University of Warsaw
Faculty of Mathematics, Informatics and Mechanics

Niklas Hellmer

Probability Meets Topological Data Analysis

PhD dissertation

Supervisor
dr hab. Paweł Dłotko
Institute of Mathematics
Polish Academy of Sciences

June 2024

DECLARATION

Author's declaration:
I hereby declare that this dissertation is my own work.

Date	Niklas Hellmer

Supervisor's declaration:
The dissertation is ready to be reviewed.

Date	dr hab. Paweł Dłotko

I read the news today - oh, boy
Four thousand holes in Blackburn, Lancashire
And though the holes were rather small
They had to count them all
Now they know how many holes it takes to fill the Albert Hall

---

A Day in the Life
The Beatles

Abstract

This thesis presents a collection of results at the interface between probability and topological data analysis (TDA). From a data driven perspective, we assume that samples lie on or near a submanifold (hence, topology enters) in high-dimensional feature space and are governed by a (usually unknown) probability distribution. The results we present illustrate different aspects of cross-fertilization between the subjects.

First, ideas from probability can be applied to the setting of topological data analysis. Specifically, we adapt the Prokhorov distance and related notions from optimal transport to persistence diagrams. These persistence diagrams are successful tools in TDA to capture multi-scale geometric and topological information of a dataset. We introduce an appropriate discrete Prokhorov distance for this setting via what we call bottleneck profiles. They generalize the bottleneck distance and satisfy bounds with respect to the Wasserstein distance, which both have been studied previously and applied with great success. In addition to the theoretical inquiry, in which we establish a stability result, we also provide algorithms and discuss numerical experiments.

Second, *vice versa*, ideas from topological data analysis can be used to solve statistical problems. Consider the set-up of n points sampled i.i.d. from a probability distribution F on $\mathbb{R}^d$. Classical statistical tools are often confined (theoretically or computationally) to the case $d = 1$, a problem we address by introducing tools from TDA which are agnostic to the ambient dimension. We study the Euler characteristic curve of the Čech complex as $n \to \infty$ in the thermodynamic limit regime. It turns out that two different probability distributions yield the same expected Euler characteristic curve in this regime if and only if their densities admit the same excess mass transform. We propose a goodness of fit test based on this result. Namely, given a sample from an unknown distribution, we test the null hypothesis that its density has a specific excess mass transform. We construct a consistent test statistic and show that the probability of type II errors vanishes exponentially as the sample size goes to infinity. Moreover, we present vast numerical experiments showcasing the superiority (in terms of test power) of this test over classical ones, even in low dimensions. The perhaps surprising result here is that while the hypothesis is stated purely in probability-theoretic language, the test statistic and its implementation are rooted in computational topology. As a case study, we apply these ideas to the processing of signals from industrial machines. Via time-delay embeddings, we obtain a spatial point pattern from a time series. We test the hypothesis of the signal being non-periodic, which corresponds to the bearing in the machine from which the measurements were taken being intact. Broken bearings cause a significant alteration of the shape of the point cloud obtained from the time-delay embedding. By combining the novel TDA approach with classical spectral techniques, we relaibly detect the presence of a deterministic periodic feature among signals with high levels of non-Gaussian noise and reduce the error rate significantly.

Third, as a synthesis, we develop a way to combine metric-topological and probabilistic information. To this end, we introduce the measure Dowker complex. This combines the classical Dowker complex of a relation, which handles bivariate or directional data, with a second filtration parameter controlling density. Informally, this is a simplicial complex in which vertices form a simplex only if there is sufficient mass near them. We establish a stability theorem bounding the interleaving distance between homology of two such complexes by the Hausdorff distance and the Prokhorov distance of the input data points. As a consequence, we obtain a version of a law of large numbers that ascertains that the interleaving distance between (the homology of) the complex built on samples and the one of the true underlying

metric probability space converges to zero in probability. In this sense, the measure Dowker complex unites topological and probabilistic information.

Keywords: Topological data analysis; persistent homology; Euler characteristic curve; Prokhorov distance; random geometric complexes; goodness-of-fit test; time delay embedding; Dowker complex.

AMS MSC 2020 classification: 55N31; 62R40; 60B99; 55U10; 05E45; 60D05; 62H15.

Streszczenie

Niniejsza rozprawa przedstawia zbiór wyników uzyskanych na styku prawdopodobieństwa i topologicznej analizy danych (TDA). W rozważanych przez nas dziedzinach zakładamy, że próbki leżą na lub w bezpośredniej bliskości pewnej rozmaitości zanurzonej w wielowymiarowej przestrzeni. Rozmieszczenie tych punktów jest determinowane przez, zwykle nieznany, rozkład prawdopodobieństwa. Naszym zadaniem jest ekstrakcja istotnych cech determinujących kształt danej próbki punktów. Wyniki przedstawione w niniejszej rozprawie ilustrują różne aspekty wzajemnego przenikania się topologii, probabilistyki i analizy danych.

W pierwszej części rozprawy pokazujemy, jak pewne idee z rachunku prawdopodobieństwa znajdują zastosowanie w topologicznej analizie danych. W szczególności adaptujemy znaną w teorii prawdopodobieństwa odległość Prokhorova i powiązane pojęcia z dziedziny optymalnego transportu do diagramów persystencji. Diagramy te są skutecznym narzędziem do przechwytywania wielkoskalowych informacji geometrycznych i topologicznych charakteryzujących dany zbiór danych. W tym rozdziale wprowadzamy nową, dyskretną wersję odległości Prokhorova, używając jako pośrednie narzędzie tzw. "profile przewężeń" (bottleneck profiles). Profile te uogólniają klasyczne odległości typu bottleneck i mają dobre własności w stosunku do klasycznych metryk Wassersteina. Praca zawiera zarówno rozważania teoretyczne, w tym własność stabilności diagramów persystencji w oparciu o nową metrykę, jak i efektywne algorytmy oraz towarzyszące im eksperymenty numeryczne.

W drugiej części rozprawy pokazujemy, jak idee topologiczne mogą być wykorzystane do rozwiązywania problemów statystycznych. Rozważamy układ n punktów próbkowanych niezależnie i identycznie z rozkładu prawdopodobieństwa F na R^d. Klasyczne narzędzia statystyczne są często ograniczone (teoretycznie lub obliczeniowo) do przypadku $d = 1$. W rozprawie pokazujemy, jak ominąć to ograniczenie, wprowadzając narzędzia z TDA, które są niezależne od wymiaru. Nasze rozwiązania bazują na krzywej charakterystyki Eulera kompleksu Čech. Pokazujemy, że dwa różne rozkłady prawdopodobieństwa dają tę samą oczekiwaną krzywą charakterystyki Eulera wtedy i tylko wtedy, gdy ich gęstości dopuszczają tę samą 'excess mass transform' (wszystkie prezentowane wyniki mają charakter asymptotyczny). Ta charakteryzacja pozwala na wyprowadzenie nowej rodziny testów zgodności statystycznej. Mając daną próbkę z nieznanego rozkładu, testujemy hipotezę zerową stanowiącą, że jej gęstość ma określoną "excess mass transform". Konstruujemy topologiczną statystykę testową i pokazujemy, że prawdopodobieństwo błędów II rodzaju zanika wykładniczo wraz z rozmiarem próbki. Ponadto, w rozprawie przedstawiamy obszerne eksperymenty numeryczne pokazujące moc naszego testu. W szczególności pokazujemy, że jest on skuteczniejszy od dostępnych opcji, nawet dla danych o niskim wymiarze.

Pokazujemy również, jak przedstawiona metodologia testów statystycznych może zostać użyta do analizy sygnałów. Używając metody włożenie z opóźnieniem czasowym, przetwarzamy dany jednowymiarowy sygnał z szeregu czasowego na trajektorię w wysokowymiarowej przestrzeni. Następnie testujemy hipotezę mówiącą, że sygnał jest nieokresowy, co odpowiada poprawnie działającej maszynie. W tym przypadku uszkodzenia mechaniczne maszyny powodują okresową składową w rozważanym szeregu czasowym, która może być wykryta przy pomocy prezentowanej przez nas techniki. W szczególności pokazujemy, że połączenie proponowanego przez nas podejścia bazującego na TDA z klasycznymi technikami analizy spektralnej daje najlepsze efekty w detekcji uszkodzeń mechanicznych maszyny.

W trzeciej części rozprawy, w ramach syntezy, opracowujemy sposób łączenia informacji metryczno-topologicznej i probabilistycznej. W tym celu wprowadzamy opartą na teorii miary wersję kompleksu

Dowkera. Łączy on klasyczny kompleks Dowkera relacji z drugim parametrem filtracji kontrolującym gęstość danych. Nieformalnie, jest to kompleks, w którym kolekcja wierzchołków tworzy sympleks wtedy i tylko wtedy, gdy są one blisko siebie oraz w ich pobliżu znajduje się wystarczająca masa pozostałych punktów. Rozdział zawiera twierdzenie o stabilności proponowanej konstrukcji, ograniczające homologie tego kompleksu przez odległość Hausdorffa między punktami i odległość Prokhorova między gęstościami. Wyprowadzamy również prawo wielkich liczb, które zapewnia, że odległość między kompleksem zbudowanym na próbkach a kompleksem prawdziwej bazowej metrycznej przestrzeni prawdopodobieństwa zbiega z prawdopodobieństwem do zera.

