizontally drawn along the interval in which the corresponding homology class persists, thus depicting where it appears and vanishes. Furthermore, heatmaps, formed by a space $(r, m, \mathcal{X}_r)$ (remind $G_p(\mathcal{X}_r) := |\beta_p(\mathcal{X}_r^m) - \beta_p(\mathcal{X}_r)|$ from Definition 44), were adopted in order to show how p-th Betti measures diverge from $\mathcal{X}_r$ whenever m new data elements are included, having r as the radius of all employed open balls, m as the number of included data, and β_p as the p-th Betti measures calculated from the simplicial complexes.

On the other hand, datasets associated with TCD were studied regarding only 0-th homology groups, i.e., their connected components. Therefore, the DBSCAN algorithm was adopted in order to provide noise filtering while clustering as datasets might present undesirable artifacts. Such algorithm regards the density of an open ball $B(\cdot, r)$ around some data point in order to merge it to another group. Formally, given such open ball and two points $c_i \in C_i$ and $c_j \in C_j$ belonging to clusters C_i and C_j, if $c_i, c_j \in B(c_i, r)$ (or conversely $c_i, c_j \in B(c_j, r)$) and $|B(c_i, r)| \geq \rho$ (or $B(c_j, r) \geq \rho$), with a pre-defined ρ, then C_i and C_j are merged.

In this sense, a tridimensional filtration, associated with the inclusions, points density, and radius, is required in order to analyze how 0-th homology groups change along the considered periods. Hence, such tridimensional filtration was constructed at each period T

disregarding the parameters $\rho^{\ncong}$ and $r^{\ncong}$ which had produced inconsistent homology classes (with respect to Equation 4.4), forming the data space $(\rho^{\cong}, r^{\cong}, G_0(\mathcal{X}_{\rho^{\cong}, r^{\cong}}))$, with $a^{\cong}$ being the set of parameters that produces consistent homology classes disregarding those associated with over-coarsening, i.e., when $\beta_0(\mathcal{X}_r^m) = 1$ for all m. Therefore, taking into account two periods T_i and T_j, $|\beta_0(\mathcal{X}_{\rho_{\cap}^{\cong}, r_{\cap}^{\cong}}, T_i) - \beta_0(\mathcal{X}_{\rho_{\cap}^{\cong}, r_{\cap}^{\cong}}, T_j)|$ calculates the divergence in the 0-th homology groups between period T_i and T_j (i.e., what we refer to as a TCD) being $\rho_{\cap}^{\cong}$ and $r_{\cap}^{\cong}$, respectively, $\rho_{T_i}^{\cong} \cap \rho_{T_j}^{\cong}$ and $r_{T_i}^{\cong} \cap r_{T_j}^{\cong}$.

5.3 Toy scenarios

We introduce our experiments to confirm the presence of the CRD and the TCD with synthetic datasets associated with the torus and the "crescent moon" geometry [1]. Such results were acquired from the successive calculus of the generalization measure (Equation 4.4) along the insertion of new data points and the increase of radii adopted to construct the corresponding simplicial complexes, being illustrated in the form of a Barcode plot or a heatmap of the space $(r, m, |\beta_p(\mathcal{X}_r^m) - \beta_p(\mathcal{X}_r)|)$. Finally, a synthetic dataset built up to represent the merging of three Gaussians along time was employed to study TCD.

5.3.1 Torus

As an introductory experiment, a torus geometry relying on internal and external radii of, respectively, $r = 0.5$ and $R = 1$ was considered in order to form the underlying topology (Z, τ_Z) from which the data samples were acquired. The torus equation:

$$f(R, r, \theta, \phi) = ((R + r\cos\theta)\cos\phi, (R + r\cos\theta)\sin\phi), \tag{5.1}$$

was adopted such that ϕ and θ assumed 400 values uniformly sampled in the interval $[0, 2\pi]$, producing the dataset illustrated in Figure 39 (see Appendix D).

The perturbed dataset was obtained after the inclusion of 100 new samples, following the same torus parameters, and, consequently, the same distribution function. A Vietoris-Rips filtration (VIETORIS, 1927) was produced with the TDA package from the R Project of Statistical Computing[2] considering radii in interval $[0, 1)$ and producing the barcode plots for the 0-th and the 1-th homology classes, as illustrated in Figures 46 and 47 (both found in Appendix E).

From such barcodes, the generalization $G_p(\mathcal{X}_r)$ was calculated for radii values from 0 to 1 with a step of 0.01 such as illustrated at the bottom charts of Figures 46 and 47. As shown in Figure 46, the minimum value for $G_0(\mathcal{X}_r)$ is given whenever $r = 0.01$ or $r \in [0.2376, 1)$ and, therefore, they produce generalization for 0-th homology classes, i.e., the number of connected

[1] The dataset of "crescent moon" geometry can be found in the package RSSL (https://cran.r-project.org/web/packages/RSSL/) of R Project for Statistical Computing.

[2] TDA package – <https://cran.r-project.org/web/packages/TDA/index.html>

components does not change along the data insertion. Complementary, when 1-th homology classes are considered, the interval of radii $[0, 0.0891)$ does not present any homology classes so that it should be disregarded. Nonetheless, the interval $[0.4257, 1)$ produces generalization for 1-th homology classes that vanishes at radius 0.88, once the torus hole is filled up.

5.3.2 Crescent Moon

The Crescent Moon dataset was produced from the RSSL package of the R Project of Statistical Computing[3] using the function *generateCrescentMoon* with parameters $n = 5,000$ ($10,000$ data points were produced as n is the number of elements per class), $d = 2$ and $\sigma = 0.5$ (see Figure 40 in Appendix D). From such dataset, a Lazy Witness algorithm was employed with 200 landmarks in order to produce a bifiltration associated with the addition of 200 new samples having the initial dataset with $5,100$ elements up to a total of $10,000$ elements while varying the radii in $[0, 3)$.

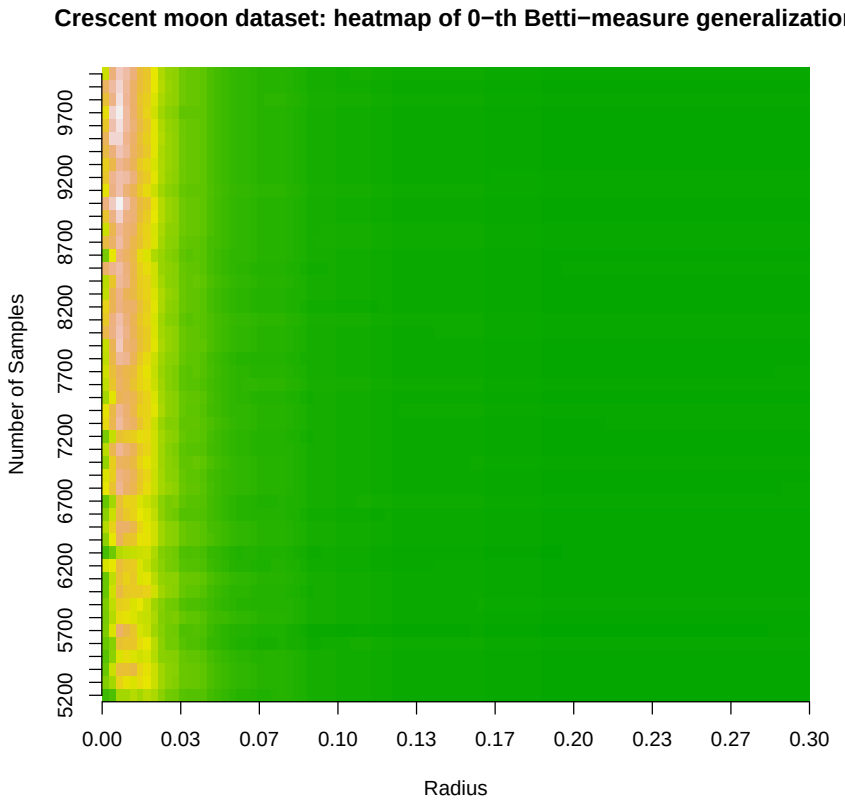


Figure 29 – The heatmap generated from the Crescent Moon dataset formed by values of the generalization measure $G_0(\mathcal{X}_r)$ along the insertion of samples and the increase of radii.

The barcode plot was generated for 0-th homology classes built up from $X^{5,100}$ and $X^{10,000}$ such as illustrated in Figure 48 (Appendix E). The generalization $G_p(\mathcal{X}_r)$ was calculated over $\mathcal{X}_r^{5,100}$ and $\mathcal{X}_r^{10,000}$ resulting in the bottom chart of Figure 48 (Appendix E), from which it

[3] https://cran.r-project.org/web/packages/RSSL/

is possible to identify the 0-th homology inconsistency when the radii are in $[0, 0.84)$. Although the interval $[0.84, 3)$ is associated with one single 0-th homology class, the crescent moon dataset is known to present two classes, hence there are data points lying between its clusters which are merging them together when Lazy Witness is performed. Therefore, some problems require the regard of density-based clustering algorithms. In addition, Figure 29 presents how $G_p(\mathcal{X}_r)$ significantly increases when over-refined metric spaces are considered and new data is sampled, such that there is no guarantee for the 0-th homology consistence, thus demonstrating the CRD for such particular scenario.

5.3.3 Topological Concept Drift on a synthetic dataset

In order to study Topological Concept Drift (TCD), we performed experiments using data sampled from three bidimensional Gaussian distributions whose mean values approximate the space center $(0,0)$ along time. In this sense, as illustrated in Figure 44 (Appendix D), $3,000$ data points were sampled from each Gaussian whose mean values were $(0,0)$, $(5,5)$, and $(20,20)$, and variances were 0.2, 1, and 2, respectively. The centroids $\mu_i(t)$ of such Gaussians approximate the origin of the coordinate system in a factor of $1/10$, i.e., at each time step, $\mu_i(t+1) = \mu_i(t) - \mu_i(t)/10$.

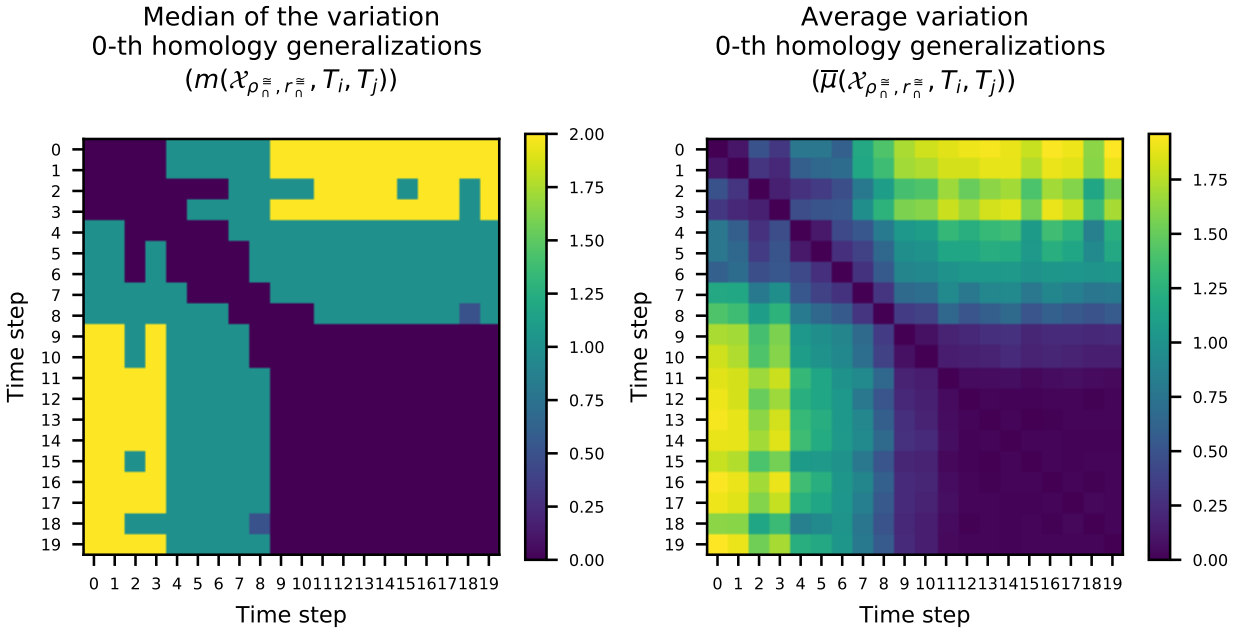


Figure 30 – Heatmaps produced from the median and average variations on 0-th homology generalizations of $\mathcal{X}_{\rho^{\cong}_{\cap}, r^{\cong}_{\cap}}(T_i)$ and $\mathcal{X}_{\rho^{\cong}_{\cap}, r^{\cong}_{\cap}}(T_i)$.

At each time step, the DBSCAN algorithm was run, as explained in the two last paragraphs of Section 5.2, with the following setting $\rho = (k)_{k=1,\ldots,25}$ and $r = (0.01 + 0.1s)_{s=1,\ldots,30}$. As illustrated in Figure 55 (Appendix F), after $T = 4$ the 0-th consistent homology groups produced in $\mathcal{X}_{\rho^{\cong}, r^{\cong}}(T_0)$ start to substantially vary. In this sense, from $T = 1$ to $T = 20$, the

median variation of 0-th homology generalizations of $\mathcal{X}_{\rho^{\cong}_{\cap},r^{\cong}_{\cap}}(T_i)$ and $\mathcal{X}_{\rho^{\cong}_{\cap},r^{\cong}_{\cap}}(T_i)$ given by:

$$m(\mathcal{X}_{\rho^{\cong}_{\cap},r^{\cong}_{\cap}},T_i,T_j) = \text{median}\left(|\beta_0(\mathcal{X}_{\rho^{\cong}_{\cap},r^{\cong}_{\cap}},T_i) - \beta_0(\mathcal{X}_{\rho^{\cong}_{\cap},r^{\cong}_{\cap}},T_j)|\right),$$

and its average variation is given by:

$$\overline{\mu}(\mathcal{X}_{\rho^{\cong}_{\cap},r^{\cong}_{\cap}},T_i,T_j) = \sum_{\rho'\in\rho^{\cong}_{\cap}}\sum_{r'\in r^{\cong}_{\cap}} \frac{|\beta_0(\mathcal{X}_{\rho',r'},T_i) - \beta_0(\mathcal{X}_{\rho',r'},T_j)|}{|r'||\rho'|},$$

which can be employed to evaluate the overall divergence of the 0-th homology generalization. Results illustrated in Figure 30 present the variations of such measure with respect to $T_i \in \{1,2,\ldots,20\}$ versus $T_j \in \{1,2,\ldots,20\}$. From such analysis, we can identify whether the overall generalization of its associated parameters lead to significant changes, thus summarizing the TCD occurred for each pair $(\rho^{\cong}, r^{\cong})$.

5.4 Dynamical systems attractors

Dynamical systems are mathematical representations that describe how points evolve along time in a geometrical space (ALLIGOOD; SAUER; YORKE, 2000). Such points are assumed to lie in a space named as state space which represents the relations between the current and previous points given by the dynamics of a system in a regime. For example, the Lorenz system is defined by the following system of equations:

$$\begin{cases} \frac{dx}{dt} = \sigma(y-x) \\ \frac{dy}{dt} = x(\rho - z) - y\,, \\ \frac{dz}{dt} = xy - \beta z \end{cases} \tag{5.2}$$

which sets how a point $p \in \mathbb{R}^3$ evolves along time, resulting in its state space (or phase space). For the sake of illustration, when setting $\sigma = 10$, $\beta = 8/3$ and $\rho = 28$, this system will present a chaotic nature.

Chaotic systems are dynamical systems which are sensitive to initial conditions and bounded. Suppose the chaotic Lorenz system, whose attractor is illustrated in Figure 41 (Appendix D), given two points $p = (x,y,z)$ and $p' = (x',y',z')$, even when $d(p,p') \leq \varepsilon$, with ε small enough, p and p' might be associated with different trajectories, therefore such system is sensitive to initial conditions. Such attractors represent how a dynamical system evolves independently of the initial points, i.e., how their trajectories behave given previous states.

For instance, as illustrated in Figure 41 (Appendix D), its attractor presents two orbits which are the geometrical structures formed by the cycling of a trajectory around an attraction point. Such class of geometrical structures can be identified by its topology as they form voids in the attractor manifold. For example, from a visual inspection, the attractor X of Figure 41 (Appendix D) presents homology groups $H_0(X) = \mathbb{Z}$ and $H_1(X) = \mathbb{Z}^2$ as it has one connected

component and two unidimensional holes. However, visual inspection becomes harder as space dimension increases.

Persistent homology can then be employed in order to study attractor orbits and, therefore, our results provide means to adequately choose the radii that have to be considered in such analysis since CRD impacts the representability and stability in the orbits modeling. Finally, assuming that elements continuously evolve in the dynamical systems trajectory, i.e., they do not "jump" (discontinuity) from a position to another, such systems will always present only one connected component and, therefore, only radii which produce such result must be considered.

5.4.1 Lorenz system

The Lorenz attractor was modeled by the Lorenz system implementation from the package *nonlinearTseries* of the R Project of Statistical Computing, setting $\sigma = 10$, $\beta = 8/3$ and $\rho = 28$ in order to produce some chaotic behavior. The initial conditions were set up as $x_0 = -13$, $y_0 = -14$ and $z_0 = 47$, and the time parameter varied from 0 to 50 with steps equal to 0.005 units. The attractor is illustrated in Figure 41 (Appendix D).

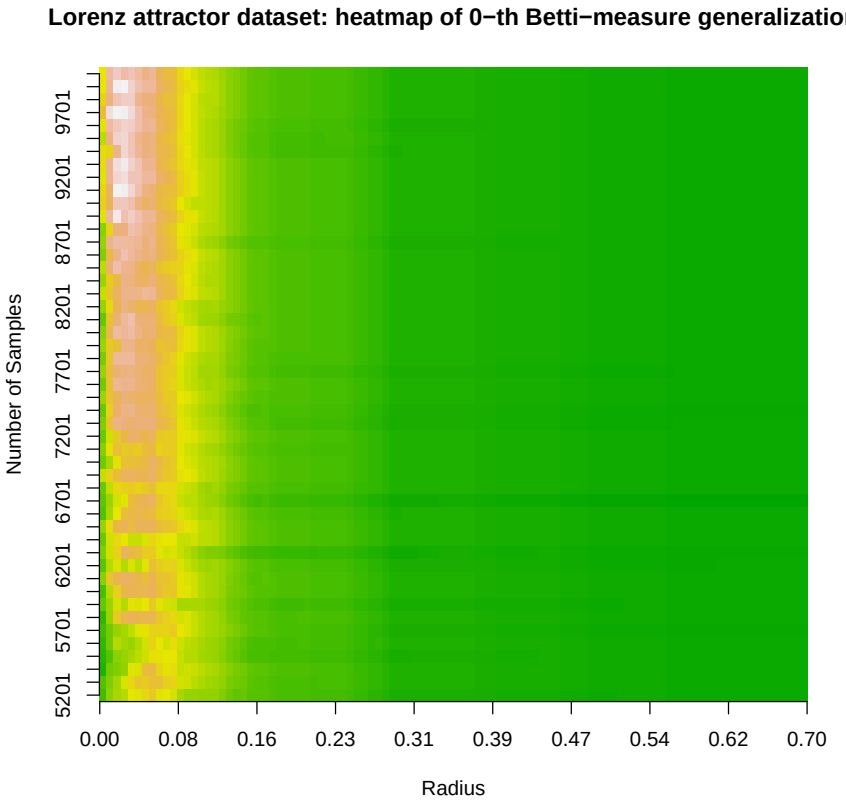


Figure 31 – The heatmap generated from the Lorenz attractor formed by values of the generalization measure $G_0(\mathcal{X}_r)$ along the insertion of samples and the increase of radius over the filtration.

The Lazy Witness algorithm was then employed considering 200 landmarks, defining a filtration with $10,000$ radii values in interval $[0,7)$ which are associated with simplicial

complexes with degree of at most one. As illustrated in Figure 49 in Appendix E, the interval $[3.829, 7)$ guarantees 0-th Betti measure generalization once it produces a single connected component. In addition, as illustrated in Figure 50 (Appendix E), the intervals $[1.2747, 1.4280)$, $[1.6037, 3.9333)$, $[4.4590, 5.5643)$, and $[5.6420, 7)$ assure the 1-th Betti measure generalization, being the three last of them the relevant/stable ones as only the interval $[3.829, 7)$ assures the 0-th Betti measure generalization. In this sense, the simplicial complex associated with the interval $[3.829, 3.9333)$ properly represents the two orbits of the considered attractor, the interval $[4.4590, 5.5643)$ is capable of representing only one and, finally, the interval $[5.6420, 7)$ does not identify any.

As illustrated in the heatmaps of Figures 31 and 32, the generalization measure diverges in the initial interval of radii when 0-th homology classes are considered, while it converges to zero in the complementary interval, confirming the CRD. Such measure presents a more complex behavior for 1-th homology classes: (i) it is constant and equals to zero when considering the first interval of radii, since no cycles were yet produced; (ii) as the radius increases, there is a fast growth in the number of 1-th homology classes; and, finally, (iii) this number slowly decays.

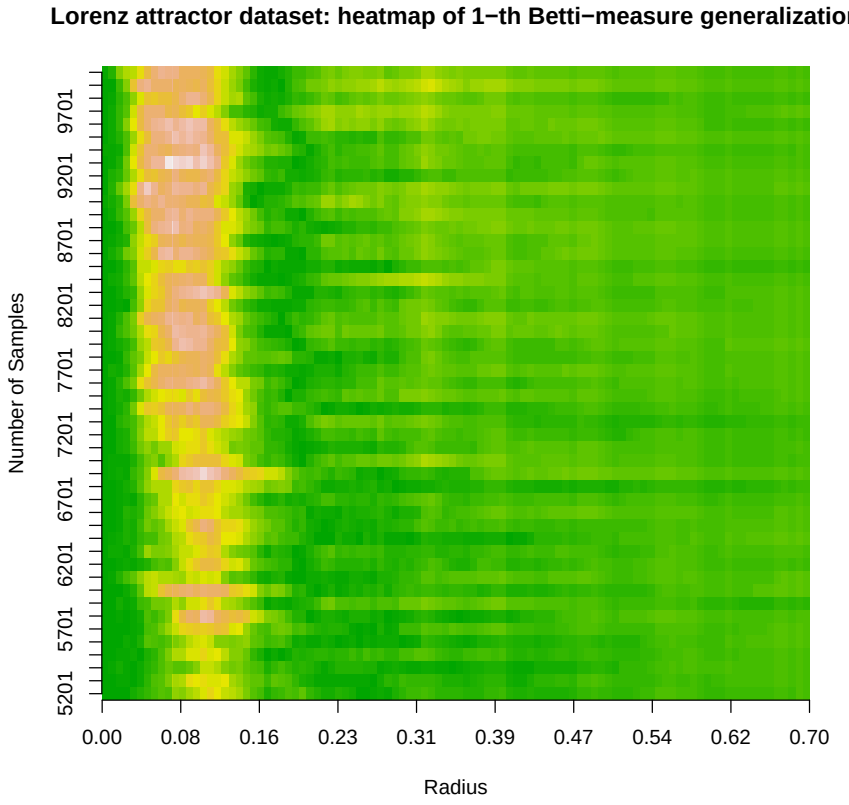


Figure 32 – The heatmap generated from the Lorenz attractor formed by values of the generalization measure $G_1(\mathcal{X}_r)$ along the insertion of samples and the increase of radius over the filtration.

5.4.2 Rössler system

Rössler system is formed by three non-linear ordinary differential equations, like the Lorenz system, such that:

$$\begin{cases} \frac{dx}{dt} = -y - z \\ \frac{dy}{dt} = x + ay \\ \frac{dz}{dt} = b + z(x - w) \end{cases}, \tag{5.3}$$

which may also depict chaotic behavior. We employed the system implemented in the package *nonlinearTseries* of the R Project of Statistical Computing, setting the parameters $a = 0.2$, $b = 0.2$ and $w = 5.7$, and the initial conditions $x_0 = -2$,$y_0 = -10$ and $z_0 = 0.2$, thus devising a chaotic Rössler system. A total of $20,000$ observations were acquired by varying time from 0 to 200 using steps of 0.01, thus generating the dataset illustrated in Figure 42 (Appendix D).

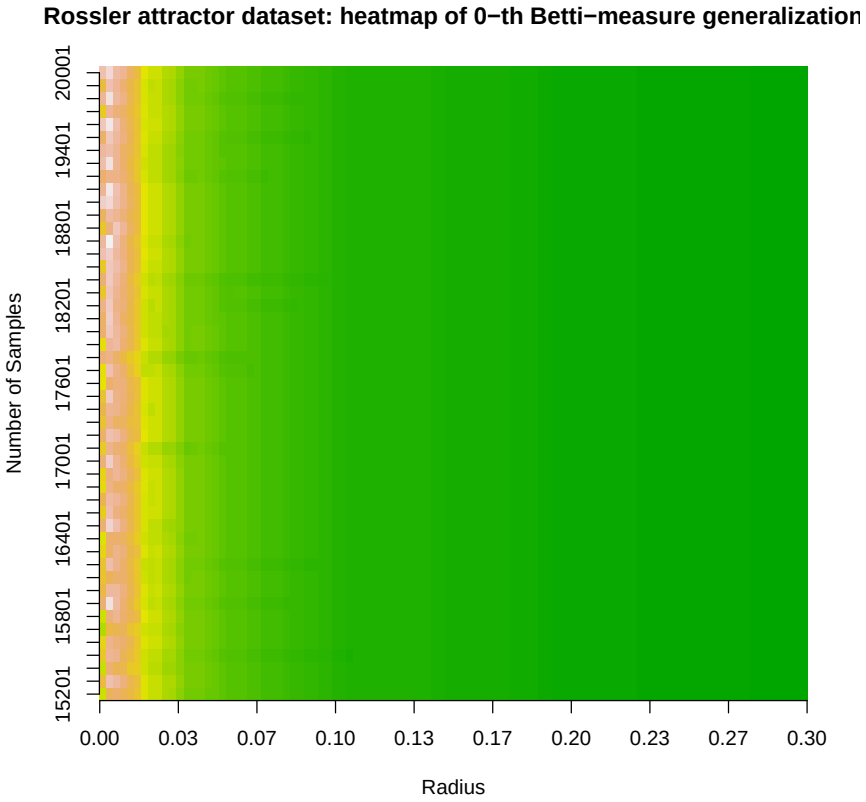


Figure 33 – The heatmap generated from the Rössler attractor formed by values of the generalization measure $G_0(\mathcal{X}_r)$ along the insertion of samples and the increase of radius over the filtration.

Considering the same approach as employed for the Lorenz system, i.e., the Lazy Witness algorithm, we relied on 200 landmarks to produce simplicial complexes of one degree at most, and their corresponding filtration was set in the radii interval $[0,5)$. Considering the 0-dimensional homology classes, the generalization $G_0(\mathcal{X}_r)$ converges to zero in the interval $[0.46,5)$, which identifies the single connected component related to this particular attractor, as illustrated in Figure 51 (Appendix E).

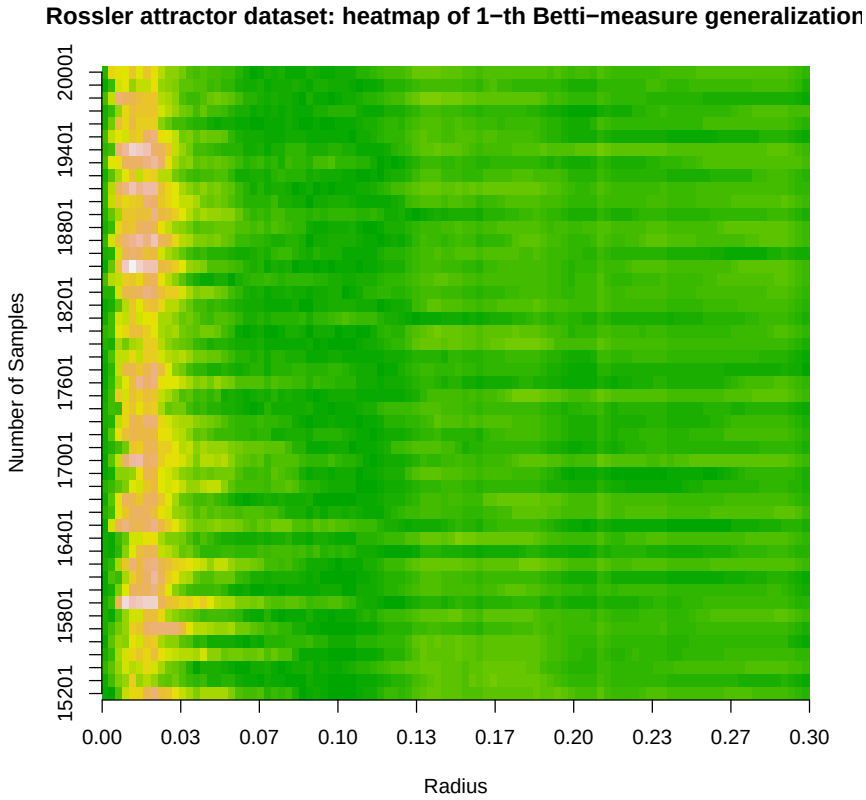


Figure 34 – The heatmap generated from the Rössler attractor formed by values of the generalization measure $G_1(\mathcal{X}_r)$ along the insertion of samples and the increase of radius over the filtration.

Taking 1-dimensional homology classes, as shown in Figure 52 (Appendix E), the intervals $[0.1050, 0.1095)$, $[2.2105, 2.3400)$, and $[2.4105, 5.000)$ guarantee $G_0(\mathcal{X}_r) \to 0$ as $m \to \infty$. In this sense, the two last intervals are considered to be relevant as $[0.1050, 0.1095)$ lies before $[0.46, 5)$ (the interval which guarantees the 0-th Betti measure generalization), and also, they are related to, respectively, the single orbit characteristic of Rössler attractors which is vanished along the filtration. Furthermore, the heatmaps of Figures 33 and 34 present the same behavior as discussed in the last paragraph of Section 5.4.1.

5.4.3 Mackey-Glass system

The Mackey-Glass system is defined after changes in a variable along time, which is set in $t - \tau$ as follows:

$$\frac{dx}{dt} = \frac{\beta x_\tau}{1 + x_\tau^n} - \gamma x. \tag{5.4}$$

Such system is applied, for instance, in the blood-cell production modeling (MACKEY; GLASS, 1977). We adopted a dataset from such system which comes with the package *frbs* of the R Project for Statistical Computing. As described in the documentation of *frbs* [4], such dataset was

[4] The documentation of *frbs* is available at <https://www.rdocumentation.org/packages/frbs/versions/3.2-0/topics/frbsData>

built considering $\beta = 0.2$, $\gamma = 0.1$, $n = 1$, $\tau = 17$, $x_0 = 1.2$ and discrete time. Moreover, it is formed after the embedding of the series in a five dimensional space given by the time-delayed states $x(t-18)$, $x(t-12)$, $x(t-6)$, $x(t)$ and, finally, $x(t+6)$, such that $(x(t-18), x(t-12), x(t-6), x(t), x(t+6))$ forms the Mackey-Glass attractor, as projected in two dimensions in Figure 43 (Appendix D).

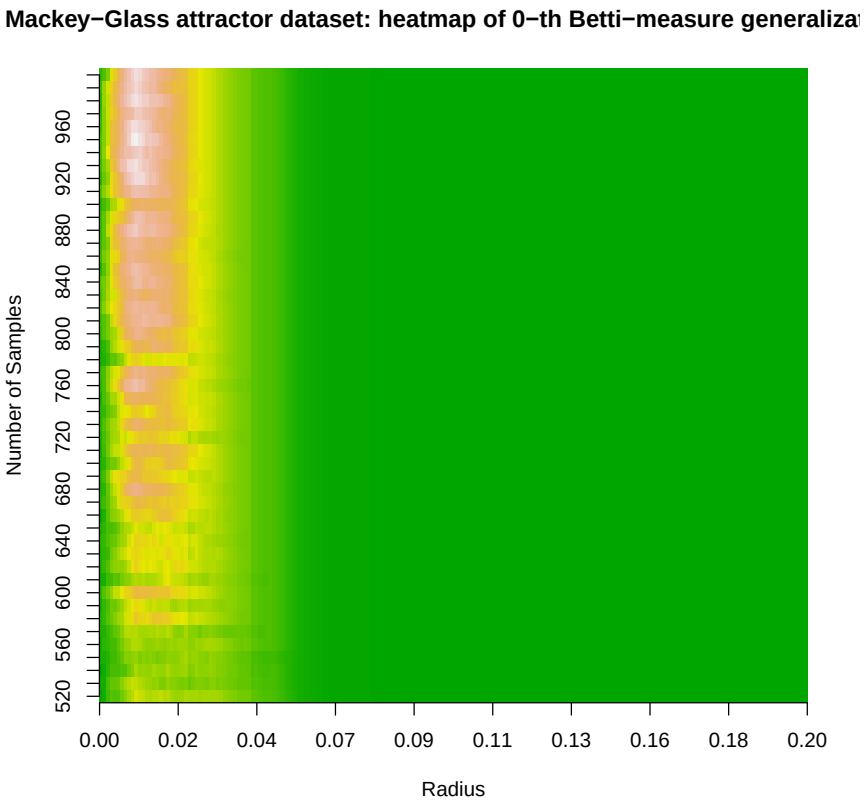


Figure 35 – The heatmap generated from the Mackey-Glass attractor formed by values of the generalization measure $G_0(\mathcal{X}_r)$ along the insertion of samples and the increase of radius over the filtration.

This dataset contains only $1,000$ observations and, therefore, the initial set X^{510} was configured with the first 510 elements and the remaining acts as data insertions. So 20 new elements were added to X^{510} until it reaches the total size of the dataset. Barcode plots considered only $X^{1,000}$ and X^{510} to be produced, and heatmaps took into account the dataset snapshots formed after every inclusion. In this sense, the generalization $|\beta_p(\mathcal{X}_r^{1,000}) - \beta_p(\mathcal{X}_r^{510})|$ is also shown using the barcode plots, and the heatmaps contain, along the data insertion, a sequence of generalizations $(|\beta_p(\mathcal{X}_r^{510+20i}) - \beta_p(\mathcal{X}_r^{510})|)_{i=1,\ldots,24}$.