Słowa kluczowe: Topologiczna analiza danych; homologie persystentne; krzywa charakterystyki Eulera; odległość Prokhorova; losowe kompleksy geometryczne; test zgodności statystycznej; włożenia z opóźnieniem; kompleks Dowkera.

Klasyfikacja AMS MSC 2020: 55N31; 62R40; 60B99; 55U10; 05E45; 60D05; 62H15.

Acknowledgement

First and foremost, I am most highly indebted to my advisor Paweł Dłotko. His steady encouragement, support, and profound insights have been invaluable in this academic journey of mine, which took me over the course of the last years through three different countries, a global pandemic and to diverse real-world applications. Also, Paweł provided coffee. Secondly, Davide Gurnari has been much more than just a great colleague and office mate. His data wizardry, tech support and memes helped me through many a struggle.

At the start of my PhD in Swansea, I profited from discussions with Jeff Giansiracusa, Nick Sale and Yue Ren. In Warsaw, the same goes for the Dioscui TDA members and alumni Michał Bogdan, Michał Lipiński, Jakub Malinowski, Bartosz Naskręcki, Jan F. Senge, Justyna Signerska-Rynkowska, Anastasios Stefanou and Rafał Topolnicki. Moreover, I thank the IMPAN staff for their support. I appreciate the hospitality of Bastian Rieck and his group at Helmholtz Munich, and also thank Ulrich Bauer's group at TU Munich for having me join their seminar while I was visiting. Bastian Rieck also created the LaTeXtemplate `mimosis` used for this thesis, which I gratefully acknowledge as well as his support debugging compatibility issues. I thank my coauthors Tobias Fleckenstein, Justyna Hebda-Sobkowicz, Jan Spaliński, Łukasz Stettner, Rafał Topolnicki, Agnieszka Wyłomańska, Radosław Zimroz. Everyone who read my papers, attended my talks and discussed with me – you know who you are. Special thanks for providing invaluable comments on drafts of this thesis go to Julian Brüggemann, Davide Gurnari and Jan Spaliński.

This work was supported through Paweł Dłotko's grant from the Dioscuri program initiated by the Max Planck Society, jointly managed with the National Science Centre (Poland), and mutually funded by the Polish Ministry of Science and Higher Education and the German Federal Ministry of Education and Research. In addition, I acknowledge the financial support from the IDUB program under actions IV.1.2 and IV.4.1 of POB3.

Contents

1 Introduction

The main paradigm of topological data analysis (TDA) is that *data has shape*, and by understanding this shape one can gain insights about the data. Probability enters the stage as the data samples are governed by an (unknown) probability distribution.

An abstract version of the basic *TDA pipeline* is depicted in Figure 1.1. We think of the data geometrically as a point cloud, in which each point is a sample whose coordinates are given by the values of the features. The assignment of topological summary statistics to the data is a Lipschitz-continous procedure – small perturbations in the data give rise to small changes in their topological summaries. This key property is also referred to as *stability*. To formalize this idea, we will need appropriate (pseudo-)metrics at each step in the pipeline. As a consequence, a probability measure governing the data gives rise to one on the topological summary. This is the context for the present study; a more thorough introduction to the themes will be given in Chapter 2. The rest of the dissertation is organized into roughly three major blocks, with Chapter 3 somewhat of a prelude, Chapters 4, 5 and 6 the centerpiece, and Chapter 7 the finale. More specifically:

- Chapter 3 takes ideas from probability theory and applies them to TDA. We adapt the Prokhorov distance, which is a classical notion in probability and optimal transport, to the setting of TDA, creating a new, more robust way to compare topological information.

- Chapters 4, 5 and 6 study the application of topological invariants to statistical problems. We first investigate under what conditions a topological invariant (the *Euler characteristic curve*) can (asymptotically) distinguish samples from two different distributions. This result is then used to construct a goodness of fit test, i.e. we can assess stochastic models via topological signatures of samples. As an application, we study point clouds which arise as state space reconstruction from measurements of vibrations of heavy duty machines in the mining industry and detect bearing failures.

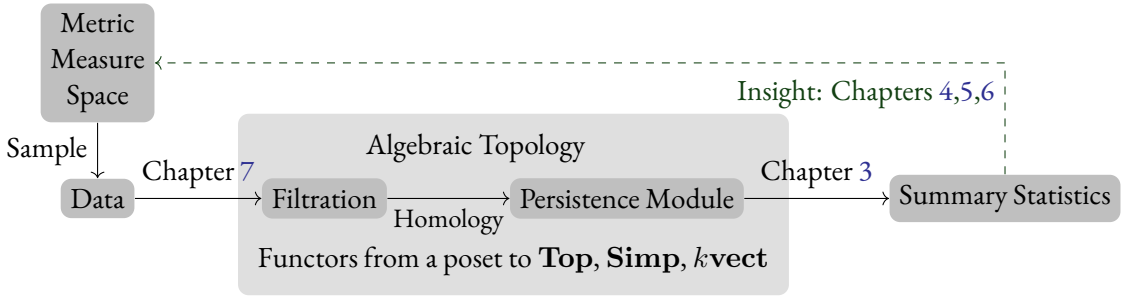


Figure 1.1: A general TDA pipeline

- Chapter 7 finally provides a framework for combining statistical and topological information. We develop a simplicial complex, the *measure Dowker bifiltered complex*, to model interactions of two point clouds in a density-sensitive and robust way.

1.1 Overview

We now give a brief overview over the contents of each chapter, highlighting key results.

1.1.1 Bottleneck Profiles and Discrete Prokhorov Metrics for Persistence Diagrams (Chapter 3)

One parameter persistence, perhaps the most well-known implementation of the TDA pipeline, takes a (often highly complex) point cloud in Euclidean space as input and produces a point cloud in the plane, the persistence diagram (PD), as output. Intuitively, persistence diagrams serve as a summary of the shape of the input data. As a consequence, one can compare different shapes indirectly, by comparing their PDs. The need for a robust and computationally efficient notion of distance for PDs arises. Classically, one uses the Bottleneck and Wasserstein distances to this end [100]. However, the Bottleneck distance only picks up the single biggest difference between PDs and the Wasserstein distance is prone to noise, as it picks up every difference no matter how small.

This fact motivates our work to search for new metrics that could balance these issues. We introduce the notion of the bottleneck profile of two PDs, which is a map $[0,\infty[\to \mathbb{N}\cup\{\infty\}$ (Definition 3.2.1). This tool summarizes metric information at varying scales and generalizes the Bottleneck distance. Also the Wasserstein distance can be, in special cases, computed from the bottleneck profile; in general, it can be bounded via a bottleneck profile.

The bottleneck profiles arises naturally in a discrete version of the Prokhorov distance, which is a classical tool in probability theory. It turns out that the Bottleneck and the Prokhorov distance are just two instances of a whole family of Prokhorov-style metrics discussed in this chapter (Definition 3.3.1). In fact, this family is parameterised by a subclass of functions $f\colon [0,\infty[\to[0,\infty[$. Not every function f gives in fact rise to a genuine metric; we examine the conditions on f in which cases it does (Definition 3.3.2, such f are called *admissible* functions). In particular, we show:

Theorem 3.3.7. *Fix an admissible function $f\colon [0,\infty[\to[0,\infty[$. The discrete f-Prokhorov metric π_f is an extended pseudometric.*

In addition to theoretical development, we discuss algorithms to compute the bottleneck profile and various Prokhorov-type distances. In particular, a computational complexity analysis of those algorithms is given:

Proposition 3.3.21. *Let $f\colon [0,\infty[\to[0,\infty[$ be monotonically increasing. Assume that the values and preimages of f can be computed in $O(1)$. Then $\pi_f(X,Y)$ can be computed in $O(n^2\log(n))$.*

We provide a run-time analysis and experiments on a number of data sets. The algorithms are provided as an open source implementation.

1.1.2 When Do Two Distributions Yield the Same Expected Euler Characteristic Curve in the Thermodynamic Limit? (Chapter 4)

The *Čech complex* $\mathcal{C}(X)_r$ of a finite point cloud $X \subset \mathbb{R}^d$ has vertices X and simplices $\sigma \subseteq X$ if the intersection $\bigcap_{x\in\sigma} \overline{B}_r(x)$ is non-empty. Here, $r \geqslant 0$ is the filtration parameter, meaning that $\mathcal{C}(X)_r \subseteq \mathcal{C}(X)_s$ whenever $r \leqslant s$. In this chapter, we are interested in the case when $X = X_n = \{x_1, \dots, x_n\}$ consists of n i.i.d. samples from some probability distribution F on $\mathbb{R}^d$. As the sample size n goes to infinity, there are three limiting regimes governing the topology of $\mathcal{C}(X_n)_{r_n}$, which are distinguished by the behaviour of $\Lambda_n = n\omega_d r_n^d$. Here, ω_d is the volume of a unit ball in $\mathbb{R}^d$ and $(r_n)_n$ is a sequence of parameters of the Čech complex. In the dense regime, $\Lambda_n \to \infty$, the Čech complex is connected; if Λ_n grows fast enough, it recovers the topology of the support of F with high probability. However, no other information about the distribution is kept. In the thermodynamic regime, $\Lambda_n \to \Lambda \in]0, \infty[$, on the other hand, we cannot recover the support of F but can hope to capture different information about the distribution. Finally, in the sparse regime, $\Lambda_n \to 0$, the Čech complex is so disconnected that it does not retain much information.