The Lazy Witness algorithm was adopted to produce simplicial complexes of at most degree one and the filtration occurred on the radius interval $[0,1)$. As illustrated in Figure 53 (Appendix E), the intervals $[0,0.0001)$ and $[0.0777,0.1)$ guarantee $G(\mathcal{X}_r) = 0$. However, the interval $[0,0.0001)$ only provides such guarantee as every simplicial complex in X^{510} and $X^{1,000}$ are represented solely by their corresponding landmarks. Therefore, the relevant radii are in

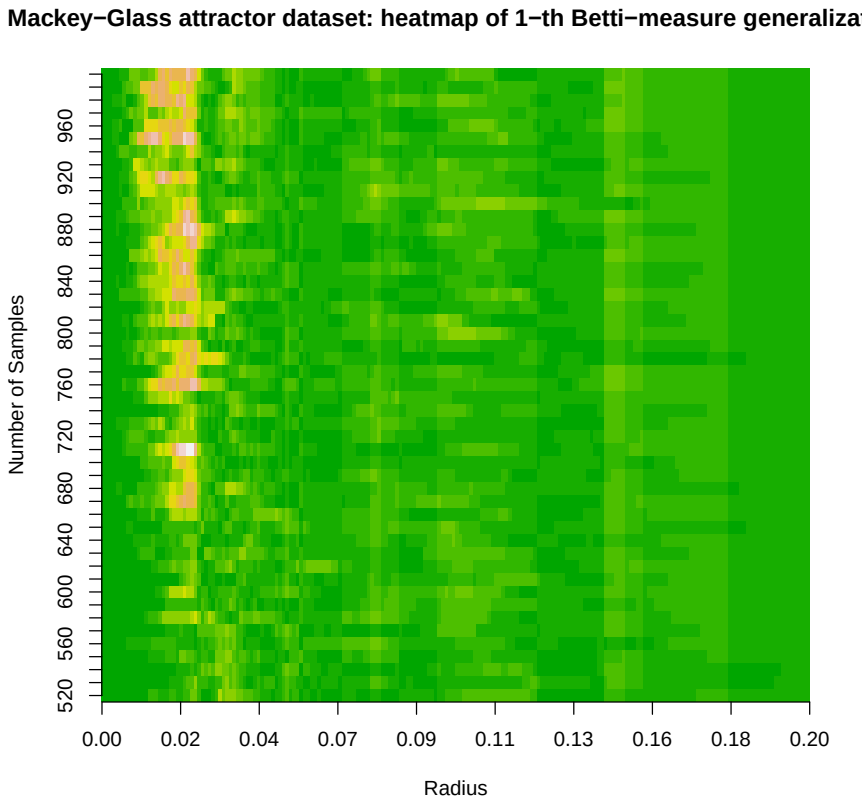


Figure 36 – The heatmap generated from the Mackey-Glass attractor formed by values of the generalization measure $G_1(\mathcal{X}_r)$ along the insertion of samples and the increase of radius over the filtration.

interval $[0.0777, 0.1)$ which identifies the single connected component of the Mackey-Glass attractor.

Considering one-dimensional homology groups, as illustrated in Figure 54 (Appendix E), our approach identifies that the intervals $[0, 0.049), [0.051, 0.0519), [0.0680, 0.0709)$, $[0.1120, 0.1229), [0.2021, 0.3070), [0.3281, 1)$ present the 1-th Betti measure generalization for the Mackey-Glass attractor, being the last three intervals the relevant ones once they are associated with the production of only one connected component. Such intervals identify, respectively, two, one, and no orbit(s) for such attractor. In addition, the heatmaps illustrated in Figures 35 and 36 present the same behavior, as explained for the Lorenz and the Rössler attractors, thus confirming once more the CRD.

5.5 Topological Concept Drift in real data

The following section details an experiment on how the vocabulary used in the abstracts of scientific papers from Machine Learning (ML) and/or Time Series (TS), extracted from Scopus [5], topologically changes from the year $2,000$ to $2,020$. Such changes correspond to

[5] https://www.scopus.com/home.uri

variations between consecutive 0-th consistent homology classes produced on top of word frequencies with respect to an initial dataset $A(2{,}000 \text{ to } 2{,}003)$ composed of documents from $2,000$ to $2,003$. Such initial set was employed to produce $\mathcal{X}_{\rho^{\cong},r^{\cong}}(T_0)$ for comparing the 0-th Betti measures such as stated in Section 5.2. Considering the ML papers, such initial dataset counts on $3,080$ abstracts, and the remaining years contain approximately $2,000$ each, except for $2,003$ ($1,689$ abstracts) and $2,020$ (889). Complementary, TS papers relies on around $2,000$ abstracts for each year but for $2,020$ (381).

5.5.1 Documents semantic changes

The main goal of this experiment was to study changes on word vocabularies from ML and TS-related paper abstracts in order to verify their correspondences along time. The similarity measure is computed from the number of connected components that the initial set $A(2{,}000 \text{ to } 2{,}003)$ has in comparison with the following years $(A(2{,}004), A(2{,}005), \ldots, A(2{,}020))$, i.e., by assessing their differences in terms of 0-th Betti measures.

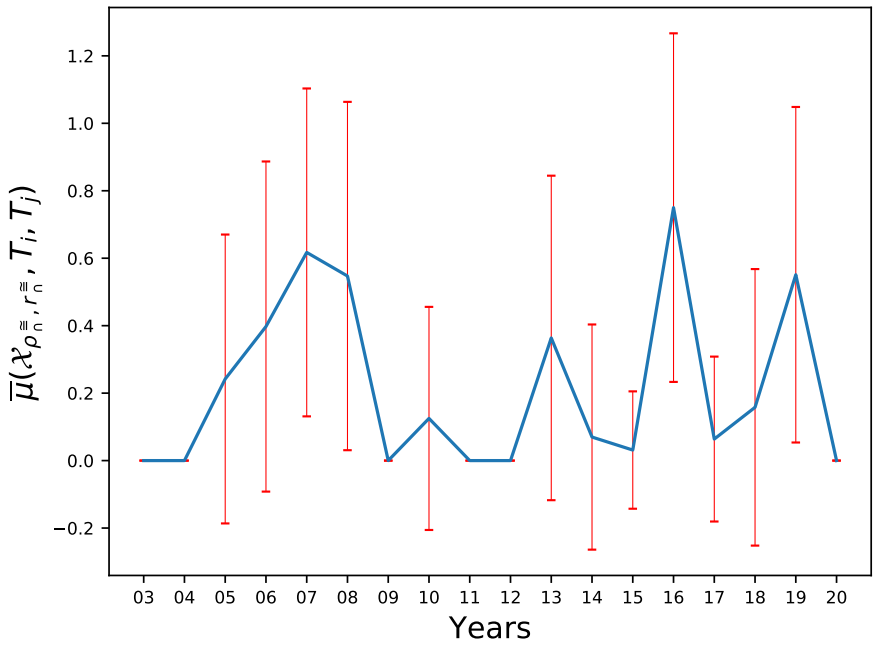


Figure 37 – Average of variations, with respect to DBSCAN parameters ρ and r, on the 0-th homology generalizations of $\mathcal{X}_{\rho^{\cong}_{\cap},r^{\cong}_{\cap}}(0)$ and $\mathcal{X}_{\rho^{\cong}_{\cap},r^{\cong}_{\cap}}(T_i)$ with the respective standard deviations represented by red vertical lines.

In order to produce such a dataset, stop-words, numbers, non-alphabetic symbols and two letter words were removed from all abstracts and the remaining words were stemmed (PORTER, 1980). Next, matrices $M_{A(2{,}000 \text{ to } 2{,}003)}$ of word frequencies were calculated for the initial collection $A(2{,}000 \text{ to } 2{,}003)$ and for all others $A(2{,}004), A(2{,}005), \ldots, A(2{,}020)$, considering set $A(\text{year})$ contains abstracts (instances) whose words are the attributes. Non-negative Matrix

Factorization (NMF) was then applied on $M_{A(\text{year})}$ to reduce data dimensionality to three while clustering data (DING; LI; PENG, 2008).

Non-negative Matrix Factorization finds matrices $H \geq 0$ and $W \geq 0$, such that $WH \approx M$ being M a non-negative matrix. In order to preserve such similarity between WH and M, the generalized Kullback-Leibler divergence:

$$D_{\text{KL}}(M|M') = \sum_{i,j}\left(M_{ij}log\frac{M_{ij}}{M'_{ij}}\right)$$

is employed as the objective function to find H and W, such that the optimal $\tilde{H}$ and $\tilde{W}$ are defined as:

$$\begin{aligned}\tilde{H} &= \arg\min_H D_{KL}(M_{A(2{,}000 \text{ to } 2{,}003)}\|WH)\\ \tilde{W} &= \arg\min_W D_{KL}(M_{A(2{,}000 \text{ to } 2{,}003)}\|WH).\end{aligned}$$

Therefore, as $\tilde{W}$ contains the clustering information (DING; LI; PENG, 2008), we transformed each $M(A(\text{year}))$ by adopting the trimmed generalized inverse $\tilde{H}_3^{-1}$ of $\tilde{H}$ such that $\dim(\tilde{H}_3^{-1}) = 3$, i.e., the transformed dataset $X(\text{year})$ is equal to $W(\text{year})\tilde{H}_3^{-1}$ (being generated from $A(\text{year})$), such as illustrated in Figure 45 (Appendix D).

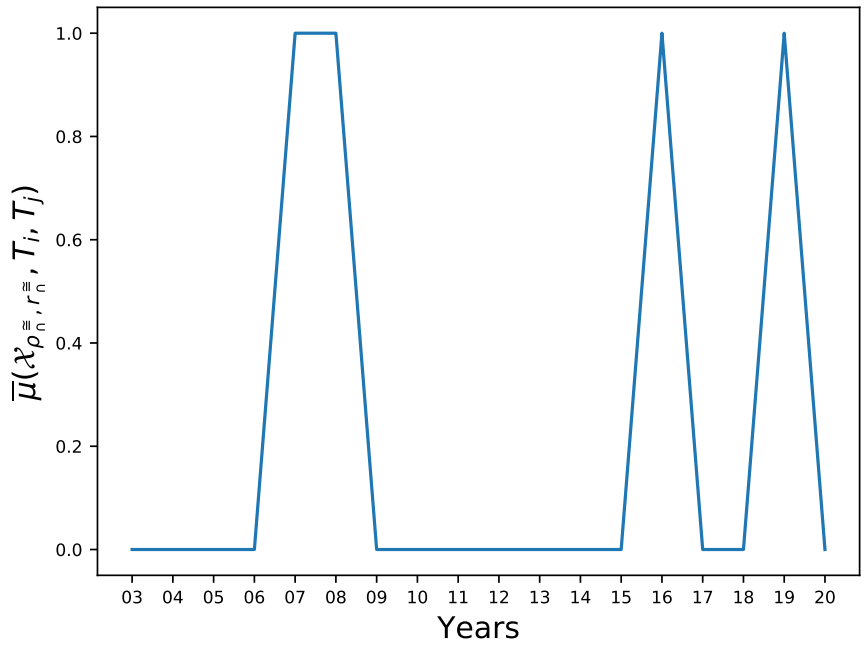


Figure 38 – Median of variations, with respect to DBSCAN parameters ρ and r, on the 0-th homology generalizations of $\mathcal{X}_{\rho_\cap^{\cong}, r_\cap^{\cong}}(0)$ and $\mathcal{X}_{\rho_\cap^{\cong}, r_\cap^{\cong}}(T_i)$.

After obtaining each dataset $X(\text{year})$, the DBSCAN algorithm was adopted to find their connected components, then the calculated 0-th homology generalizations were compared to those of $X(2{,}000 \text{ to } 2{,}003)$, such as discussed in Section 5.2. By analyzing the average differences (w.r.t. to parameters $\rho_\cap^{\cong}$ and $r_\cap^{\cong}$) between the 0-th homology generalization of $\mathcal{X}_{\rho_\cap^{\cong}, r_\cap^{\cong}}(0)$ and

$\mathcal{X}_{\rho^{\cong}_{\cap}, r^{\cong}_{\cap}}(T_i)$, as illustrated in Figure 37, from the years $2,004$ to $2,007$ the overall changes in 0-th homology classes increase in relation to the initial dataset, while decreasing from $2,007$ to $2,009$. Along $2,010$, $2,013$, $2,016$ and $2,019$, topological changes were also identified with $2,010$ being associated with the minimum variation and $2,016$ with the maximum. In the same form, if the median is considered, years $2,007$, $2,008$, $2,016$ an $2,019$ present TCD as illustrated in Figure 38. Such detection can allow the study of unexpected changes in a system such as the vocabulary of a knowledge area or social group.

In fact, performing a deeper study over such dataset, by adopting clusters to analyze the most frequent words present in the abstracts of $2,004$, $2,007$, $2,016$, and $2,019$ as illustrated in Figures 56, 57 and 58, we conclude that the aforementioned TCDs were given not specifically by some approximation of the vocabulary employed in ML and TS-related abstracts. Instead, there is a significant number of papers related with health sciences which consider ML and TS techniques in their approaches as there are word clusters in which terms such as “surgery” and ’ ‘disease” are noticeable present. Hence, the associated TCDs represent the semantic approximation between health sciences and ML/TS papers.

5.6 Final considerations

This chapter experimentally demonstrate the existence of Coarse-Refinement Dilemma (CRD) in DC and HC by illustrating the divergence of the generalization measure (Equation 4.4) along data inclusions and topological refinement. Such methodology allows the choice of a collection of clusters which present p-homology consistence for any pre-defined value of p. Such clusters are adequate to represent the data topology as long as the trivial one, produced by over-coarsening, is disregarded in the model. In this sense, the presence of CRD was demonstrated in the clustering of toy datasets and the Lorenz, Rössler and Mackey-Glass dynamical system attractors. For instance, the evaluation of the stability of clusters leads to an adequate identification of topological features as the orbits of such attractors.

In sequence, assuming that data are collected from i.i.d. distribution, if such model presents only stable clusters, variations in their corresponding topological spaces represent unexpected changes in the underlying and unknown topological distribution or in the associated probability distribution, i.e., the system from which data is acquired presented some anomalous perturbations, which we defined as Topological Concept Drift (TCD). In this context, this chapter also presented experiments to detect the TCD in a synthetic scenario, as discussed in Section 5.3.3, and in terms of document semantic changes in Section 5.5. In the latter, we identified the TCD produced by the vocabulary approximation between health sciences-related papers and the time series and machine learning-related ones.

CHAPTER

6

CONCLUDING REMARKS

The Data Clustering (DC) and the Hierarchical Clustering (HC) problems consist in finding partitions that adequately represent the structural space similarities of datasets. In addition to the efforts to formalize DC and HC settings, Carlsson and Mémoli (2010) propose a statistical approach, based on ultrametric divergences between a model built from some dataset and a reference one, to ensure unsupervised learning stability, at first in terms of data order permutations and, next, while considering data distributions. Besides their relevant contributions, such study is restricted to deal with data endowed in the complete metric space what has motivated us, in this PhD thesis, to extend the clustering problem to more general spaces, most precisely the topological one, in order to redefine the DC and the HC problems. Such problems are typically formalized from metric spaces even when other spaces are more adequate, e.g., when data corresponds to network structures, geometrical forms, graphs, among others; therefore this first contribution provides the ability to model structural similarities without assuming a metric nature but instead allowing a less restrictive problem space representation.

From such foundation, we proceeded with the second contribution by defining over-refined and over-coarsed topological spaces endowing topologies that, respectively: (i) are refined enough to devise inconsistent results when new data is sampled from the same probability distribution; and (ii) are coarsed enough to always produce consistent features while disregarding any data detail (i.e., they are always topologically equivalent to the S^d sphere, as discussed in Section 4.2). Consequently, the existence of such topologies imposes the conditions for which consistency is ensured. In this sense, there is a trade-off in choosing an appropriate refinement for a clustering model which we named Coarse-Refinement Dilemma (CRD, see Chapter 4).

In order to study the CRD, we modeled the HC problem as a bifiltration (CARLSSON; ZOMORODIAN, 2009), i.e., bidirectional topological inclusions, regarding HC hierarchies while inserting data samples. Note the DC model is associated with a particular filtration contained in such structure that corresponds to a specific HC level. This framework comprises our third contribution thus allowing the employment of persistent homology in order to analyze how

topological features change not only along HC levels but also throughout the inclusion of new data. Changes occurring along such inclusions are modeled as martingales in order to permit their probabilistic representations from which we obtained a new consistency result derived from Azuma (1967)'s inequality. As a practical outcome, clustering models can be compared and appropriately refined in order to represent structural space similarities while leading to stable results.

There are other subsequent contributions derived from our consistency result, to mention: (i) fourth contribution – the definition of lower and upper bounds for Carlsson and Mémoli (2010)'s metric space consistency; (ii) fifth contribution – the confrontation of Kleinberg (2002)'s richness axiom once it leads to non-consistent clustering results, i.e., such axiom allows the production of over-refined and over-coarse partitions thus not being, respectively, stable and representative. This fact justifies the adoption of the 0-th homology consistency as a sufficient property regarding scale-invariance, partitions consistency, and locality as these are sustained by isomorphisms of H_0; and, finally (iii) sixth contribution – the definition of Topological Concept Drift (TCD) from changes in topological features provided data inclusions unrelated to over-refined topological spaces.

Experiments were performed to explore the CRD over five datasets: (i) the torus; (ii) the crescent moon; (iii) the Lorenz attractor; (iv) the Rössler attractor and, finally, (v) the Mackey-Glass attractor. We confirmed the usefulness of CRD to characterize orbits of dynamical systems attractors by defining stable p-th homology classes that are more likely to correspond to meaningful features. Opening room for future work, we noticed the need for adapting the Lazy Witness algorithm which eventually found an incoherent number of connected components for the crescent moon dataset due to its particular type of noise.

Moreover, we evaluated TCD over consistent clusterings produced from a synthetic (three moving Gaussian distributions) and a real-world dataset formed by paper abstracts extracted from Scopus [1]. Results allowed to depict the feasibility in detecting TCD but there is room for additional work to reduce the computational processing demanded when substantial data is taken into account.

All results expressed in this PhD thesis were condensed into two scientific journal papers in which the first encloses initial aspects published in 2019 in the Expert Systems With Applications, number 117, pages 90-102 under the title "On learning guarantees to unsupervised concept drift detection on data streams" with the cooperation of Rodrigo F. de Mello, Carlos H. Grossi Ferreira, and Albert Bifet. The second paper contains all the theoretical and experimental contributions detailed along most of this PhD thesis, which is still under review in the journal Expert Systems With Applications.

Finally, this PhD thesis still motivates future work on: (i) the study of how the sparsity in

[1] Scopus web site – <https://www.scopus.com/home.uri>

p-th Betti measures, associated with data insertions, impacts consistency; (ii) the formalization of the asymptotic behavior of the p-homology generalization as neighborhood topologies are refined or coarsed; (iii) the demonstration of a relation between 0-th homology consistency and the theoretical results of the Statistical Learning Theory (SLT), therefore associating the Coarse-Refinement Dilemma (CRD) with the Bias-Variance Dilemma in an attempt to closely connect unsupervised and supervised learning; and, finally, (iv) the application of our TCD detection method in additional real-world problems from the industry.

BIBLIOGRAPHY

ACKERMAN, M.; BEN-DAVID, S.; LOKER, D. Towards property-based classification of clustering paradigms. In: LAFFERTY, J. D.; WILLIAMS, C. K. I.; SHAWE-TAYLOR, J.; ZEMEL, R. S.; CULOTTA, A. (Ed.). **Advances in neural information processing systems 23**. Curran Associates, Inc., 2010. p. 10–18. Available: <http://papers.nips.cc/paper/4101-towards-property-based-classification-of-clustering-paradigms.pdf>. Citations on pages 23, 26, 27, 63, 65, and 69.

ALLIGOOD, K.; SAUER, T.; YORKE, J. **Chaos: an introduction to dynamical systems**. Springer New York, 2000. (Textbooks in mathematical sciences). ISBN 9780387946771. Available: <https://books.google.com.br/books?id=48YHnbHGZAgC>. Citation on page 93.

ALTMAN, N. S. An introduction to kernel and nearest-neighbor nonparametric regression. **American statistician**, American Statistical Association, v. 46, n. 3, p. 175–185, Aug. 1992. ISSN 0003-1305. Available: <https://www.jstor.org/stable/2685209?seq=1>. Citation on page 40.

ARROW, K. J. **Social choice and individual values**. Yale University Press, 1951. ISBN 9780300179316. Available: <http://www.jstor.org/stable/j.ctt1nqb90>. Citation on page 64.

AZUMA, K. Weighted sums of certain dependent random variables. **Tohoku mathematical journal**, Tohoku University, Mathematical Institute, v. 19, n. 3, p. 357–367, 1967. ISSN 2186-585X. Available: <https://doi.org/10.2748/tmj/1178243286>. Citations on pages 29, 57, 81, 104, and 121.

BEN-DAVID, S.; ACKERMAN, M. Measures of clustering quality: a working set of axioms for clustering. In: KOLLER, D.; SCHUURMANS, D.; BENGIO, Y.; BOTTOU, L. (Ed.). **Advances in neural information processing systems 21**. Curran Associates, Inc., 2009. p. 121–128. Available: <http://papers.nips.cc/paper/3491-measures-of-clustering-quality-a-working-set-of-axioms-for-clustering.pdf>. Citations on pages 23, 63, and 65.

BEN-HUR, A.; HORN, D.; SIEGELMANN, H. T.; VAPNIK, V. Support vector clustering. **J. Mach. Learn. Res.**, JMLR.org, v. 2, p. 125–137, Mar. 2002. ISSN 1532-4435. Citation on page 63.

BOUSQUET, O.; ELISSEEFF, A. Stability and generalization. **Journal of machine learning research**, JMLR, v. 2, p. 499–526, Mar. 2002. ISSN 1532-4435. Available: <http://www.jmlr.org/papers/volume2/bousquet02a/bousquet02a.pdf>. Citations on pages 23 and 26.

BURRIS, S.; SANKAPPANAVAR, H. P. **A course in universal algebra**. Springer-Verlag New York, 1981. ISSN 0072-5285. Available: <http://www.math.uwaterloo.ca/~snburris/htdocs/ualg.html>. Citation on page 55.

CAMPELLO, R. J. G. B.; MOULAVI, D.; SANDER, J. Density-based clustering based on hierarchical density estimates. In: PEI, J.; TSENG, V. S.; CAO, L.; MOTODA, H.; XU, G.

(Ed.). **Advances in knowledge discovery and data mining**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013. p. 160–172. ISBN 978-3-642-37456-2. Available: <https://link.springer.com/chapter/10.1007/978-3-642-37456-2_14>. Citations on pages 30 and 41.

CARLSSON, G. Topology and data. **Bulletin of the American Mathematical Society**, American Mathematical Society, v. 46, n. 2, p. 255–308, Jan. 2009. ISSN 0273-0979. Available: <https://www.ams.org/journals/bull/2009-46-02/S0273-0979-09-01249-X/>. Citations on pages 15, 28, 68, 76, 88, and 89.

CARLSSON, G.; MÉMOLI, F. Characterization, stability and convergence of hierarchical clustering methods. **Journal of machine learning research**, JMLR, v. 11, p. 1425–1470, 2010. ISSN 1533-7928. Available: <http://www.jmlr.org/papers/volume11/carlsson10a/carlsson10a.pdf>. Citations on pages 13, 14, 23, 26, 27, 31, 42, 52, 63, 66, 67, 68, 71, 72, 83, 84, 85, 86, 103, 104, 117, and 118.

CARLSSON, G.; ZOMORODIAN, A. The theory of multidimensional persistence. **Discrete & computational geometry**, Springer, v. 42, n. 1, p. 71–93, Jul 2009. ISSN 1432-0444. Available: <https://link.springer.com/article/10.1007/s00454-009-9176-0>. Citations on pages 13, 14, 28, 30, 54, 79, and 103.

CHAZAL, F.; COHEN-STEINER, D.; GLISSE, M.; GUIBAS, L. J.; OUDOT, S. Y. Proximity of persistence modules and their diagrams. In: **Proceedings of the Twenty-fifth Annual Symposium on Computational Geometry**. New York, NY, USA: ACM, 2009. (SCG '09), p. 237–246. ISBN 978-1-60558-501-7. Available: <https://dl.acm.org/doi/10.1145/1542362.1542407>. Citations on pages 29 and 68.

CHERNOFF, H. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. **Annals of mathematical statistics**, Institute of Mathematical Statistics, v. 23, p. 493–509, 1952. ISSN 0003-4851. Available: <https://projecteuclid.org/download/pdf_1/euclid.aoms/1177729330>. Citations on pages 57 and 59.

COHEN-STEINER, D.; EDELSBRUNNER, H.; HARER, J. Stability of persistence diagrams. **Discrete & computational geometry**, Springer, v. 37, n. 1, p. 103–120, Jan 2007. ISSN 1432-0444. Available: <https://link.springer.com/article/10.1007/s00454-006-1276-5>. Citations on pages 29 and 68.

CORTES, C.; VAPNIK, V. Support-Vector Networks. In: **Machine Learning**. [S.l.: s.n.], 1995. p. 273–297. Citation on page 63.

DEFAYS, D. An efficient algorithm for a complete link method. **The computer journal**, Oxford University Press, v. 20, n. 4, p. 364–366, 1977. ISSN 1460-2067. Available: <https://academic.oup.com/comjnl/article/20/4/364/393966>. Citation on page 41.

DING, C.; LI, T.; PENG, W. On the equivalence between Non-Negative Matrix Factorization and probabilistic Latent Semantic Indexing. **Computational statistics & data analysis**, Elsevier Science Publishers B. V., NLD, v. 52, n. 8, p. 3913–3927, Apr. 2008. ISSN 0167-9473. Available: <https://doi.org/10.1016/j.csda.2008.01.011>. Citation on page 101.

EDELSBRUNNER, H.; HARER, J. Persistent homology -— a survey. In: **Surveys on discrete and computational geometry**. American Mathematical Society, 2008. v. 453, p. 257. ISBN 978-0-8218-8132-3. Available: <https://bookstore.ams.org/conm-453>. Citations on pages 51 and 68.

ESTER, M.; KRIEGEL, H.-P.; SANDER, J.; XU, X. A density-based algorithm for discovering clusters in large spatial databases with noise. In: **Proceedings of the Second International Conference on Knowledge Discovery and Data Mining**. AAAI Press, 1996. (KDD'96), p. 226–231. ISBN 978-1-57735-004-0. Available: <https://dl.acm.org/doi/10.5555/3001460.3001507>. Citations on pages 30 and 40.

FERENTINOS, K. On Tchebycheff's type inequalities. **Trabajos de estadística e investigación operativa**, Instituto de Investigación Operativa y Estadística., v. 33, n. 1, p. 125–132, 1982. ISSN 0041-0241. Available: <http://eudml.org/doc/40674>. Citation on page 59.

FORMAGGIA, L.; SALERI, F.; VENEZIANI, A. Some fundamental tools. In: **Solving numerical PDEs: problems, applications, exercises**. Milano: Springer Milan, 2012. chap. 1, p. 3–15. ISBN 978-88-470-2412-0. Available: <https://www.springer.com/gp/book/9788847024113>. Citations on pages 13 and 36.

FURIHATA, M.; TAKEUCHI, T. Gleason grading. In: SCHWAB, M. (Ed.). **Encyclopedia of cancer**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011. p. 1548–1551. ISBN 978-3-642-16483-5. Available: <https://link.springer.com/referenceworkentry/10.1007%2F978-3-642-16483-5_2415>. Citation on page 69.

GHOSH, B. K. Probability inequalities related to Markov's theorem. **The american statistician**, Taylor & Francis, v. 56, n. 3, p. 186–190, 2002. ISSN 1537-2731. Available: <https://www.tandfonline.com/doi/abs/10.1198/000313002119>. Citation on page 57.

GROMOV, M.; LAFONTAINE, J.; PANSU, P. **Structures métriques pour les variétés riemanniennes**. CEDIC/Fernand Nathan, 1981. (Textes mathématiques). ISBN 9782712407148. Available: <https://books.google.com.br/books?id=TxN0QgAACAAJ>. Citation on page 26.

____. **Metric structures for Riemannian and non-Riemannian spaces**. Birkhäuser, 1999. (Progress in mathematics - Birkhäuser). ISBN 9780817638986. Available: <https://books.google.com.br/books?id=kkVAAQAAIAAJ>. Citation on page 67.

GUARASCIO, M.; MANCO, G.; RITACCO, E. Deep learning. In: RANGANATHAN, S.; GRIBSKOV, M.; NAKAI, K.; SCHÖNBACH, C. (Ed.). **Encyclopedia of bioinformatics and computational biology**. Oxford: Academic Press, 2019. p. 634 – 647. ISBN 978-0-12-811432-2. Available: <http://www.sciencedirect.com/science/article/pii/B978012809633820352X>. Citations on pages 13 and 38.

HAR-PELED, S.; JONES, M. On separating points by lines. In: CZUMAJ, A. (Ed.). **Proceedings on 29th Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2018**. Association for Computing Machinery, 2018. p. 918–932. ISBN 0-89871-513-X. Available: <https://experts.illinois.edu/en/publications/on-separating-points-by-lines>. Citation on page 25.

HATCHER, A. **Algebraic topology**. Cambridge: Cambridge University Press, 2000. ISBN 978-0521795401. Available: <https://pi.math.cornell.edu/~hatcher/AT/AT.pdf>. Citations on pages 13, 14, 28, 45, 46, 47, 48, 49, 51, and 75.

HOEFFDING, W. Probability inequalities for sums of bounded random variables. **Journal of the American Statistical Association**, American Statistical Association, v. 58, n. 301, p. 13–30, March 1963. ISSN 1537-274X. Available: <http://www.jstor.org/stable/2282952?> Citations on pages 57, 59, and 60.

KAUFMAN, L.; ROUSSEEUW, P. **Clustering by means of medoids**. Faculty of Mathematics and Informatics, 1987. (Delft University of Technology : reports of the Faculty of Technical Mathematics and Informatics). Available: <https://books.google.com.br/books?id=HK-4GwAACAAJ>. Citation on page 83.

KHASAWNEH, F. A.; MUNCH, E. Chatter detection in turning using persistent homology. **Mechanical systems and signal processing**, Elsevier, v. 70-71, p. 527 – 541, 2016. ISSN 0888-3270. Available: <http://www.sciencedirect.com/science/article/pii/S0888327015004598>. Citations on pages 13 and 29.

KIMURA, M.; OBAYASHI, I.; TAKEICHI, Y.; MURAO, R.; HIRAOKA, Y. Non-empirical identification of trigger sites in heterogeneous processes using persistent homology. **Scientific reports**, Nature, v. 8, n. 1, p. 3553, February 2018. ISSN 2045-2322. Available: <https://europepmc.org/articles/PMC5824834>. Citation on page 69.

KLEIN, S. T. On the connection between hamming codes, heapsort and other methods. **Information processing letters**, Elsevier, v. 113, n. 17, p. 617 – 620, 2013. ISSN 0020-0190. Available: <http://www.sciencedirect.com/science/article/pii/S0020019013001555>. Citations on pages 13 and 35.

KLEINBERG, J. An impossibility theorem for clustering. In: **Proceedings of the 15th International Conference on Neural Information Processing Systems**. Cambridge, MA, USA: MIT Press, 2002. (NIPS'02), p. 463–470. Available: <https://www.cs.cornell.edu/home/kleinber/nips15.pdf>. Citations on pages 23, 26, 27, 31, 63, 64, 69, 72, 83, 86, 104, and 117.

KOHONEN, T. Self-organized formation of topologically correct feature maps. **Biological cybernetics**, Springer-Verlag Ltd. Germany, v. 43, n. 1, p. 59–69, Jan 1982. ISSN 1432-0770. Available: <https://link.springer.com/article/10.1007/BF00337288>. Citation on page 83.

KOTZ, S.; IBRAGIMOV, I.; HAS'MINSKII, R. **Statistical estimation: asymptotic theory**. Springer New York, 2013. (Stochastic modelling and applied probability). ISBN 9781489900272. Available: <https://books.google.com.br/books?id=D4zSBwAAQBAJ>. Citation on page 59.

LAWSON, P.; SHOLL, A. B.; BROWN, J. Q.; FASY, B. T.; WENK, C. Persistent homology for the quantitative evaluation of architectural features in prostate cancer histology. **Scientific reports**, Nature, v. 9, n. 1, Dec 2019. ISSN 2045-2322. Available: <https://www.nature.com/articles/s41598-018-36798-y>. Citations on pages 68 and 69.

LIBBRECHT, M.; NOBLE, W. Machine learning applications in genetics and genomics. **Nature reviews. Genetics**, Nature, v. 16, p. 321–332, 05 2015. Available: <https://www.nature.com/articles/nrg3920>. Citation on page 23.

LUXBURG, U. von; SCHÖLKOPF, B. Statistical learning theory: models, concepts, and results. In: GABBAY, D. M.; HARTMANN, S.; WOODS, J. (Ed.). **Inductive logic**. North-Holland, 2011, (Handbook of the history of logic, v. 10). p. 651 – 706. Available: <http://www.sciencedirect.com/science/article/pii/B9780444529367500161>. Citations on pages 14, 23, 24, 38, 57, and 73.

MACKEY, M.; GLASS, L. Oscillation and chaos in physiological control systems. **Science**, American Association for the Advancement of Science, v. 197, n. 4300, p. 287–289, 1977. ISSN 0036-8075. Available: <https://science.sciencemag.org/content/197/4300/287>. Citation on page 97.

MACQUEEN, J. Some methods for classification and analysis of multivariate observations. In: **Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, volume 1: Statistics**. Berkeley, Calif.: University of California Press, 1967. p. 281–297. ISSN 0097-0433. Available: <https://projecteuclid.org/euclid.bsmsp/1200512992>. Citation on page 83.

MATOUSEK, J. **Lectures on discrete geometry**. Berlin, Heidelberg: Springer-Verlag, 2002. ISBN 0387953744. Available: <https://www.springer.com/gp/book/9780387953731>. Citations on pages 13 and 35.

MCDIARMID, C. On the method of bounded differences. In: SIEMONS, J. (Ed.). **Surveys in combinatorics, 1989: invited papers at the Twelfth British Combinatorial Conference**. Cambridge University Press, 1989, (London Mathematical Society lecture note series). p. 148–188. Available: <https://www.cambridge.org/core/books/surveys-in-combinatorics-1989/on-the-method-of-bounded-differences/AABA597B562BDA7D89C6077E302694FB>. Citation on page 26.

MELLO, R.; PONTI, M. **Machine learning: a practical approach on the Statistical Learning Theory**. Springer International Publishing, 2018. ISBN 978-3-319-94988-8. Available: <https://www.springer.com/gp/book/9783319949888>. Citations on pages 13, 23, 24, 25, 38, and 57.

MELLO, R. F. de; MANAPRAGADA, C.; BIFET, A. Measuring the shattering coefficient of decision tree models. **Expert systems with applications**, Elsevier, v. 137, p. 443 – 452, 2019. ISSN 0957-4174. Available: <http://www.sciencedirect.com/science/article/pii/S0957417419304919>. Citation on page 25.

MELLO, R. F. de; VAZ, Y.; GROSSI, C. H.; BIFET, A. On learning guarantees to unsupervised concept drift detection on data streams. **Expert systems with applications**, Elsevier, v. 117, p. 90 – 102, 2019. ISSN 0957-4174. Available: <http://www.sciencedirect.com/science/article/pii/S0957417418305682>. Citations on pages 14, 58, and 59.

MUKHERJEE, S.; NIYOGI, P.; POGGIO, T.; RIFKIN, R. Learning theory: stability is sufficient for generalization and necessary and sufficient for consistency of empirical risk minimization. **Advances in computational mathematics**, Springer Nature Switzerland AG, v. 25, n. 1, p. 161–193, 2006. ISSN 1572-9044. Available: <http://dx.doi.org/10.1007/s10444-004-7634-z>. Citation on page 23.