This raises the question what properties of the distribution are in fact captured by the topology of the Čech complex in the thermodynamic limit. To this end, Vishwanath et al. [157] have recently introduced the concept of "$\mathcal{F}$-equivalence", which provides a sufficient condition for probability distributions to have Čech complexes which are indistinguishable by means of topological invariants in this asymptotic regime. Specifically, two probability density functions $f, g\colon \mathbb{R}^d \to [0, \infty[$ are $\mathcal{F}$-equivalent if they admit the same excess mass $\hat{f} = \hat{g}$, where

$$\hat{f}(t) := \int_{\mathbb{R}^d} \mathbb{1}_{[t,\infty[}(f(x))\, f(x) \mathrm{d}x. \tag{1.1}$$

The main result of the present chapter is to show that this condition is indeed also necessary in the setting of expected Euler characteristic curves. The two preceding statements can be succinctly combined into the following theorem:

Theorem 4.0.1. *Let F, G be probability distributions on $\mathbb{R}^d$ with densities with respect to the Lebesgue measure f, g which are bounded. The following are equivalent:*

i) The excess mass transforms agree: $\hat{f}(t) = \hat{g}(t)$ for all $t > 0$,

ii) in the thermodynamic limit, the expected Euler characteristic curves agree: $\overline{\chi}_F(\Lambda) = \overline{\chi}_G(\Lambda)$ for all $\Lambda > 0$.

The implication $i) \Rightarrow ii)$ was established by Vishwanath et al. [157]; the subject of this chapter is to show the perhaps surprising implication $ii) \Rightarrow i)$. This is Theorem 4.3.1 below.

1.1.3 Topology-Driven Goodness-of-Fit Tests in Arbitrary Dimensions (Chapter 5)

Goodness-of-fit (GoF) testing is one of the standard tasks in statistics. The testing procedure can be stated in the one-sample or two-sample setting. In case of the one-sample problem, we observe a sample of m independent realizations $\{x_1, \dots, x_m\}$ of a d-dimensional random vector X with an

unknown distribution function G, i.e. $x_i \sim G$. The task is to test whether G is equal to a specific distribution F, i.e. we would like to test the following null hypothesis H_0 against the alternative H_1:

$$H_0 : G = F \text{ vs. } H_1 : G \neq F. \tag{1.2}$$

In the setting of the two-sample problem we are given two independent samples consisting of m and n ($m \neq n$ in general) independent realizations of d-dimensional random vectors X and Y with unknown distribution functions F and G, respectively. This means $X = \{x_1, \ldots, x_m\}, x_i \sim F$ and $Y = \{y_1, \ldots, y_n\}, y_j \sim G$, while the hypothesis is the same as in (1.2).

In the light of the preceding chapter, we consider a more general notion of equivalence, replacing the equal sign above by the relation of having the same excess mass (Equation 1.1 on the previous page; Definition 4.1).

We are interested in the setting in which the underlying distribution is continuous. In this case, prominent GoF tests for samples from $\mathbb{R}$ rely on the empirical distribution function, see [57, Chapter 4]. These include, in the one dimensional case, the Kolmogorov-Smirnov, Cramér-von-Mises and Anderson-Darling tests. In higher dimensions, Kolmogorov-Smirnov leads to Fasano-Franceschini[73] and Peacock[122] tests; a general case was considered by Justel [95]. A multivariate version of Cramér-von-Mises was proposed by Chiu and Liu[48]. Since those tests are based on the empirical distribution function, their generalization to $\mathbb{R}^d$ for $d \geqslant 2$ is conceptually and computationally difficult. Moreover, we are not aware of an efficient implementation of a general goodness-of-fit test for high dimensional samples.

To tackle this challenge we propose to replace the cumulative distribution function by the *Euler characteristic curve (ECC)* [85, 130, 160], a tool from computational topology that provides a signature of the considered sample. To a given sample X, this notion associates a function $\chi(X) \colon [0, \infty) \to \mathbb{Z}$, which can serve as a stand-in for the empirical distribution function in arbitrary dimension. Subsequently, for one-sample tests, inspired by the Kolmogorov-Smirnov test, we define the test statistic to be the supremum distance between the ECC of the sample and the expected ECC for the distribution. This topologically driven testing scheme will be referred to as "TopoTests" for short.

The key characteristic of any goodness of fit test is its power, i.e. the type II error should be small, under the requirement that the type I error is fixed at level α. We show that the proposed test satisfies this condition and that it performs very well in practical cases.

Theorem 5.1.4. *For fixed α, the probability of a type II error goes to 0 exponentially as $n \to \infty$.*

In particular, even restricted to one dimensional samples, its power is comparable to those of the standard GoF tests.

1.1.4 Damage Identification in Rolling Element Bearings Using Topological Data Analysis (Chapter 6)

In this chapter, we consider the practical example of condition monitoring of complex heavy-duty machines in the mining industry. In the era of Industry 4.0, monitoring processes and systems becomes the main ingredient and the necessary condition for the successful delivery of the final product. With the unprecedented increase in measuring and storage capabilities, efficient protocols for the extraction

of useful information and its direct application into the decision-making pipeline remain a major challenge. Predictive maintenance is a good example of such research direction.

Consider a complex-design heavy duty machine operating under time-varying load and speed conditions. Vibration or acoustic measurement from such a machine may be considered as a mixture of informative signal p and non-informative signal s (later called noise). Presence of the informative signal p is an evidence of a malfunction of the machine. The ability to detect the presence of a signal p in high-amplitude noise s allows appropriate maintenance action to take place. The variable load speed and the unknown, mostly non-Gaussian, distribution of s present additional challenges in local fault detection. Although s may sometimes be modeled by Gaussian noise, there are important cases where this assumption does not hold. In this chapter, an automatic continuous monitoring technique that is agnostic to the type of distribution from which s is sampled is presented. As a result, the proposed approach is resistant to the difficulties that often arise in methods based on assumptions about the distribution of the analyzed signal.

We adapt the tools of TDA to the analysis of the signal $s + p$ and determine the existence of a non-zero component p in the observed signal. A significant advantage of the proposed approach is that knowledge about the distribution of the signal s (considered as a general disturbance signal) is not needed. Thus, the method could be suitable for different machines and various speed/load conditions.

In this chapter, the ability of the proposed approach for local damage detection is tested first on synthetic examples, then on rolling element bearings on a test rig (laboratory conditions), and finally on the acutal acoustic signal from the belt conveyor system operated in the mining company. It should be noted that, under laboratory conditions, the method was tested for different levels of speed, simultaneously for faulty and healthy bearings. Industrial data contained several data sets for healthy and faulty bearings with a non-Gaussian distribution of non-informative components (noise). The efficiency of the method was also analyzed using synthetic data with Monte Carlo simulations for a wide range of Signal-to-Noise Ratio (SNR) and level of non-Gaussianity expressed by parameter α - a stability index in α-stable distribution. The proposed method has also been compared with state-of-the-art methods, commonly used in bearing diagnostics, i.e., spectral kurtosis [9], conditional variance-based (CVB) selector [88], as well as infograms [8, 89].

1.1.5 Density Sensitive Bifiltered Dowker Complexes via Total Weight (Chapter 7)

In TDA, persistent homology of Čech or Vietoris–Rips complexes is a standard tool to extract information about the shape of data. There are some shortcomings to this standard approach, notably its lack of sensitivity to density and against bivariate or directional data. Addressing these issues has been a focus of recent research. On the one hand, one can introduce a second filtration parameter to capture information about density [27], just like the proximity parameter of Čech or Rips controls metric information. On the other hand, Dowker complexes have received attention in applications involving directional [50] or bivariate [166] data. In this chapter, we combine the two approaches into what we call the *measure Dowker bifiltration* $\mathcal{MD}$ (Definition 7.1.6). We build on the total weight function of Robinson [134], which we rephrase using counting measures and then generalise to arbitrary measures. Roughly speaking, the idea is to construct a complex in which data points form a simplex only if there is sufficient mass near it; the mass can be the point cloud itself, a second point cloud or some ambient measure like Lebesgue's. We elaborate on the relation between our construction and other density

sensitive bifiltrations as well as Dowker duality in section 7.1: Notably, it turns out that our bifiltration is an instance of Sheehy's multicover bifiltration:

Theorem 7.1.4. *Let $R \subseteq X \times Y$ be a relation satisfying certain finiteness conditions. Then we have a weak equivalence of filtrations $|\mathcal{D}(X,Y,R)_\bullet| \simeq |\mathcal{S}(\mathcal{D}(Y,X,R^\top))_\bullet|$, where $\mathcal{S}$ is the subdivision filtration (Definition 2.2.35). Moreover, the weak equivalence is natural with respect to filtrations of relations.*

In section 7.2, we prove a robustness theorem for the measure Dowker complex of a finite metric space endowed with its empirical probability measure. In addition, we prove a stability theorem ascertaining that the change in homology of the measure Dowker bifiltration is upper-bounded by the maximum of Hausdorff distance between the data points and Prokhorov distance between the measures:

Theorem 7.2.4. *Suppose (Z,d) is a Polish space, endowed with Borel Σ-algebra $\mathfrak{B}(Z)$. Let $X_1, X_2 \in \mathfrak{B}(Z)$ and let μ_1, μ_2 be measures on $(Z, \mathfrak{B}(Z))$. Then for any $k \in \mathbb{N}$, we have*

$$d_I(H_k(\mathcal{MD}(X_1,\mu_1)), H_k(\mathcal{MD}(X_2,\mu_2))) \leqslant \max(\{d_H(X_1,X_2), d_{Pr}(\mu_1,\mu_2)\}),$$

where d_H is the Hausdorff distance (Definition 2.1.1) and d_{Pr} is the Prokhorov metric (Definition 2.1.6).