MUNKRES, J. **Topology**. Prentice Hall, Incorporated, 2000. (Featured titles for topology series). ISBN 9780131816299. Available: <https://books.google.com.br/books?id=XjoZAQAAIAAJ>. Citations on pages 27 and 28.

PATEL, J.; SHAH, S.; THAKKAR, P.; KOTECHA, K. Predicting stock market index using fusion of machine learning techniques. **Expert systems with applications**, Elsevier, v. 42, n. 4, p. 2162 – 2172, 2015. ISSN 0957-4174. Available: <http://www.sciencedirect.com/science/article/pii/S0957417414006551>. Citation on page 23.

PORTER, M. F. An algorithm for suffix stripping. **Program: electronic library and information systems**, v. 14, n. 3, p. 130–137, 1980. ISSN 0033-0337. Available: <http://dblp.uni-trier.de/db/journals/program/program14.html#Porter80>. Citation on page 100.

RAHMAWATI, D.; HUANG, Y. Using c-support vector classification to forecast dengue fever epidemics in taiwan. In: **2016 International Conference on System Science and Engineering (ICSSE)**. IEEE, 2016. p. 1–4. ISSN 2325-0925. Available: <https://ieeexplore.ieee.org/document/7551552>. Citation on page 23.

ROSENBLATT, F. **The Perceptron, a perceiving and recognizing automaton Project Para**. Cornell Aeronautical Laboratory, 1957. (Report: Cornell Aeronautical Laboratory). Report No. 85-460-1. Available: <https://blogs.umass.edu/brain-wars/files/2016/03/rosenblatt-1957.pdf>. Citation on page 38.

ROSENBLATT, F.; BUFFALO, N. Y. Cornell Aeronautical Lab inc. **Principles of neurodynamics. Perceptrons and the theory of brain mechanisms**. Arlington Hall Station, Arlington 12, Virginia, United States: Armed Services Technical Information Center, 1961. Report No. 1196-G-8. Available: <https://apps.dtic.mil/dtic/tr/fulltext/u2/256582.pdf>. Citation on page 38.

RUMELHART, D. E.; HINTON, G. E.; WILLIAMS, R. J. Learning internal representations by error propagation. In: RUMELHART, D. E.; MCCLELLAND, J. L. (Ed.). **Parallel distributed processing: explorations in the microstructure of cognition, volume 1: Foundations**. Cambridge, MA: MIT Press, 1986. p. 318–362. ISBN 978-0262181235. Available: <https://web.stanford.edu/class/psych209a/ReadingsByDate/02_06/PDPVolIChapter8.pdf>. Citation on page 38.

SAJDA, P. Machine learning for detection and diagnosis of disease. **Annual review of biomedical engineering**, Annual Reviews, v. 8, n. 1, p. 537–565, 2006. ISSN 1545-4274. Available: <https://www.ncbi.nlm.nih.gov/pubmed/16834566>. Citation on page 23.

SCHOLKOPF, B.; SMOLA, A. J. **Learning with kernels: Support Vector Machines, regularization, optimization, and beyond**. Cambridge, MA, USA: MIT Press, 2001. ISBN 0262194759. Available: <https://ieeexplore.ieee.org/book/6267332>. Citation on page 38.

SIBSON, R. SLINK: an optimally efficient algorithm for the single-link cluster method. **The computer journal**, Oxford University Press, v. 16, n. 1, p. 30–34, jan 1973. ISSN 0010-4620. Available: <https://academic.oup.com/comjnl/article/16/1/30/434805>. Citation on page 41.

SILVA, V. D.; CARLSSON, G. Topological estimation using witness complexes. In: **Proceedings of the First Eurographics Conference on Point-Based Graphics**. Aire-la-Ville, Switzerland, Switzerland: Eurographics Association, 2004. (SPBG'04), p. 157–166. ISBN 3-905673-09-6. Available: <https://dl.acm.org/doi/10.5555/2386332.2386359>. Citations on pages 15, 88, and 89.

STEINBRUCH, A.; WINTERLE, P. **Álgebra linear**. Pearson Makron Books, 1987. ISBN 9780074504123. Available: <https://books.google.com.br/books?id=q36CPgAACAAJ>. Citations on pages 13 and 34.

STEINHAUS, H. Sur la division des corps matériels en parties. In: **Bulletin de l'Académie Polonaise des Sciences**. Polish Academy of Sciences, 1956. IV, n. 12, p. 801–804. Available: <http://www.laurent-duval.eu/Documents/Steinhaus_H_1956_j-bull-acad-polon-sci_division_cmp-k-means.pdf>. Citation on page 83.

TAKIYAMA, A.; TERAMOTO, T.; SUZUKI, H.; YAMASHIRO, K.; TANAKA, S. Persistent homology index as a robust quantitative measure of immunohistochemical scoring. **Scientific reports**, Nature, v. 7, 12 2017. ISSN 2045-2322. Available: <https://www.nature.com/articles/s41598-017-14392-y>. Citation on page 69.

VAPNIK, V. N. **The nature of Statistical Learning Theory**. New York, NY, USA: Springer-Verlag New York, Inc., 1995. ISBN 0-387-94559-8. Available: <https://www.springer.com/gp/book/9780387987804>. Citations on pages 23, 24, 25, 38, and 57.

VIETORIS, L. Über den höheren Zusammenhang kompakter Räume und eine Klasse von zusammenhangstreuen Abbildungen. **Mathematische annalen**, Springer-Verlag Ltd. Germany, v. 97, p. 454–472, 1927. ISSN 1432-1807. Available: <http://eudml.org/doc/182647>. Citations on pages 88, 89, and 90.

VILLE, J. **Étude critique de la notion de collectif**. Gauthier-Villars, 1939. (Monographies des probabilités). Available: <http://www.numdam.org/issue/THESE_1939__218__1_0.pdf>. Citations on pages 29 and 81.

WANG, F.; ZHEN, Z.; WANG, B.; MI, Z. Comparative study on KNN and SVM based weather classification models for day ahead short term solar PV power forecasting. **Applied sciences**, MDPI, v. 8, p. 28, 12 2017. ISSN 2076-3417. Available: <https://www.mdpi.com/2076-3417/8/1/28>. Citations on pages 13 and 41.

WANG, P.; JIANG, A.; LIU, X.; SHANG, J.; ZHANG, L. LSTM-based EEG classification in motor imagery tasks. **IEEE transactions on neural systems and rehabilitation engineering**, IEEE, v. 26, n. 11, p. 2086–2095, 2018. ISSN 1558-0210. Available: <https://ieeexplore.ieee.org/document/8496885>. Citation on page 23.

WERBOS, P. J. **Beyond regression: new tools for prediction and analysis in the behavioral sciences**. Phd Thesis (PhD Thesis) — Harvard University, 1974. Available: <https://www.researchgate.net/publication/35657389_Beyond_regression_new_tools_for_prediction_and_analysis_in_the_behavioral_sciences>. Citation on page 38.

Appendix

APPENDIX

A

NOTATION

$x, \ldots, z$ (Lower case Latin letters from x to z) : Variables;
$f, \ldots, h$ (Lower case Latin letters from f to h) : Functions;
αx or αd (Lower case Greek letters associated with a variable x or function d) : Scalar multiplication;
X (Upper case Latin letters) : Sets;
$\mathbb{R}^p, \mathbb{N}^p$: Real and natural sets of dimension p;
Γ : Partition of a set S defined as in (KLEINBERG, 2002);
(X, d) or (X, d_X) : Metric space associated with a set X of elements and a distance function d (when such function is considered to be general) or d_X (when such function is considered to be associated with X);
(Z, τ_Z) : The unknown underlying topological space (considered in the Data Clustering and Hierarchical Clustering problems);
(X, τ_X) : The topological space built on top of a dataset X;
(Ω, τ_Ω) : The unknown universal topological space (in which (Z, τ_Z) is included) associated with a Borel measure;
$\mathcal{N}(x)$: Neighborhood of x;
η : Map $\eta : x \mapsto \mathcal{N}(x)$;
$\mathcal{X}$: Topological space built on top of X by a neighborhood topology;
$\sim$: Equivalence relation;
$\sim_r$: r-relation (CARLSSON; MÉMOLI, 2010);
$\mathcal{N}_i(X)$: A neighborhood topology associated with a level i of a hierarchical clustering model;
η_i : Map $\eta : x \mapsto \mathcal{N}_i(x)$;
i, j : Indices related with persistence along topological coarse-refinement;
t_i : Element of the pre-image of a tame function associated with an index i;
k : An iterator;
Δ^p : Simplex of dimension p;
σ : A singular p-th dimensional simplex (a continuous map from a simplex to a topological space);
∂ : The boundary operator;
C_p : A p-dimensional simplicial complex;
$\mathrm{im}(\cdot)$: Image of a map;
$\ker(\cdot)$: Kernel of a map;
H_p : p-dimensional homology group;
$\mathbb{Z}^p$: Abelian group of p-tuple of integers endowed with a direct sum;
$\mathbb{Q}$: Field of rational numbers;
β_p : p-th Betti number (or measure);
$\mathbb{E}(\cdot)$: Expected value;
$\mathcal{F}(X, \eta)$: Filtration of neighborhood topological spaces associated with the map η and dataset X;
$\mathbf{f}_p^{i,j}$ and $\mathbf{g}_p^{l,q}$: p-dimensional persistence functions associated with, respectively, the coarse-refinement of a topological space and the inclusion of new data;
$\mathcal{X}_i$ and $\mathcal{X}_i^m$: Neighborhood topological spaces associated, respectively, with $\mathcal{N}_i(X)$ and $\mathcal{N}_i(X^m)$;
$\imath_X$ and $\imath_\mathcal{N}$: Inclusions $\mathcal{X}_i^{l<q} \subseteq \mathcal{X}_i^q$ and $\mathcal{X}_i^q \subseteq \mathcal{X}_j^q$ respectively;
l, q : Indices related with the persistence along data inclusions;
$\mathcal{H}$: Collection of abelian groups;
μ : A measure function;
P_μ : Probability function endowed with a measure func-

tion μ;
P_{β_p} : Probability function endowed with a p-th Betti measure;
$\mathcal{O}(\cdot)$: Big-O notation;
$\Omega(\cdot)$: Big-Ω notation;
Ω : Universe set;
M,N : Cardinalities;
(X, u_d) : Ultrametric space;
(X, d_X, μ_X) : Measurable metric space (mm-space);
$\mathcal{S}_{mm}(X)$: Measurable metric space (mm-space) associated with the dataset X;
$B(x, r)$: Open ball of radius r centered in x;
$\text{supp}(\cdot)$: Support of a measure function;
d_H : Hausdorff distance;
d_{GH} : Gromov-Hausdorff distance;
$\mathfrak{L}$: Carlsson and Mémoli (2010)'s adapted single linkage function;
$G_p(\mathcal{X})$: Generalization measure of a topological space $\mathcal{X}_r$ in terms of its p-th homology group;
$G(f)$: Generalization measure of a function f;
$\mathcal{X}$: Set of instances associated with Supervised Machine Learning;
$\mathcal{Y}$: Set of labels associated with Supervised Machine Learning;
$\mathcal{F}$: Space of admissible functions;
$R(f)$: Expected risk of f;
$R_{\text{emp}}(f)$: Empirical risk of f;
$m(\cdot)$: Median;
$\mu(\cdot)$: Mean;
$\overline{\mu}(\cdot)$: Estimated average;
$\rho^{\cong}$: Parameter ρ that assure p-th homology consistency.

APPENDIX

B

DICTIONARY OF TERMS

Stability: Characteristic of the bounded absolute difference between a measure applied over a set of variables (typically random) and its perturbation (given by variable permutations, removals or inclusions);
Consistency: Characteristic of an estimation which approximates its expected value as the number of random variables increases;
***p*-Homology consistency**: Consistency for the estimation of the rank of p-homology groups;
***p*-simplex**: p-dimensional generalization of triangles;
Boundary Operator: Homomorphism applied over p-chains producing $p-1$-chains;
Chain complex: Sequence of p-chains produced by the successive applications of the Boundary Operator;
Voids: "Empty space" inside a topological space;
Holes: "Empty space" which traverses a topological space;
Connected Component: Maximal connected subsets which cannot be partitioned into two disjoint nonempty subsets;
Homology group: Algebraic representation for Topological Space which considers its connected components, voids, and holes;
Homology class: Algebraic representation of cycles in a Homology group;
Over-refinement: Refinement of a topological space which produces unstable p-homology groups;
Over-coarse: Refinement of a topological space which vanishes relevant information such as number of connected components, holes, and voids;
Coarse-refinement Dilemma: Dilemma in choosing an appropriate refinement in order to represent relevant topological features while avoiding instabilities on the p-homology groups;
Dendrogram: Representation of a hierarchical clustering model responsible for mapping a the set of grouped elements into the minimal radius which merged them all together;
Filtration: Successive inclusions of topological spaces;
Multifiltration: Inclusions of topological spaces modified by more than one degree (e.g., radius of neighborhood topologies);
Persistence: Interval in which a homology class persists along a filtration;
Persistence Function: Function which maps a homology group of a included topological space into the homology group of a larger topological space;
Multidimensional Persistence: Persistence calculated on top of a multifiltration; ***p*-th Betti number**: The rank of a homology group;
Topological Data Analysis: Scientific area responsible for studying topological properties of data points.

APPENDIX

C

PROOFS

Proof of Lemma 6 1. Given a topological space $\mathcal{X}_i$ built on top of a dataset X by a clustering algorithm, the one $\mathcal{X}_i^m$ built on top of the perturbation of X, X^m (such that $X^m = X \cup \{x'_1, \ldots, x'_m\}$) with i.i.d. samples), and their corresponding p-th Betti measure $\beta_p(\cdot)$ associated with a martingale or a supermartingale process, by Azuma's Inequality (AZUMA, 1967):

$$P\left(|\beta_p(\mathcal{X}_i^m) - \beta_p(\mathcal{X}_i)| > \varepsilon\right) \leq 2\exp\left(\frac{-\varepsilon^2}{2\sum_{q=1}^{m} c_{q,p}^2}\right). \tag{C.1}$$

If Equation C.2, which calculates the absolute difference between the p-th Betti numbers associated to $\mathcal{X}_i$ and $\mathcal{X}_i^m$, holds, generalization is guaranteed:

$$\forall m,\ |\beta_p(\mathcal{X}_i^m) - \beta_p(\mathcal{X}_i)| = 0. \tag{C.2}$$

Considering Equation 4.3, the sequence $\beta_p(\mathcal{X}_i^q) - \beta_p(\mathcal{X}_i)$ with $q = 1, \ldots, m$ also forms a martingale as:

$$\begin{aligned} \mathbb{E}\left[\beta_p(\mathcal{X}_i^q) - \beta_p(\mathcal{X}_i)\right] &= \mathbb{E}\left[\beta_p(\mathcal{X}_i^q)\right] - \beta_p(\mathcal{X}_i) \\ &= \beta_p(\mathcal{X}_i^{q-1}) - \beta_p(\mathcal{X}_i). \end{aligned}$$

Therefore, in terms of the average p-th Betti measure differences, with $|\beta_p(\mathcal{X}_i^q) - \beta_p(\mathcal{X}_i)| < c_{q,p}$, probabilistic convergence is defined as:

$$P\left(\sum_{q=1}^{m} |\beta_p(\mathcal{X}_i^q) - \beta_p(\mathcal{X}_i)| > m\varepsilon\right) \leq 2\exp\left(\frac{-m^2\varepsilon^2}{2\sum_{q=1}^{m} c_{q,p}^2}\right). \tag{C.3}$$

Proof of Corollary 3 1. Consider the Inequality 4.5 of Lemma 6 which can be written as:

$$P\left(\sum_{q=1}^{m} |\beta_p(\mathcal{X}_i^q) - \beta_p(\mathcal{X}_i)| \leq m\varepsilon\right) > 2\exp\left(\frac{-m^2\varepsilon^2}{2\sum_{q=1}^{m} c_{q,p}^2}\right).$$

By substituting $P\left(\sum_{q=1}^{m} |\beta_p(\mathcal{X}_i^q) - \beta_p(\mathcal{X}_i)| \leq m\varepsilon\right) = \delta$, we have:

$$\delta > 2\exp\left(\frac{-m^2\varepsilon^2}{2\sum_{q=1}^{m} c_{q,p}^2}\right). \tag{C.4}$$

Now, let $\overline{c}_p$ be the maximum value of $(c_{q,p})_{q=1\ldots,m}$, then it follows from Inequality C.4 that:

$$\delta > 2\exp\left(\frac{-m^2\varepsilon^2}{2\sum_{q=1}^{m} c_{q,p}^2}\right) \geq 2\exp\left(\frac{-m^2\varepsilon^2}{2m\overline{c}_p^2}\right), \tag{C.5}$$

and, consequently,

$$\begin{aligned} \delta &> 2\exp\left(\tfrac{-m^2\varepsilon^2}{2m\overline{c}_p^2}\right) \\ \ln\delta &> \ln 2 - \tfrac{m\varepsilon^2}{2\overline{c}_p^2} \\ \ln 2 - \ln\delta &< \tfrac{m\varepsilon^2}{2\overline{c}_p^2} \\ \overline{c}_p &< \sqrt{\tfrac{m\varepsilon^2}{2(\ln 2 - \ln\delta)}}. \end{aligned}$$

Since $\delta = 1$ implies consistency, then:

$$\overline{c}_p < \varepsilon\sqrt{m/2\ln 2} \tag{C.6}$$

guarantees it.