Moreover, we present an algorithm (Algorithm 7.1) to compute the measure Dowker bifiltered complex. We discuss its runtime complexity and make an open source implementation available on github[1]. We carry out several experiments showcasing applications to protein-ligand binding affinity prediction, clustering and dimensionality reduction of gene expression data and random hypergraphs of Erdös–Renyi type in section 7.3.

Relevant related work includes the study of functorial Dowker duality motivated by TDA by [37, 50, 134]. The total weight filtration of a Dowker complex was introduced by Robinson [134] and has also been studied in [152], where it is noted that this filtration is in general different from the one of the dual Dowker complex. Another approach to bifiltered Dowker complexes [24] was developed in parallel to this work. Applications of Dowker complexes include protein-ligand binding affinity prediction [113, 114], spatial patterns in the tumor microenvironment [166], music theory [79] and time series and dynamical systems analysis [81].

For the stability and robustness of two-parameter persistence, [27] is our main reference and inspiration; the work of Scoccola and Rolle [136, 142] is also of note.

1.2 Author's Contributions

Chapter 3 is based on an article [62] with Paweł Dłotko, who supervised the work. Chapter 4 is joint work with Tobias Fleckenstein, with equal contributions made by both authors. Chapter 5 is a slightly revised version of a joint article [63] with Paweł Dłotko, Łukasz Stettner and Rafał Topolnicki; R.T. and the author of this thesis are co-lead authors. Chapter 6 is coauthored with Justyna Hebda-Sobkowicz, Agnieszka Wyłomańska, Radosław Zimroz and Paweł Dłotko; N. H. is the lead author and responsible for the TDA results. Chapter 7 is joint work with Jan Spaliński, with N.H. being the leading author. More detailed statements of contributions are given at the start of each chapter.

[1] https://github.com/nihell/pyDowker

2 Background

What is now known as topological data analysis (TDA) emerged from parallel discoveries marked by the seminal articles by Edelsbrunner–Letscher–Zomorodian [70], Robins [133] and Carlsson–Zomorodian [168], though its roots trace back all the way to Vietoris [153]. Notably, the main motivation for Robins [132] was in non-linear dynamical systems: Can we infer the topology of the attractor (and thus information about the dynamics) from finite samples (e.g. obtained through numerical simulation)? Robins also laid the groundwork for random topology [131], although topics like random Voronoi tessellations, boolean models and their connectivity and Euler characteristic are classical in stochastic geometry [49]. This chapter aims to give an overview of the relevant background for the remainder of the thesis. It cannot, however, survey the compete area that is TDA today with its various interactions. For a textbook introduction with a focus on computations, see [69]; for a comprehensive treatment on the background concerning algebraic topology, there are many references available [34, 145].

2.1 Metric Measure Spaces

To compare two finite point clouds (or, more general subsets) A, B in some ambient metric space (X, d), the Hausdorff distance is a natural choice. As a prerequisite, introduce the distance to a subset $A \subseteq X$ as

$$d_A \colon X \to [0, \infty[, \qquad d_A(x) = \inf_{a \in A} d(a, x).$$

Definition 2.1.1. Let (X, d) be a metric space and $A, B \subseteq X$ be compact subsets. The *Hausdorff distance* between A and B is

$$d_H(A, B) = \sup_{x \in X} |d_B(x) - d_A(x)|.$$

Definition 2.1.2. Let (X, d) a metric space, $A \subseteq X$ and $\varepsilon > 0$. Define the *ε-thickening* of A to be

$$A^\varepsilon = \{x \in X \colon \exists a \in A \text{ such that } d(x, a) \leqslant \varepsilon\}.$$

The following alternative characterization is also commonly used [121, chapter 7, § 45]:

Proposition 2.1.3. $d_H(A, B) = \inf\{\varepsilon > 0 \colon A \subseteq B^\varepsilon \text{ and } B \subseteq A^\varepsilon\}$.

Next, we want to consider measures on metric spaces. To that end, we will impose the assumption that whenever we deal with measures, the underlying metric space is *Polish*, i.e. complete and separable, without explicitly mentioning it every time.

Recall that a *Borel Σ-algebra* $\mathcal{F}$ over a set X is a family of subsets of X satisfying the following conditions:

- $X \in \mathcal{F}$,
- $A \in \mathcal{F} \Rightarrow X \backslash A \in \mathcal{F}$,
- $A_i \in \mathcal{F}$ for all $i \in \mathbb{N} \Rightarrow \bigcup_{i=0}^{\infty} A_i \in \mathcal{F}$.

For a metric space (X, d), let $\mathfrak{B}(X)$ denote its Borel Σ-algebra; this is the smallest Σ-algebra containing all open and closed sets. Recall furthermore that a *measure* on a set X with Σ-algebra $\mathcal{F}$ is a function with values in the extended non-negative real numbers, $\mu \colon \mathcal{F} \to [0, \infty]$, subject to the following conditions:

- $\mu(\varnothing) = 0$,
- for any countable family of disjoint sets $\{A_i\}_{i \in \mathbb{N}}$ with $A_i \in \mathcal{F}$ for all $i \in \mathbb{N}$, we have

$$\mu\left(\bigcup_{i=0}^{\infty} A_i\right) = \sum_{i=0}^{\infty} \mu(A_i).$$

All measures we will consider are to be understood with respect to the Borel Σ-algebra of a metric space. As an example, consider the *Dirac measure* δ_x of a point $x \in X$ which is given by

$$\delta_x \colon \mathfrak{B}(X) \to [0, \infty], \qquad \delta_X(A) = \begin{cases} 1 & \text{if } x \in A, \\ 0 & \text{otherwise.} \end{cases}$$

For a finite metric space X, we consider the counting measure $\mu_X = \sum_{x \in X} \delta_X$ and the empirical probability measure $\nu_X = 1/|X| \sum_{x \in X} \delta_X$. A metric space with a Borel measure on it is known as a *metric measure space*; if this measure has total mass equal 1 it is a *metric probability space*. Given a continuous map $\varphi \colon X \to Y$ and a measure μ on X, denote its push-forward from X to Y by $\varphi_{\#}\mu$, it is defined by

$$\varphi_{\#}\mu(A) = \mu(\varphi^{-1}(A)).$$

Definition 2.1.4. A *coupling* of two measures μ, η on X is a measure γ on $X \times X$ whose marginals are μ and η, respectively. That is, $\pi^1_{\#}\gamma = \mu$ and $\pi^2_{\#}\gamma = \eta$ where $\pi^1, \pi^2 \colon X \times X \to X$ are the two canonical projections.

The following definition is non-standard but will be handy in Theorem 2.2.40 and Chapter 3.

Definition 2.1.5. Let μ, η be probability measures on a common metric space (X, d). The *Prokhorov profile* between μ and η is the function

$$\begin{aligned} \Pi_{\mu,\eta} \colon [0, \infty[&\to [0, 1] \\ \varepsilon &\mapsto \inf_{\gamma} \gamma(\{(x_1, x_2) \in X \times X \colon d(x_1, x_2) \geqslant \varepsilon\}), \end{aligned}$$

where γ ranges over all couplings of μ and η.

In words, the Prokhorov profile is a function that assigns to each ε the minimal amount of mass that needs to be transported over a distance $\geqslant \varepsilon$ in order to transform μ into η. We can now define the Prokhorov metric by intersecting a Prokhorov profile with a straight line:

Definition 2.1.6. Let μ, η be probability measures on a common metric space (X, d), let $l \geqslant 0$. The *l-Prokhorov metric* between μ and η is

$$d_{Pr_l}(\mu, \nu) \mapsto \inf\{\varepsilon > 0 \colon \Pi_{\mu,\eta}(\varepsilon) < l\varepsilon\}.$$

As pointed out in [154, chapter 27], there is a strong analogy between the Hausdorff distance in the theory of metric spaces and the Prokhorov distance in probability. The following alternative characterisation is often given as a definition of the Prokhorov metric, see for instance [128, (3.2.24)]. The equivalence of the two characterizations is due to Strassen's theorem [155, Remark 1.29].

Proposition 2.1.7.

$$d_{Pr_l}(\mu, \nu) = \inf\{\varepsilon > 0 \colon \mu(A) \leqslant \eta(A^\varepsilon) + l\varepsilon \text{ and } \eta(A) \leqslant \mu(A^\varepsilon) + l\varepsilon \text{ for all closed } A \subseteq X\}$$

Perhaps even more popular than Prokhorov as an optimal transport distance is the Wasserstein metric, also known as earth mover distance. This alias is due to the informal idea that the distance is the minimal cost of transporting one measure into another, with the cost being the product of mass and distance.