Proof of Theorem 7 1. Defining $\overline{\Delta\beta}_{p,i} = \sum_{q=1}^{m} |\beta_p(\mathcal{X}_i^q) - \beta_p(\mathcal{X}_i)|$:

$$\begin{aligned} P\left(\sup_{i\in\mathbb{R}} \overline{\Delta\beta}_{p,i} > m\varepsilon\right) &= P\left(\bigvee_{i\in\mathbb{R}} (\overline{\Delta\beta}_{p,i}) > m\varepsilon\right) \\ &\leq \sum_{i\in\mathbb{R}} P\left(\overline{\Delta\beta}_{p,i} > m\varepsilon\right) \\ &\leq 2M\exp\left(\tfrac{-m\varepsilon^2}{2\tilde{c}_p^2}\right), \end{aligned}$$

in which M is the number of critical points that a tame function f produces and $\tilde{c}_p = \max_{i\in\mathbb{R}} \overline{c}_{p,i}$.

APPENDIX

D

DATASET IMAGES

This appendix contains the illustrations of all datasets employed on the experiments of this paper, which are: the bidimensional torus, the Crescent Moon dataset, the Lorenz Attractor, the Rössler Attractor and the MacKey-Glass Attractor.

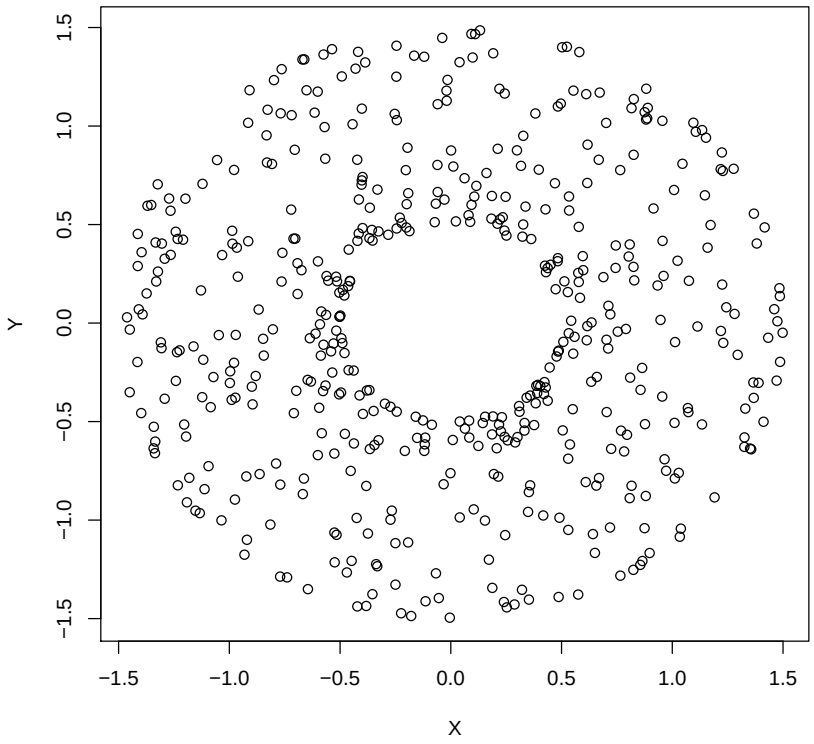


Figure 39 – A two-dimensional torus formed by a point cloud of 500 samples generated using Equation 5.1 with parameters $R = 1, r = 0.5, \phi \in [0, 2\pi]$ and $\theta \in [0, 2\pi]$.

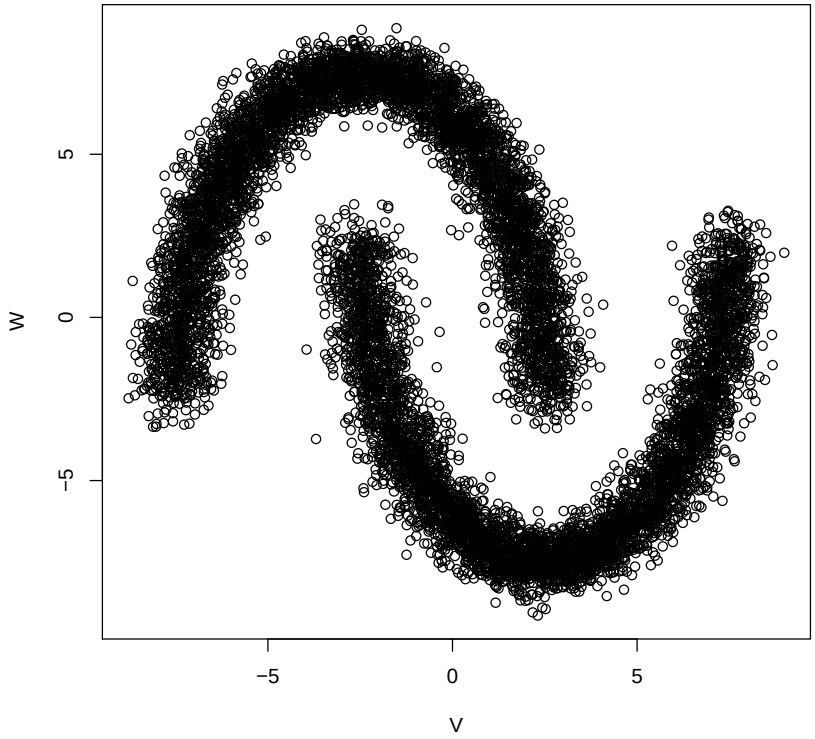


Figure 40 – The Crescent Moon dataset produced using the function *generateCrescentMoon* of the RSSL package from the R Project for Statistical Computing, adopting the parameters: $n = 5,000, d = 2$ and $\sigma = 0.5$.

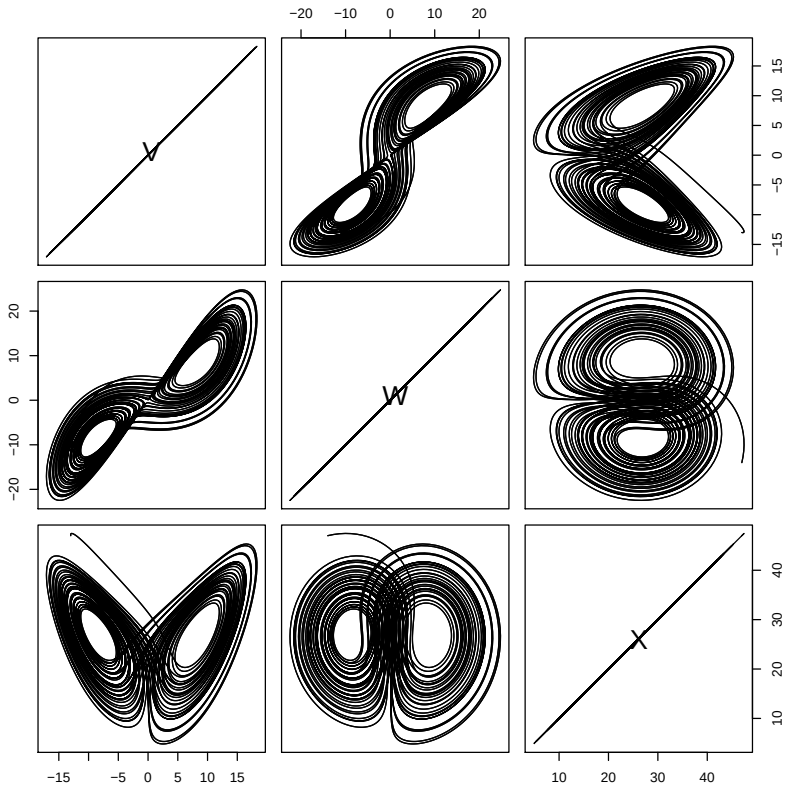


Figure 41 – The Lorenz attractor generated using the *lorenz* function of the package nonlinearTseries from the R Project for Statistical Computing with the parameters: $\sigma = 10, \beta = 8/3$ and $\rho = 28$ with initial conditions given by $x_0 = -13, y_0 = -14$ and $z = 47$ (plotted using lines to improve visualization).

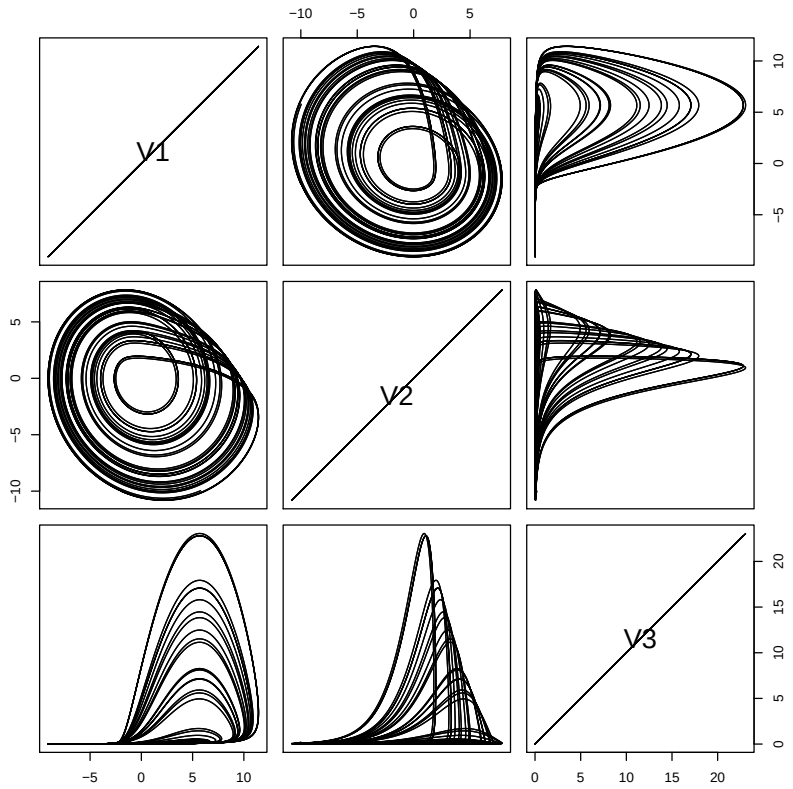


Figure 42 – The Rössler attractor generated using the function *rossler*, implemented in package nonlinearTseries from the R Project for Statistical Computing, with the parameters: $a = 0.2, b = 0.2$ and $w = 5.7$ with initial conditions given by $x_0 = -2, y_0 = -10$ and $z_0 = 0.2$ (plotted using lines to improve visualization).

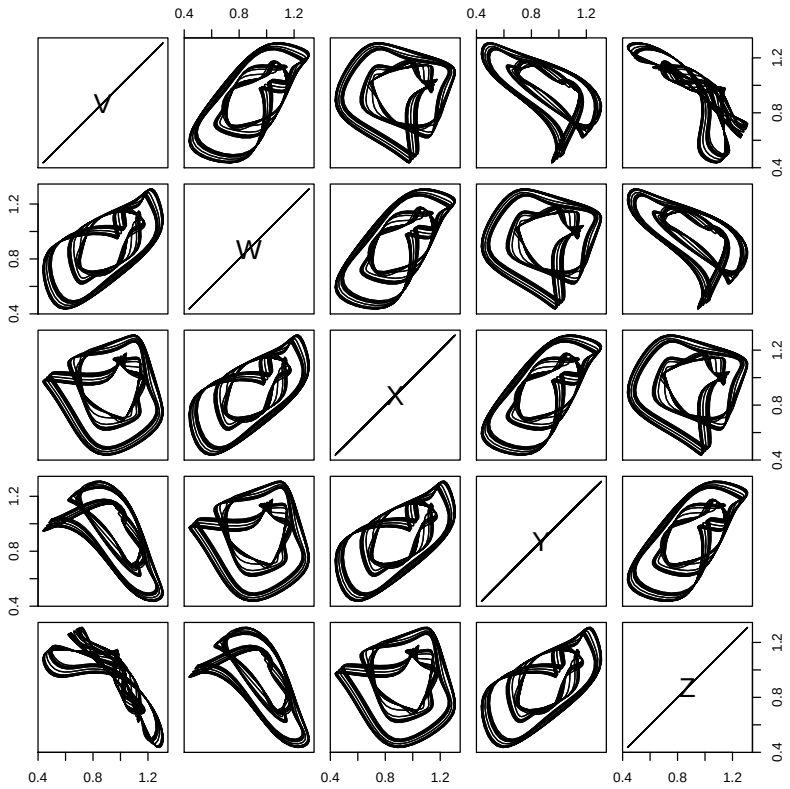


Figure 43 – The Mackey-Glass attractor available with the package frbs from the R Project for Statistical Computing (plotted using lines to improve visualization).

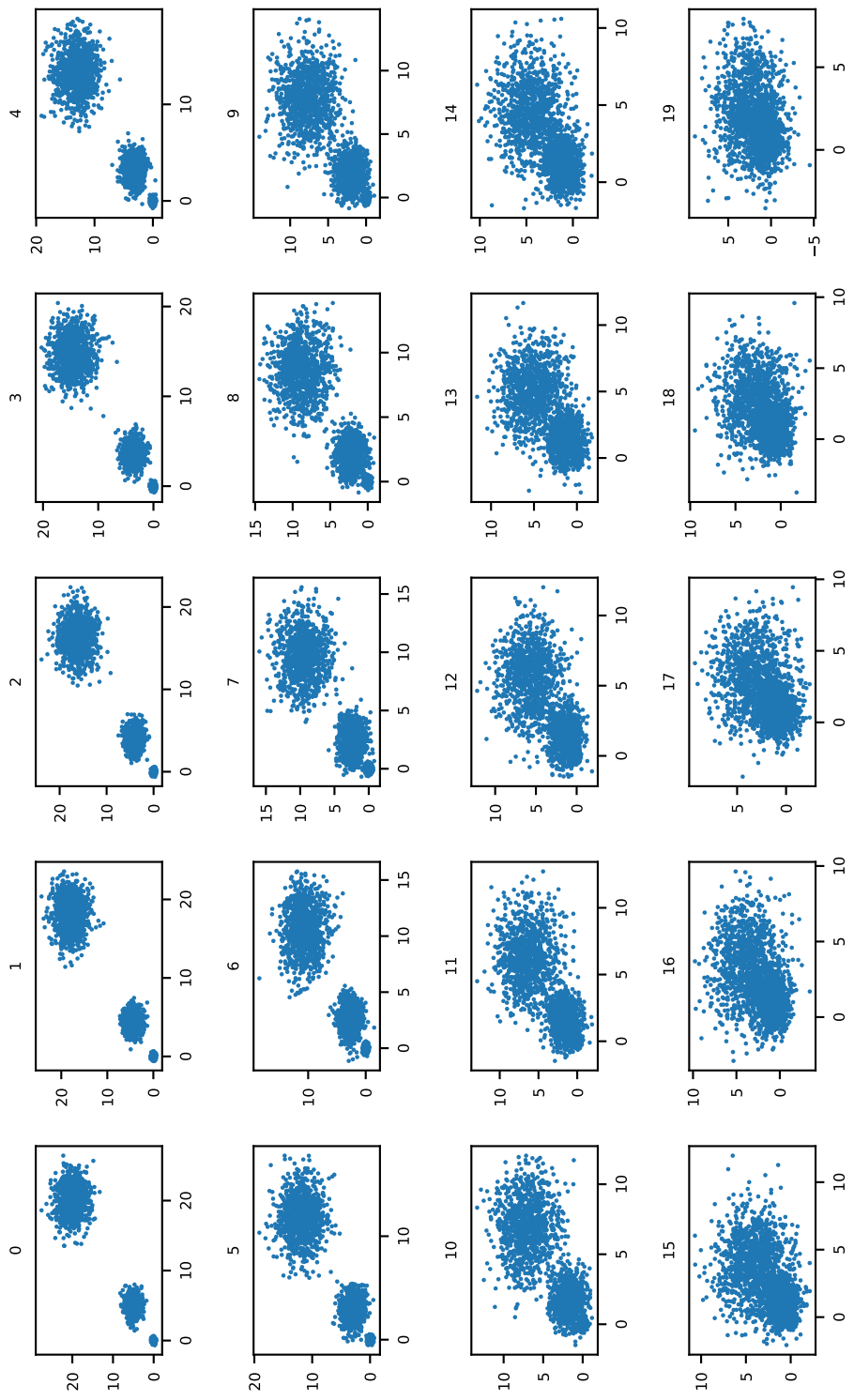


Figure 44 – The synthetic dataset of Section 5.3.3 corresponding to three Gaussians approaching one another.

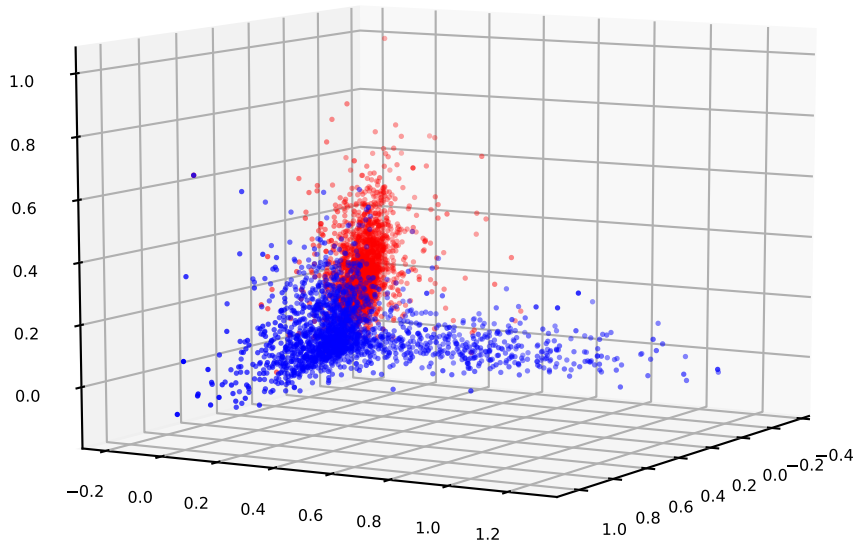


Figure 45 – The dataset $X(2,010)$ considered in Section 5.5 after transformation applied by Non-negative Matrix Factorization. Red points are the data associated with abstracts whose main subject is "Machine Learning" while blue ones are associated with "Time Series".