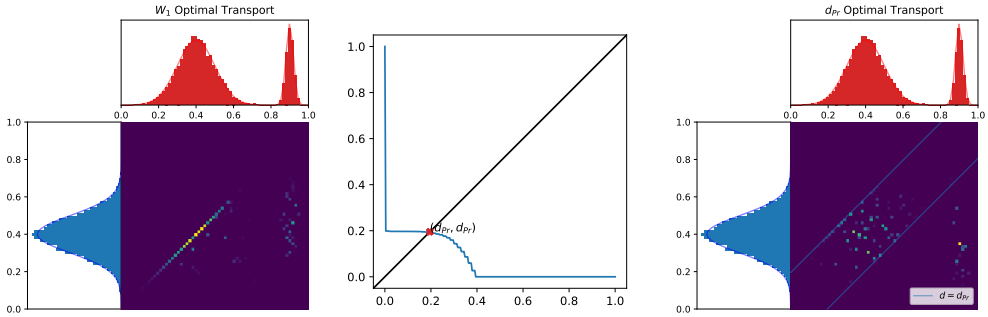
Definition 2.1.8. Let μ, η be probability measures on a common metric space (X, d), let $p \geqslant 1$. The *p-Wasserstein metric* between μ and η is

$$d_{W_p}(\mu, \eta) \mapsto \inf_{\gamma} \left(\int_{X \times X} d(x_1, x_2)^p \, \mathrm{d}\gamma(x_1, x_2) \right)^{1/p},$$

where γ ranges over all couplings of μ and η.

For convenience, we use the shorthand notations $d_{Pr} = d_{Pr_1}$ and $d_W = d_{W_1}$.

Example 2.1.9. Consider a normal distribution $\mu = \mathcal{N}(0.4, 0.1)$ and its perturbation given by a Gaussian mixture $\eta = 0.8 \cdot \mathcal{N}(0.4, 0.1) + 0.2 \cdot \mathcal{N}(0.9, 0.02)$. We sample 10000 points from each distribution and compute histograms on 64 equispaced bins in $[0, 1]$. For the 1-Wasserstein distance, one has to roughly transport a mass of 0.2 over a distance of 0.5 (namely from 0.4 to 0.9), leading to around 0.1 as value of the metric between the histograms. See Figure 2.1a for the optimal transport plan, which keeps most of the mass in their bins. For Prokhorov on the other hand, let us consider the Prokhorov profile shown in Figure 2.1b: for $0 < \varepsilon < 0.2$, its value stays around 0.2, indicating only a mass of 0.2 needs to be transported over a distance longer than such ε. It then decreases and hits 0 at around $\varepsilon = 0.4$, which means no mass needs to be transported over a distance greater than 0.4. The Prokhorov distance can be visually seen as the ε for which the blue graph of the Prokhorov profile intersects the black graph of the identity ; this point is marked in red in the figure. Roughly, only a mass of 0.2 is transported over a distance greater than 0.2, causing this to be the value of the Prokhorov metric. See Figure 2.1c for an optimal transport plan. The band between the blue diagonal lines shows the pairs of bins which are less than d_{Pr} apart, allowing mass to be moved between them at no cost. As a consequence, opposed to the picture in the Wasserstein case, almost no mass is kept in its bin. Now if one were to consider more generally $\eta = 0.8 \cdot \mathcal{N}(0.4, 0.1) + 0.2 \cdot \mathcal{N}(m, 0.02)$, for some $m \in \mathbb{R}$, the preceding discussion would apply *mutatis-mutandis*, the Wasserstein distance can increase beyond any limit when $|m|$ becomes large, whereas the Prokhorov distance remains around



(a) (b) (c)

Figure 2.1: An illustration of 1-Wasserstein and Prokhorov optimal transport distances using histograms as described in Example 2.1.9. In the 1-Wasserstein transport plan (a), most mass stays in its bin. The Prokhorov distance can be visualized via the intersection of the graphs of the Prokhorov profile and the identity (b). An optimal transport plan for the Prokhorov distance may move mass between nearby bins (between the blue lines) at no cost.

0.2 because only this amount of mass is moved across a large distance. This provides some intuition that the Prokhorov distance is more robust against outliers than Wasserstein.

In the context of data, we might not always be provided with a canonical embedding into a common ambient space; instead, we are just given an abstract finite metric space. Correspondingly, some of the topological invariants we consider are built on the finite metric structure of the data, without requiring any ambient space. In order to quantify the stability of these constructions, one needs to compare such different metric (measure) spaces, as they form the input to our pipeline. The crucial idea is to map them into a common space.

Definition 2.1.10. Let (X_1, d_1), (X_2, d_2) be two metric spaces. Their *Gromov-Hausdorff distance* is

$$d_{GH}((X_1, d_1), (X_2, d_2)) = \inf_{X_1 \xrightarrow{\varphi} Z \xleftarrow{\psi} X_2} d_H(\varphi(X_1), \psi(X_2)),$$

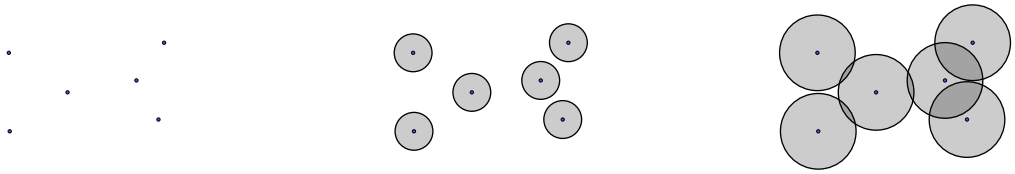
where the infimum ranges over all isometric embeddings into a common metric space (Z, d), in which the Hausdorff distance is evaluated.

We consider the analogous definitions for metric measure spaces:

Definition 2.1.11 ([2]). Let (X_1, d_1, μ_1) and (X_2, d_2, μ_2) be two metric probability spaces.

1. Their *Gromov-Prokhorov distance* [86] is

$$d_{GPr}(\mu_1, \mu_2) = \inf_{X_1 \xrightarrow{\varphi} Z \xleftarrow{\psi} X_2} d_{Pr}(\varphi_\#(\mu_1), \psi_\#(\mu_2))$$



(a) $r = 0$ (b) $r = 1$ (c) $r = 2$

Figure 2.2: The offset filtration of six points in the plane at three different values of r. For $r = 0$, we just have the points themselves. As we increase r to 1, the balls are still all disjoint, at $r = 2$ they have all merged and a hole is about to be formed on the left.

2. Their *Gromov-Hausdorff-Prokhorov distance* [2] is

$$d_{GHPr}((X_1, d_1, \mu_1), (X_2, d_2, \mu_2)) = \inf_{X_1 \xrightarrow{\varphi} Z \xleftarrow{\psi} X_2} \max\{d_H(\varphi(X_1), \psi(X_2)), d_{Pr}(\varphi_\#(\mu_1), \psi_\#(\mu_2))\},$$

where the infimum both times ranges over all isometric embeddings into a common Polish metric space (Z, d), in which the Hausdorff and Prokhorov distances are evaluated.

2.2 Filtered Spaces and Complexes

In the prototypical TDA setting, we consider a finite set of points $X \subseteq \mathbb{R}^d$. Its topology might seem uninteresting on its own at first as it is discrete. The key idea is to consider balls of increasing radius around all the points and study how the topology of this nested family of spaces changes throughout. A nested family of spaces is called a *filtration*, keeping track of topological features is a concept known as *persistence*.

We can regard the union of balls of increasing radius as a functor from the poset $[0, \infty[$ regarded as a category with a unique morphism $r \to s$ whenever $r \leqslant s$ to the category **Top** of (compactly generated weakly Hausdorff) spaces.

Definition 2.2.1. For a subset of a metric space $X \subseteq (Z, d)$, the *offset filtration* is

$$\begin{aligned} \mathcal{O}(X)\colon [0, \infty[&\to \mathbf{Top} \\ r &\mapsto \bigcup_{x \in X} \overline{B}_r(x) \\ r \leqslant s &\mapsto \bigcup_{x \in X} \overline{B}_r(x) \hookrightarrow \bigcup_{x \in X} \overline{B}_s(x). \end{aligned}$$

See Figure 2.2 for an illustration. The crucial observation now is that the offset filtration is described via sublevels of the distance-to-X function,

$$\mathcal{O}_r(X) = \{z \in Z\colon d_X(z) \leqslant r\}.$$

Now, if we are given another finite set $Y \subseteq \mathbb{R}^d$, we obtain a commutative diagram as follows for any $\delta > d_H(X, Y)$ by the definition of the Hausdorff distance:

$$\begin{array}{ccccc} \mathcal{O}(X)_r & \hookrightarrow & \mathcal{O}(X)_{r+\delta} & \hookrightarrow & \mathcal{O}(X)_{r+2\delta} \\ & \searrow\nearrow & & \searrow\nearrow & \\ \mathcal{O}(Y)_r & \hookrightarrow & \mathcal{O}(Y)_{r+\delta} & \hookrightarrow & \mathcal{O}(Y)_{r+2\delta}. \end{array} \tag{2.1}$$

Such a diagram is called a *δ-interleaving* (cf. Definition 2.2.12 below). Motivated by this example, we phrase the theory of interleavings in an abstract categorical framework in Section 2.2.1.

Before we get to that, we note the constructions presented thus far are not tractable on a computer. As a remedy, a convenient way to encode topological information in a finite-combinatorial manner is given by abstract simplicial complexes.

Definition 2.2.2. An *abstract simplicial complex* K is a collection of non-empty sets such that $\varnothing \neq \sigma \subseteq \tau$ and $\tau \in K \Rightarrow \sigma \in K$. Its elements are called *simplices*. If $\sigma \subseteq \tau \in K$, we say σ is a *face* of τ and τ is a *coface* of σ. The *dimension* of a simplex σ is $\dim(\sigma) = |\sigma| - 1$, where $|\cdot|$ denotes the cardinality. The set of d-dimensional simplices is $K^{(d)}$; the 0-dimensional simplices are called *vertices*, the set of vertices is denoted[1] K_0. Furthermore, the *k-skeleton* is

$$\mathrm{sk}^k(K) = \{\sigma \in K : \dim(\sigma) \leqslant k\}.$$

A *simplicial map* between simplicial complexes $f : K \to K'$ is a function between the vertex sets $f : K_0 \to K'_0$ such that for each $\sigma \in K$, we have $f(\sigma) \in K'$. We form the category of abstract simplicial complexes and maps, denoted **Simp**, as follows:

- Its objects are abstract simplical complexes;
- its morphisms are simplicial maps,
- composition is defined via composition of maps of vertex sets.

Geometric realization is the procedure to obtain a topological space from a combinatorial abstract simplical complex. Consider the *standard geometric n-simplex*

$$|\Delta^n| = \left\{ x = (x_1, \ldots, x_{n+1})^\top : \sum_{i=1}^{n+1} x_i = 1 \text{ and all } x_i \geqslant 0 \right\} \subseteq \mathbb{R}^{n+1}.$$

Definition 2.2.3. Let K be an abstract simplicial complex and fix a total ordering on its vertex set. The *geometric realization* of K is the topological space given by the quotient

$$|K| = \left(\coprod_{\sigma \in K} |\Delta^{\dim(\sigma)}| \right) / \sim,$$

[1] note the subtlety that $K^{(0)} = \{\{v\} : v \in K_0\}$

where the equivalence relation is generated as follows: Suppose $\sigma = [x_0, \ldots, x_n]$ is an abstract n-dimensional simplex in K with vertices in order, i.e. $x_0 < \ldots < x_n$. For any $i \in \{0, \ldots, n\}$, we identify the copy of $|\Delta^{n-1}|$ indexed by $[x_0, \ldots, \hat{x_i}, \ldots, x_n]$ (i.e. the face of σ obtained by removing x_i) under $\sim$ with the subspace of $|\Delta^n|$ where $x_i = 0$.

Important examples of simplicial complexes arise as so-called nerves, flag complexes and Dowker complexes:

Definition 2.2.4. Let $\mathcal{U} = \{U_i\}_{i \in I}$ be a cover of a topological space X. Its *nerve* $\mathrm{Nrv}(\mathcal{U})$ is the abstract simplicial complex which has $\sigma = [U_{i_0}, \ldots, U_{i_k}]$ as a k-simplex if and only if $U_{i_0} \cap \ldots \cap U_{i_k} \neq \varnothing$.

The importance of nerves is in large part due to the following result, which has a long tradition in topology. We are focusing on closed covers, following [14, Theorem D].

Theorem 2.2.5 (Nerve Theorem)**.** *Let $\mathcal{A} = \{A_i\}_{i \in I}$ be a closed cover of a compactly generated Hausdorff space X satisfying all of the following assumptions:*

i) Every finite intersection of elements of $\mathcal{A}$ is either empty or contractible.

ii) $\mathcal{A}$ is locally finite, i.e. for any point in X there is a neighborhood that only intersects finitely many of the A_i.

iii) $\mathcal{A}$ is locally finite-dimensional, i.e. for every $i \in I$ there exists $k_i \in \mathbb{N}$ such that whenever $i \in J \subseteq I$, if $\bigcap_{j \in J} A_j \neq \varnothing$ then $|J| \leqslant k_i$.

iv) If $T \subseteq I$ is such that $A_T := \bigcap_{t \in T} A_t \neq \varnothing$, then the latching space $L(T) = \bigcup_{T \subsetneq J \subseteq I} A_J$ is a closed subspace of A_T and the pair $(A_T, L(T))$ satisfies the homotopy extension property.

Then there exist a space Z and natural homotopy equivalences $|\mathrm{Nrv}(\mathcal{A})| \xleftarrow{\simeq} Z \xrightarrow{\simeq} X$.

The last condition demands some elaboration. Recall the definition what it means for a pair of spaces (X, A), where $A \subseteq X$, to have the homotopy extension property (HEP): Suppose we have $H\colon A \times [0,1] \to Y$ and $\tilde{H}_0\colon X \to Y$, where Y is any topological space, such that $H(\cdot, 0) = \tilde{H}_0|_A$. Then the HEP guaratees that there exists an extension $\tilde{H}\colon X \times [0,1] \to Y$ satisfying $\tilde{H}|_{A \times [0,1]} = H$ and $\tilde{H}(\cdot, 0) = \tilde{H}_0$. In the setting of A being a closed subspace, having the HEP is equivalent ([15, Prop. 5.13]) to (X, A) being an NDR[2]-pair, which means that there exist continnuous maps $u\colon X \to [0,1]$ and $h\colon X \times [0,1] \to X$ such that:

- $u^{-1}(0) = A$,
- $h(\cdot, 0) = \mathrm{id}_X$,
- $h(a, t) = a$ for all $a \in A$ and $t \in [0,1]$,
- $h(x, 1) \in A$ for all x such that $u(x) < 1$.

Let us put this into action for the case of Euclidean balls in the offset filtration:

[2] the abbreviation stands for *neighborhood deformation retract*

Lemma 2.2.6. *Let $X \subseteq \mathbb{R}^d$ be a finite subset. For $x \in X$, set $A_x = \overline{B}_r(x)$ and $A_I = \bigcap_{x\in I} A_x$ for $I \subseteq X$. Whenever $I \subseteq J \subseteq X$, we have an NDR-pair (A_I, A_J).*

Proof. Set $u\colon A_I \to [0,1],\ u(x) = \min\{d_{A_I}(x), 1\}$. Now, finite intersections of Euclidean balls are compact and convex, therefore each point in $x \in A_I$ has a unique closest point $p_{A_J}(x)$ in A_J, and p_{A_J} is continuous. Furthermore, note that p_{A_J} is the identity on A_J. Finally, set $h\colon A_I \times [0,1] \to A_I,\ h(x,t) = (1-t)\cdot \mathrm{id}_{A_I} + t \cdot p_{A_J}(x)$. □

The naturality of homotopy equivalences is useful in TDA in the context of filtrations. This becomes clear in the example of the offset filtration, which admits a canonical cover by the balls of radius r around the data points: The nerve of this cover is known as the Čech complex:

Definition 2.2.7. The *Čech complex* $\mathcal{C}(X)_r$ of a subset X of a metric space (Z,d) has simplices

$$X \supseteq \sigma \in \mathcal{C}(X)_r \Leftrightarrow \bigcap_{x\in\sigma} \overline{B}_r(x) \neq \varnothing \text{ and } |\sigma| < \infty.$$

It assembles into a filtration $\mathcal{C}(X)\colon [0,\infty[\to \mathbf{Simp}$,

For finite subsets of Euclidean space $X \subseteq \mathbb{R}^d$, the Čech filtration recovers the homotopy type of the offset filtration by virtue of the nerve theorem (see Figure 2.3a). This is because Euclidean space satisfies the following condition by [14, Cor. 5.16] and Lemma 2.2.6.

Definition 2.2.8 (adapted from [27, Definition 1.3]). A metric space is called *good* if any finite set of closed balls satisfies the assumptions of Theorem 2.2.5.

This is not just true for any fixed r, but actually an equivalence of filtrations, which we will define next (Definition 2.2.9). That notion will allow us to conclude that the interleaving diagram of the offset filtration gives rise to an interleaving diagram of the simplicial homology of the associated Čech complexes, which is what we study in Section 2.3.

2.2.1 Interlude: A Categorical Perspective on Filtrations and Interleavings

As we have seen in the motivating example of the offset filtration and the Čech complex, the language of functors and commutative diagrams appears quite naturally in the study of persistence. More variants of filtrations and interleavings appear throughout this thesis (and the TDA literature in general), therefore it is handy to set up an abstract categorical framework to handle them all. Most of this material is adapted from [27, Section 2.5]

The offset filtration is indexed by a non-negative real parameter r, which we visually think of as the radius of the balls centered at the data points. Categorically speaking, we regard $[0,\infty[$ as a category: its objects are the non-negative real numbers and there is a unique morphism $r \to s$ whenever $r \leqslant s$. Reflexivity of the $\leqslant$-relation gives rise to identity morphisms; transitivity defines composition in this category. Note that this recipe turns any poset into a category, we have not used anything specific to $[0,\infty[$. Later on, we will consider two-parameter filtrations, also called *bifiltrations*. Instead of the poset $[0,\infty[$ we will consider the cartesian product $]0,\infty[^{op} \times [0,\infty[$. These, too, are interpreted as categories arising from a poset, where the partial order is given by $(m,r) \leqslant (m',r')$ if and only if $m \geqslant m'$ and $r \leqslant r'$. We will write T to denote any of the categories given by the posets $[0,\infty[$ or $]0,\infty[^{op}$ or $\mathbb{R}^{op}$ or $]0,\infty[^{op} \times [0,\infty[$ or $\mathbb{R}^{op} \times [0,\infty[$.