APPENDIX

E

BARCODE PLOTS

This appendix contains the barcode plots produced by the datasets: Torus, Crescent Moon, Lorenz Attractor, Rössler Attractor and MacKey-Glass Attractor. They are produced considering the dataset with and without perturbations and the generalization measure $|\beta_p(\mathcal{X}_{r_i}) - \beta_p(\mathcal{X}^m_{r_i})|$ is also considered in the illustrations.

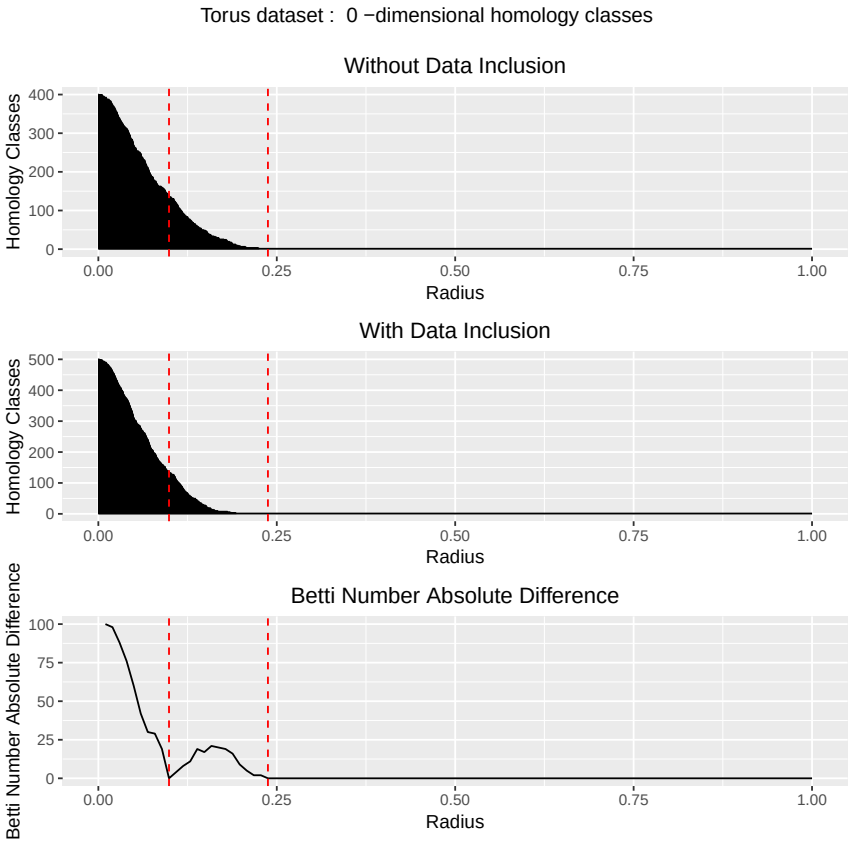


Figure 46 – Graphs produced from 0-dimensional homology classes which correspond to, from top to bottom: (i) Barcode plots generated over the torus with 400 samples; (ii) Barcode plots generated over the perturbed dataset with 500 samples; and, finally, (iii) the values for the generalization measure $G_0(\mathcal{X}_r)$. Red-dashed lines mark the initial value of the intervals which ensure $G_0(\mathcal{X}_r) = 0$.

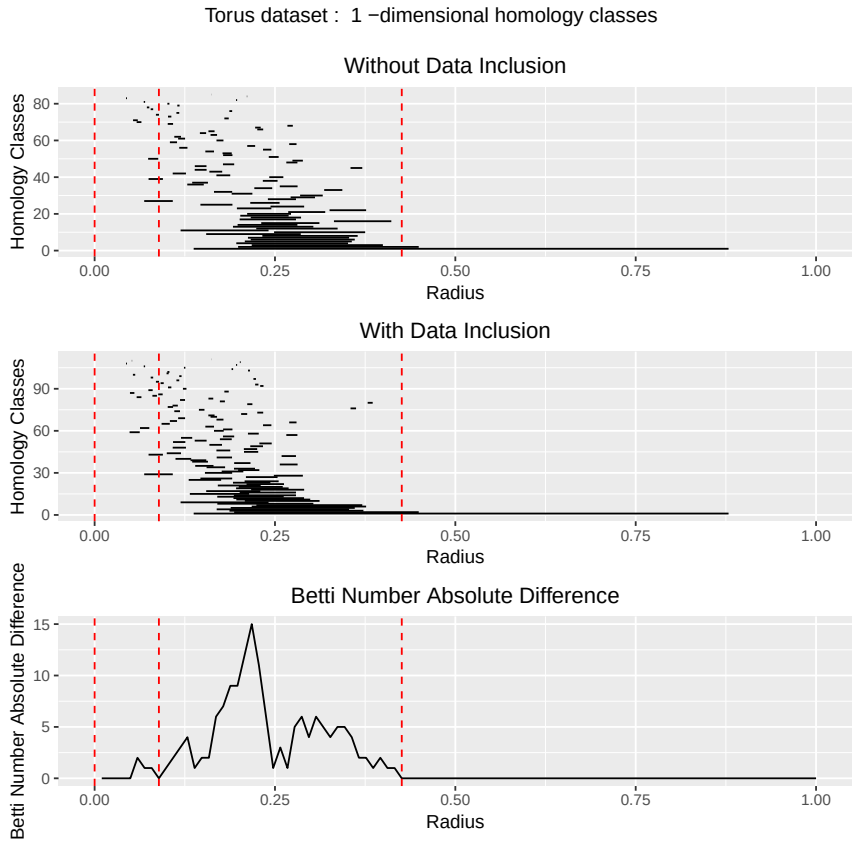


Figure 47 – Graphs produced from one-dimensional homology classes which correspond to, from top to bottom: (i) Barcode plots generated over the torus with 400 samples; (ii) Barcode plots generated over the perturbed dataset with 500 samples; and, finally, (iii) the values for the generalization measure $G_1(\mathcal{X}_r)$. Red-dashed lines mark the initial value of the intervals which ensure $G_1(\mathcal{X}_r) = 0$.

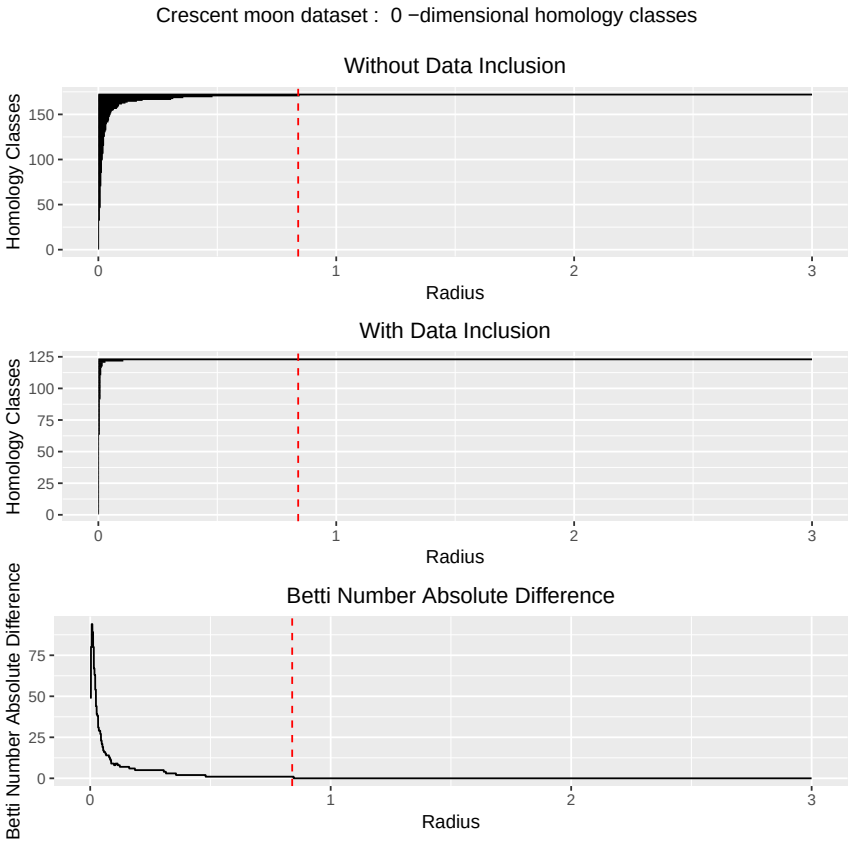


Figure 48 – Graphs produced from 0-dimensional homology classes which correspond to, from top to bottom: (i) Barcode plots generated using the Crescent Moon dataset with $5,100$ samples; (ii) Barcode plots generated using the perturbed dataset with $10,000$ samples; and, finally, (iii) the values for the generalization measure $G_0(\mathcal{X}_r)$. Red-dashed lines mark the initial value of the intervals which ensure $G_0(\mathcal{X}_r) = 0$.

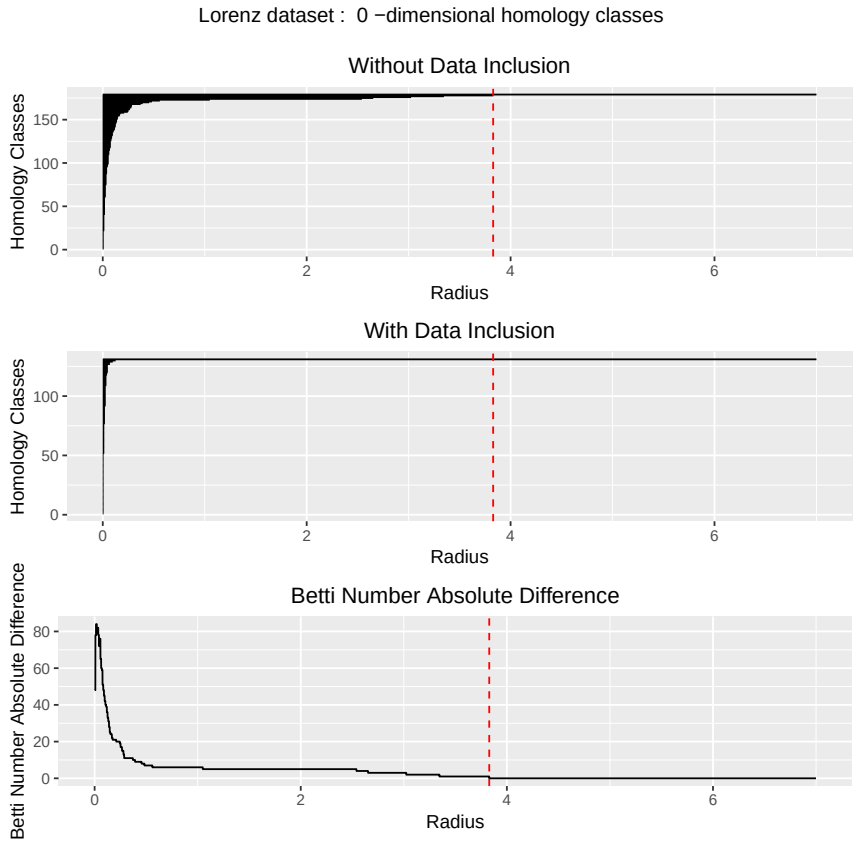


Figure 49 – Graphs produced from 0-dimensional homology classes which correspond to, from top to bottom: (i) Barcode plots generated using the Lorenz attractor with $5,101$ samples; (ii) Barcode plots generated using the perturbed dataset with $10,001$ samples; and, finally, (iii) the values for the generalization measure $G_0(\mathcal{X}_r)$. Red-dashed lines mark the initial value of the intervals which ensure $G_0(\mathcal{X}_r) = 0$.

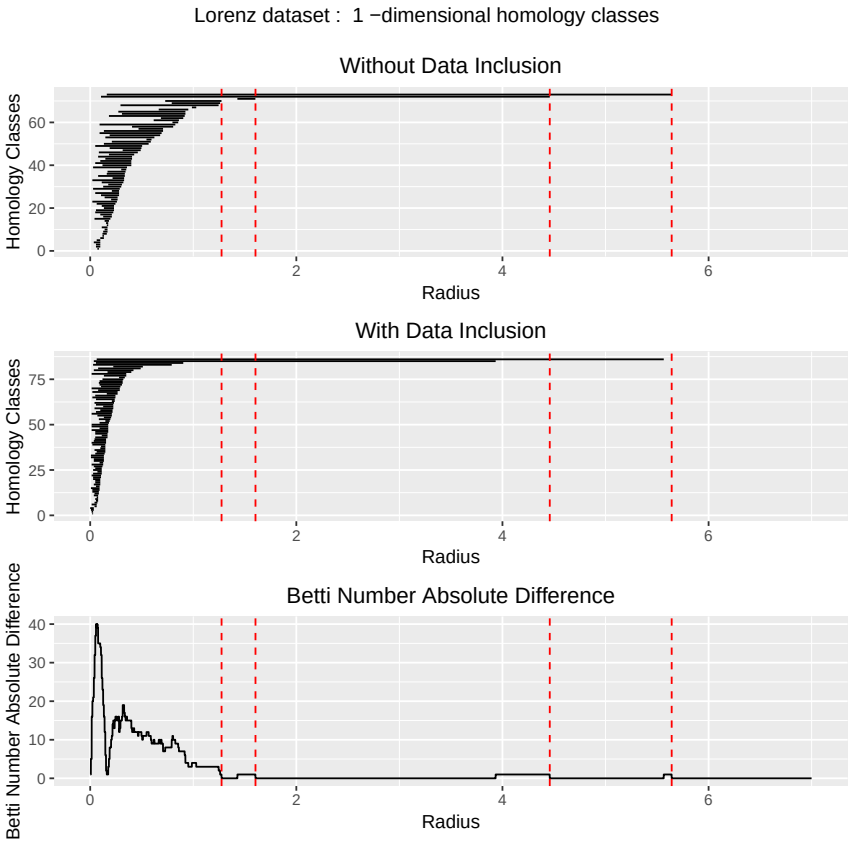


Figure 50 – Graphs produced from one-dimensional homology classes which correspond to, from top to bottom: (i) Barcode plots generated using the Lorenz attractor with $5,101$ samples; (ii) Barcode plots generated using the perturbed dataset with $10,001$ samples; and, finally, (iii) the values for the generalization measure $G_1(\mathcal{X}_r)$. Red-dashed lines mark the initial value of the intervals which ensure $G_1(\mathcal{X}_r) = 0$.

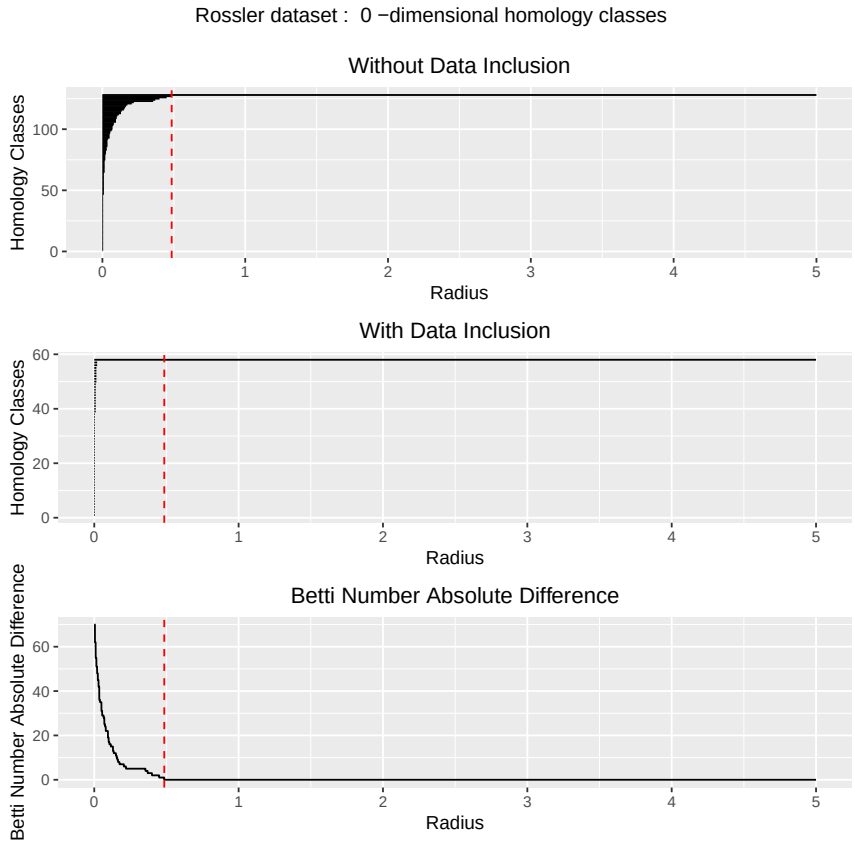


Figure 51 – Graphs produced from 0-dimensional homology classes which correspond to, from top to bottom: (i) Barcode plots generated using the Rössler attractor with $5,101$ samples; (ii) Barcode plots generated using the perturbed dataset with $10,001$ samples; and, finally, (iii) the values for the generalization measure $G_0(\mathcal{X}_r)$. Red-dashed lines mark the initial value of the intervals which ensure $G_0(\mathcal{X}_r) = 0$.