Definition 2.2.9. If $\mathbf{C} \in \{\mathbf{Top}, \mathbf{Simp}\}$, a ($T$-indexed) *filtration* is a functor $F\colon T \to \mathbf{C}$ such that all morphisms in T get mapped to inclusions in $\mathbf{C}$. We also call $F : T \to \mathbf{Top}$ a *(bi)filtered space* and $F\colon T \to \mathbf{Simp}$ a *(bi)filtered complex*.

Two filtrations $F, F'\colon T \to \mathbf{Top}$ are said to be *objectwise equivalent* if there is a natural transformation $\eta\colon F \Rightarrow F'$ such that all components $\eta_t\colon F_t \xrightarrow{\simeq} F'_t$ are homotopy equivalences. Furthermore, F, F' are called *weakly equivalent*, written $F \simeq F'$, if they are connected via a zig-zag of objectwise equivalences. That is, there is a sequence of filtrations

$$F = F^0, F^1, \ldots, F^n = F',$$

such that for each $i \in \{1, \ldots, n\}$ there is an objectwise equivalence $\eta_i\colon F^{i-1} \Rightarrow F^i$ or $\eta_i\colon F^i \Rightarrow F^{i-1}$.

This kind of equivalence is what the functorial nerve lemma guarantees.

When dealing with functors defined on a category given by a poset $F\colon T \to \mathbf{C}$, we will write $F_t := F(t)$ and $F_{s \leqslant t} := F(s \leqslant t)$, where $s, t \in T$. An important example of filtrations is by sublevelsets.

Definition 2.2.10. Let X be a topological space, T a partially ordered set and $f\colon X \to T$ be any function. The *sublevel filtration* is

$$\mathsf{S}^{\downarrow}(f)\colon T \to \mathbf{Top},\ \mathsf{S}^{\downarrow}(f)_r = \{x \in X\colon f(x) \leqslant r\}.$$

Indeed, observe that $\mathcal{O}(X) = \mathsf{S}^{\downarrow}(d_X)$, the offset filtration is the sublevel filtration of the distance-to-X function.

If a function $f\colon K \to T$ defined on a simplicial complex K is monotonic (using the face poset), it gives rise to a sublevel filtration in the analogous way.

There is a standard way to compare two functors defined on a common poset via interleavings, as introduced by Chazal [43], generalizing the diagram (2.1) we observed for the offset filtration.

Definition 2.2.11. Let T be any of the posets $[0, \infty[, \mathbb{R}^{op}, \mathbb{R}^{op} \times [0, \infty[$. A poset automorphism $\alpha\colon T \to T$ is called a *forward shift* if $t \leqslant \alpha(t)$ for all $t \in T$.

The most important example will be the *δ-shift* for $\delta > 0$, this is

$$\begin{aligned}
[\delta]\colon [0, \infty[&\to [0, \infty[, & r &\mapsto r + \delta,\\
[\delta]\colon \mathbb{R}^{op} &\to \mathbb{R}^{op}, & r &\mapsto r - \delta,\\
[\delta]\colon \mathbb{R}^{op} \times [0, \infty[&\to \mathbb{R}^{op} \times [0, \infty[, & (m, r) &\mapsto (m - \delta, r + \delta).
\end{aligned}$$

Note that we cannot define the δ-shift only on $]0, \infty[^{op}$ because it would be undefined for $t \leqslant \delta$. Focusing on $T = [0, \infty[$ and a δ-shift $\alpha = [\delta]$, we write $F \circ \alpha = F[\delta]$. Explicitly, we have $(F[\delta])_t = F_{t+\delta}$, and the δ-shift turns into an endofunctor of the functor category $\mathbf{C}^{[0,\infty[}$ as

$$f\colon F \to G \text{ induces } f[\delta]\colon F[\delta] \to G[\delta] \text{ via } f[\delta]_t = f_{t+\delta}.$$

In words, we shift the non-negative real line by δ before applying F, explaining the name. Moreover, the maps $F_{t \leqslant t+\delta}\colon F_t \to F_{t+\delta} = F[\delta]_t$ assemble into a morphism in the functor category (i.e., a

natural transformation) $\varphi_F^\delta \colon F \to F[\delta]$ called the *canonical shift map* . In the example of the offset filtration, the canonical shift map is simply the inclusion of the balls of radius r inside the balls of radius $r + \delta$.

The following abstract definition encapsulates both the one- and the two-parameter setting.

Definition 2.2.12. Let (T, T') be either of the following pairs of posets:

$$([0, \infty[, [0, \infty[); \qquad (\mathbb{R}^{op} \times [0, \infty],]0, \infty[^{op} \times [0, \infty[).$$

Let $\alpha, \beta \colon T \to T$ be forward shifts. The (α, β)-*interleaving category* of T' is $\mathbf{I}(T', \alpha, \beta)$ and has $T' \times \{0, 1\}$ as objects and morphisms $(r, i) \to (s, j)$ if and only if either

- $i = j$ and $r \leqslant s$,
- $i = 0, j = 1$ and $\alpha(r) \leqslant s$,
- $i = 1, j = 0$ and $\beta(r) \leqslant s$.

Composition in $\mathbf{I}(T', \alpha, \beta)$ is defined by the requirement that between any two objects, there is at most one morphisms (i.e. the category is thin). We can include T' in $\mathbf{I}(T', \alpha, \beta)$ in two ways, namely $E_i \colon T' \to T' \times \{i\}$ via the identity on T' for $i \in \{0, 1\}$. Given tow functors $F, G \colon T' \to \mathbf{C}$, an (α, β)-interleaving is a functor $Z \colon \mathbf{I}(T', alpha, \beta)$ such that

$$F = Z \circ E_0 \text{ and } G = Z \circ E_1.$$

Again restricting to the special case of δ-shifts, if F, G are $([\delta], [\delta])$-interleaved, we say they are δ interleaved, for short. In the one-parameter setting, a δ-interleaving amounts to saying that there are natural transformations

$$f \colon F \to G[\delta], \ g \colon G \to F[\delta]$$

such that

$$g[\delta] \circ f = \varphi_F^{2\delta} \text{ and } f[\delta] \circ g = \varphi_G^{2\delta}.$$

Moreover, look back at Diagram 2.1, which we can now interpret as a functor from the interleaving category $\mathbf{I}([0, \infty[, [\delta], [\delta]$. With this at hand, we can define a distance:

Definition 2.2.13. Let $F, G \colon T \to \mathbf{C}$. Their *interleaving distance* is

$$d_I(F, G) = \inf\{\delta > 0 \colon F, G \text{ are } \delta\text{-interleaved}\}.$$

Observe that in case of f, g being mutually inverse isomorphisms, one obtains an interleaving distance of 0. Thus, one can intuitively think of this distance as measuring how far from isomorphic two functors are, although there is a caveat: While this construction is symmetric and satisfies the triangle inequality (because the composition of a δ-shift with an ε-shift is a $(\delta + \varepsilon)$-shift), there can be non-isomorphic filtrations at interleaving distance zero. For instance, if we replacee closed by open balls in the offset filtration, we get a different filtration which is δ-interleaved with the original one for every $\delta > 0$. We saw an example of an interleaving in diagram (2.1): the union of closed r-balls around X is contained in the union of closed $(r + \delta)$-balls around Y when $\delta > d_H(X, Y)$. As the offset filtration is weakly equivalent to the Čech filtration, the next definition arises naturally:

Definition 2.2.14 ([27, Definition 2.37]). Let $F, G\colon \mathbb{R}^d \to \mathbf{Top}$. Their *homotopy interleaving distance* is

$$d_{HI}(F,G) = \inf\{\delta > 0 \colon \exists F' \simeq F,\ G' \simeq G \text{ such that } F', G' \text{ are } \delta - \text{interleaved}\},$$

where $\simeq$ denotes a weak equivalence of functors $T \to \mathbf{Top}$ (Definition 2.2.9).

The following kind of Lipschitz-continuity is a prototypical example of what is called a *stability result* in TDA literature. We put all the effort in phrasing the preceding constructions in the language of category theory so that the proof is now very easy.