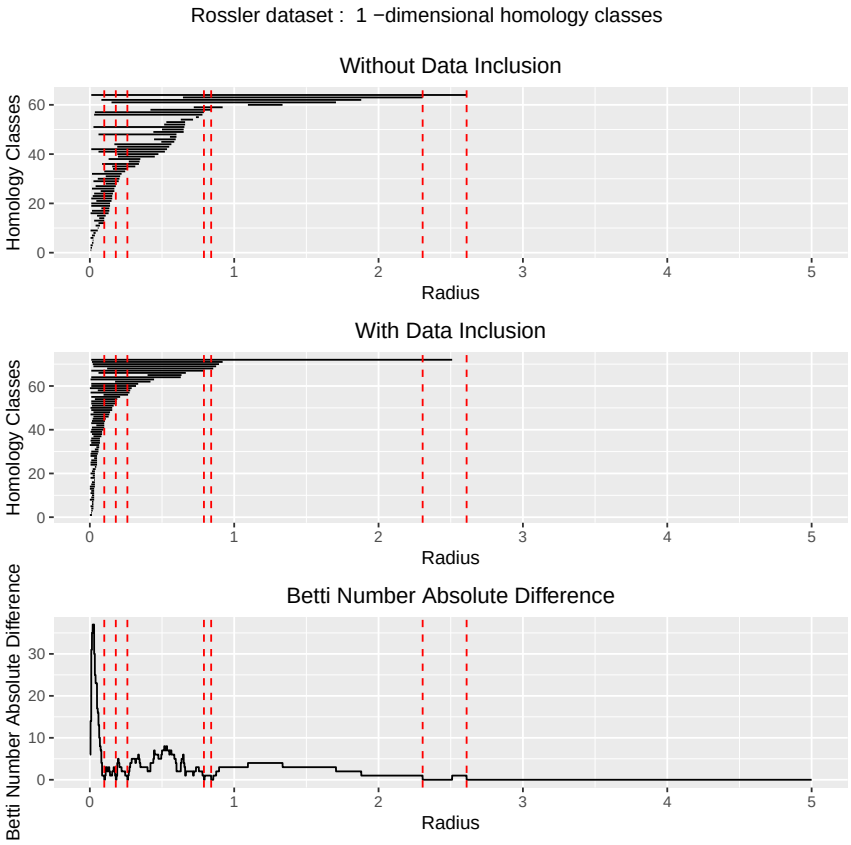


Figure 52 – Graphs produced from one-dimensional homology classes which correspond to, from top to bottom: (i) Barcode plots generated using the Rössler attractor previously described with $5,101$ samples; (ii) Barcode plots generated using the perturbed dataset with $10,001$ samples; and, finally, (iii) the values for the generalization measure $G_1(\mathcal{X}_r)$. Red-dashed lines mark the initial value of the intervals which ensure $G_1(\mathcal{X}_r) = 0$.

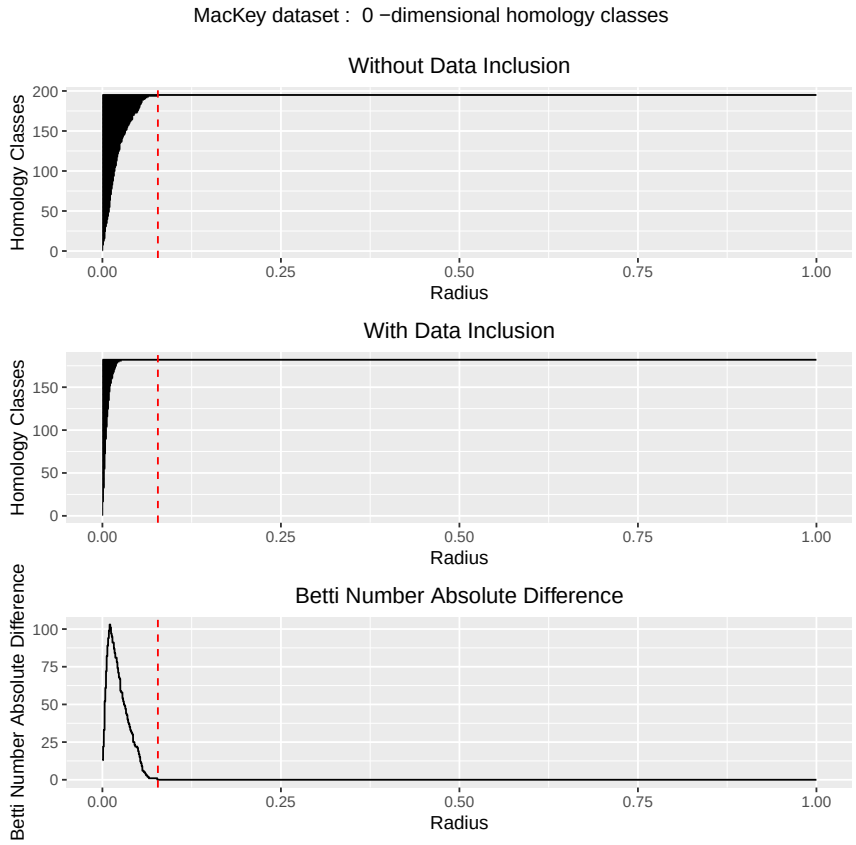


Figure 53 – Graphs produced from 0-dimensional homology classes which correspond to, from top to bottom: (i) Barcode plots generated using the Mackey-Glass attractor with 500 samples; (ii) Barcode plots generated using the perturbed dataset with $1,000$ samples; and, finally, (iii) the values for the generalization measure $G_0(\mathcal{X}_r)$. Red-dashed lines marks the initial value of the intervals which ensure $G_0(\mathcal{X}_r) = 0$.

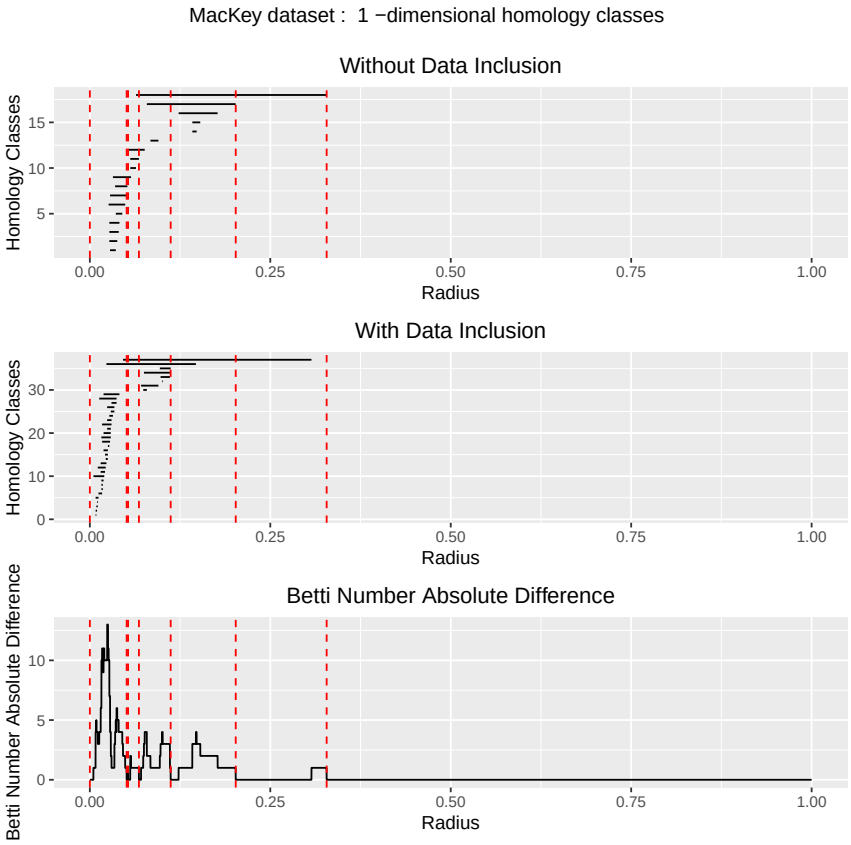


Figure 54 – Graphs produced from one-dimensional homology classes which correspond to, from top to bottom: (i) Barcode plots generated using the Mackey-Glass attractor previously described with 500 samples; (ii) Barcode plots generated using the perturbed dataset with $1,000$ samples; and, finally, (iii) the values for the generalization measure $G_1(\mathcal{X}_r)$. Red-dashed lines marks the initial value of the intervals which ensure $G_1(\mathcal{X}_r) = 0$.

APPENDIX

F

TOPOLOGICAL CONCEPT DRIFT GENERALIZATION MEASUREMENTS

This appendix contains the chart with the generalization divergence calculated for the Topological Concept Drift experiment detailed in Section 5.3.3.

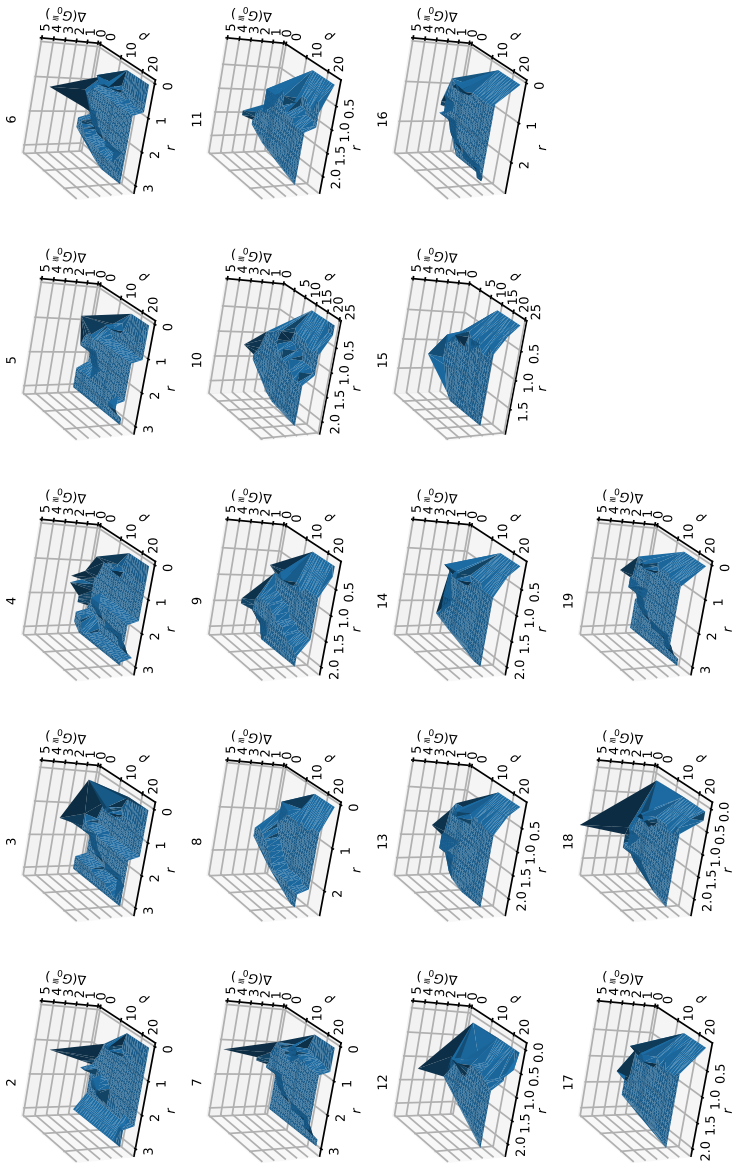


Figure 55 – Generalization divergences regarding epochs (as discussed in Section 5.3.3) along parameters r and ρ.

APPENDIX

G

WORD CLUSTERS OF DOCUMENT SEMANTIC CHANGES

This appendix contains word clusters associated with the document semantic changes detailed in Section 5.5.

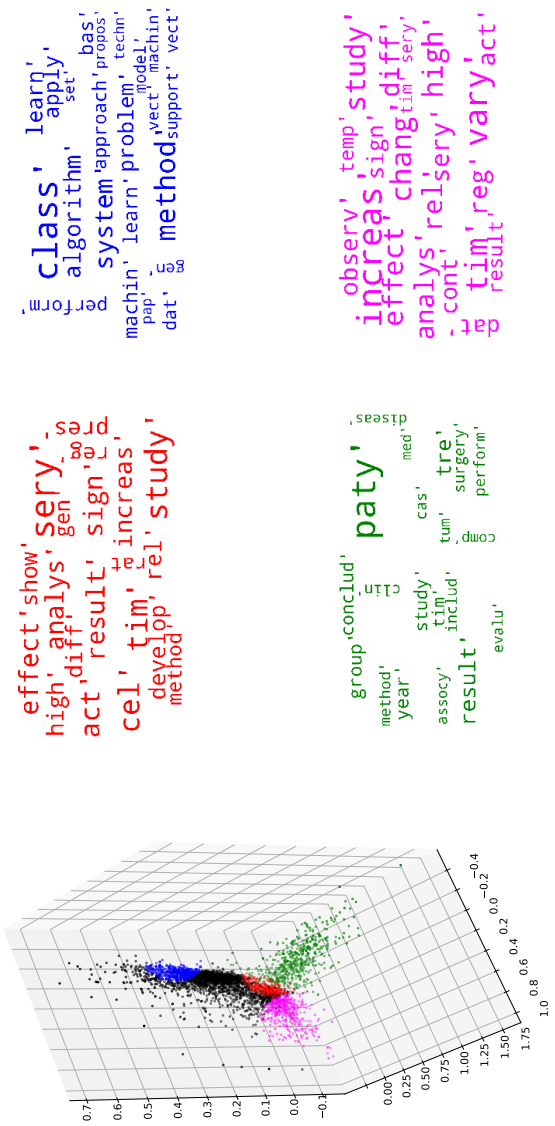


Figure 56 – Word clusters related with regions defined over the features $X(2,004)$ produced by Non-negative Matrix Factorization applied over the abstracts of the year $2,004$, as defined in Section 5.5.

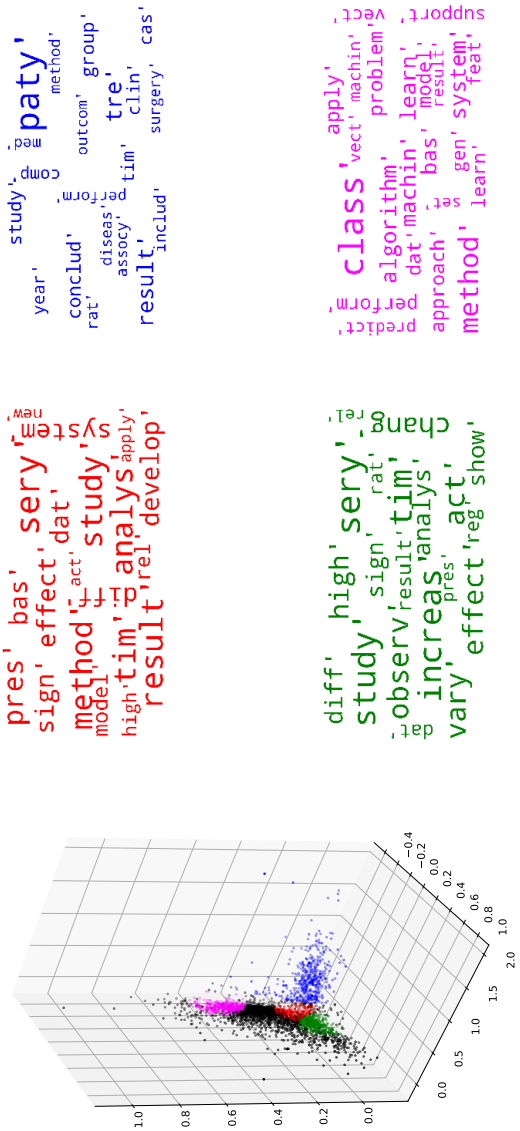


Figure 57 – Word clusters related with regions defined over the features $X(2,007)$ produced by Non-negative Matrix Factorization applied over the abstracts of the year $2,007$, as defined in Section 5.5.

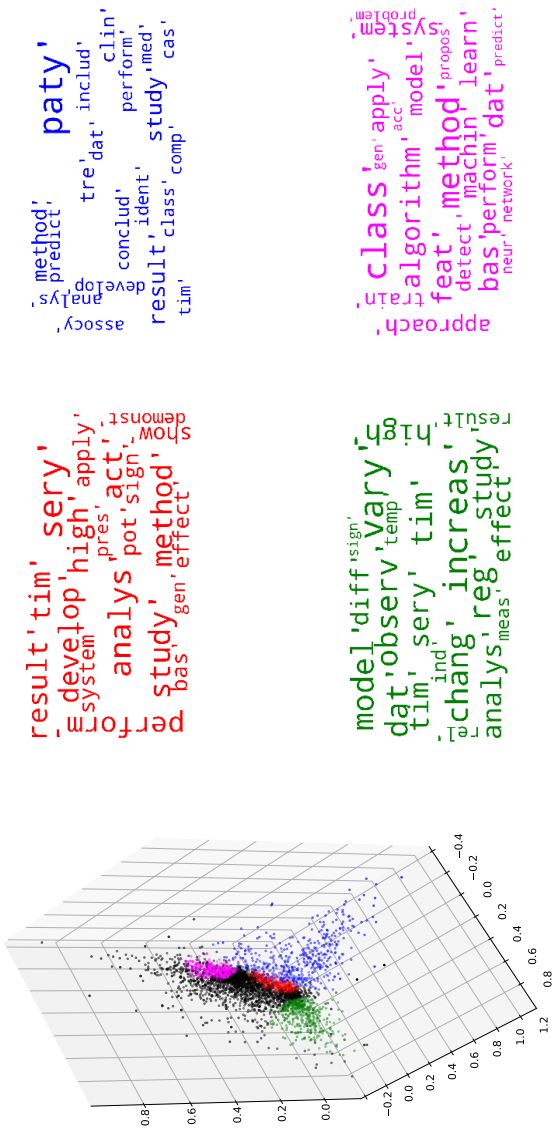


Figure 58 – Word clusters related with regions defined over the features $X(2,016)$ produced by Non-negative Matrix Factorization applied over the abstracts of the year $2,016$, as defined in Section 5.5.


ICMC
USP
SÃO CARLOS