Proposition 2.2.15. *Let X be a topological space or a simplicial complex, let $f, g\colon X \to T$ be functions (monotonic[3] ones, if X is a simplicial complex). Then*

$$d_I(\mathsf{S}^{\downarrow}(f), \mathsf{S}^{\downarrow}(g)) \leqslant \sup_{x \in X} \|f(x) - g(x)\|_\infty.$$

Proof. By assumption, we have inclusions

$$\mathsf{S}^{\downarrow}(f) \hookrightarrow \mathsf{S}^{\downarrow}(g)[\delta],\ \mathsf{S}^{\downarrow}(g) \hookrightarrow \mathsf{S}^{\downarrow}(f)[\delta].$$

Indeed, if $f(x) \leqslant t$ then $g(x) \leqslant t + \delta$ in the one-parameter case and $g(x) \leqslant t + (-\delta, \delta)^\top$ in the two-parameter case (and analogously for the roles of f and g exchanged). As the interleaving maps are inclusions, they compose to form the structural inclusions; explicitly, the following diagrams commute:

$$\begin{array}{ccc} \{f(x) \leqslant t\} & \hookrightarrow & \{g(x) \leqslant t + \delta\} \\ & \searrow & \downarrow \\ & & \{f(x) \leqslant t + 2\delta\} \end{array} \qquad \begin{array}{ccc} \{g(x) \leqslant t\} & \hookrightarrow & \{f(x) \leqslant t + \delta\} \\ & \searrow & \downarrow \\ & & \{g(x) \leqslant t + 2\delta\} \end{array},$$

where in the two-parameter setting, $\leqslant$ is understood as the product partial order and we use shorthand notation $\delta := (-\delta, \delta)$. □

This stability result of sublevel filtrations goes back to [52].

2.2.2 One-Parameter Filtrations

We are now going to give more examples of filtrations that appear later on in this thesis.

First, recall that taking the sublevel sets of a distance to a point cloud yields the offset filtration (Definition 2.2.1), a union of balls:

$$\mathcal{O}(X) = \mathsf{S}^{\downarrow}(d_X).$$

Hence, given two finite subsets $X, Y \subset \mathbb{R}^d$, Proposition 2.2.15 yields the inequality in the following line,

$$d_I(\mathcal{O}(X), \mathcal{O}(Y)) \leqslant \sup_{z \in \mathbb{R}^d} |d_X(z) - d_Y(z)| = d_H(X, Y),$$

[3] with respect to the partial orders, given by inclusion of simplices in the complex on the domain, and by the cartesian product of the $\leqslant$-orders on the codomain

whereas the equality is the very definition of the Hausdorff distance (Definition 2.1.1). This estimate is a rigorous way to say that small perturbations in the input point clouds only lead to small perturbations in the offset filtration.

For computations, it is beneficial to further reduce the size of the complex. To this end, one introduces a classical notion from computational geometry, the Voronoi tesselation.

Definition 2.2.16. Let $X \subset \mathbb{R}^d$ be a finite subset of Euclidean space. The *Voronoi cell* of a point $x \in X$ is

$$\mathrm{Vor}_X(x) = \{y \in \mathbb{R}^d \colon \forall x \neq x' \in X \colon d(x,y) \leqslant d(x',y)\}.$$

In words, the Voronoi cell of x consists of those points for which there is no other point in X that is closer.

Definition 2.2.17. Let $X \subset \mathbb{R}^d$ be a finite subset of Euclidean space. The *Alpha complex* [68] $\mathcal{A}(X)_r$ of X has simplices (we do not have to impose an additional finiteness condition as X is already finite)

$$X \supseteq \sigma \in \mathcal{A}(X)_r \Leftrightarrow \bigcap_{x\in\sigma} \left(\overline{B}_r(x) \cap \mathrm{Vor}_X(x)\right) \neq \varnothing;$$

this gives rise to a filtration $\mathcal{A}(X) \colon [0,\infty[\to \mathbf{Simp}$.

Again, an application of the nerve theorem shows that

$$|\mathcal{A}(X)_r| \simeq \bigcup_{x\in X} \left(\overline{B}_r(x) \cap \mathrm{Vor}_X(x)\right) = \bigcup_{x\in X} \overline{B}_r(x) \simeq |\mathcal{C}(X)_r|;$$

and again this is not just for fixed r but an equivalence of filtrations.

As a consequence, we obtain stability of Čech and Alpha filtrations in the homotopy interleaving distance: If $X, Y \subset \mathbb{R}^d$ are finite subsets, we have

$$d_{HI}(\mathcal{C}(X), \mathcal{C}(Y)) \leqslant d_H(X,Y) \text{ and } d_{HI}(\mathcal{A}(X), \mathcal{A}(Y)) \leqslant d_H(X,Y).$$

A second important recipe for constructing simplicial complexes are *flag complexes* also known as *clique complexes*.

Definition 2.2.18. Let $G = (V,E)$ be a graph. Its *clique complex* has k-simplices equal to the $k+1$ cliques (i.e. complete subgraphs) of G.

Definition 2.2.19. A simplicial complex is a *flag complex* if it is the clique complex of its 1-skeleton.

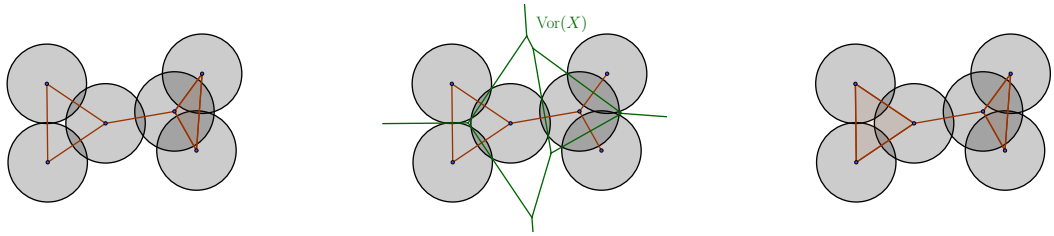
Definition 2.2.20. Let (X,d) be a metric space. Its *Vietoris-Rips complex* is the abstract simplicial complex $\mathcal{R}(X)_r$ with simplices

$$X \supseteq \sigma = [x_0, \dots x_k] \in \mathcal{R}(X)_r \Leftrightarrow d(x_i, x_j) \leqslant r;$$

this gives rise to a functor $\mathcal{R}(X) \colon [0,\infty[\to \mathbf{Simp}$.

If $X \subseteq \mathbb{R}^d$ is a finite subset, one easily observes the following relation between Čech and Vietoris-Rips complexes:

$$\mathcal{C}(X)_r \subseteq \mathcal{R}(X)_{2r}, \qquad VR(X)_r \subseteq \mathcal{C}(X)_r;$$



(a) Čech (b) Alpha (c) Vietoris-Rips

Figure 2.3: Three different constructions of filtered simplicial complexes with a fixed sample as vertex set: Čech and Alpha at scale r, and Vietoris-Rips at scale $2r$ (from left to right). Observe how Alpha and Čech capture the topology of the union of balls, which is the offset filtration.

indeed there are even stronger, dimension-dependent, bounds known [58, Theorem 2.5]. The above inclusions constitute the first instance (in this thesis) of an interleaving for which the shifts are not just given by δ-shifts.

While we cannot simply appeal to the nerve theorem, Vietoris-Rips complexes still satisfy a stability theorem:

Theorem 2.2.21 ([25, 43])**.** *For finite metric spaces X, Y, we have*

$$d_{HI}(\mathcal{R}(X), \mathcal{R}(Y)) \leqslant d_{GH}(X, Y).$$

Example 2.2.22. Figure 2.3 illustrates Čech and Alpha (at a common filtration value r) and Rips (at scale $2r$) complexes drawn in brown on a fixed point cloud. Observe their differences: The left triangle is filled in the Rips complex, because it is a flag complex complex and all the sides are present. The three sides are also present in the Alpha and Čech complexes, because the intersection of the r-balls is non-empty and even meets the Voronoi edge. However, there is no threefold intersection among the three balls on the left, whence that triangle is missing here. There is a filled triangle on the right in the Rips complex. This is also filled in the Čech complex because the r balls not just intersect pairwise, but there is actually a non-empty triple intersection. However, this triple intersection does not meet the Voronoi diagram; neither does the intersection of the top and the bottom ball. Therefore, the Alpha complex is a proper subcomplex of the Čech complex, missing the rightmost edge and the 2-simplex. Note that Alpha and Čech complex are, of course, homotopy equivalent.

The third prototypical construction of simplicial complexes is due to Dowker [65].

Definition 2.2.23. Given two sets X, Y, a *relation* $R \subseteq X \times Y$ is a subset of the product. The category **Rel** has triples (X, Y, R) as objects, where $R \subseteq X \times Y$. Its morphisms are

$$f = (f_X, f_Y) \colon (X, Y, R) \to (X', Y', R'),$$

where $f_X \colon X \to X'$ and $f_Y \colon Y \to Y'$ are maps such that $(x, y) \in R$ implies $(f_X(x), f_Y(y)) \in R'$. Composition is defined component-wise.

We will usually identify a relation with its indicator matrix, which is a binary matrix with row labels given by X and column labels Y. The entry at (x, y) is 1 if $(x, y) \in R$ and 0 otherwise.

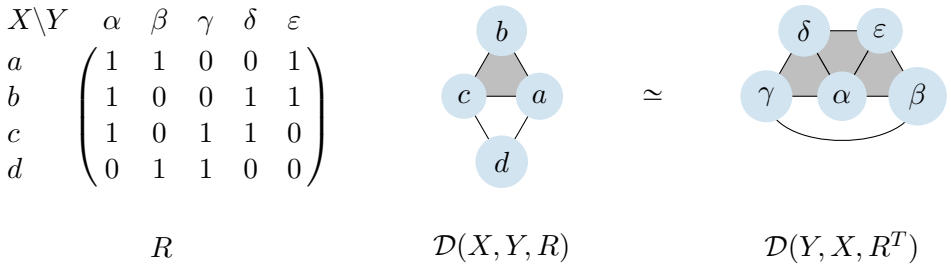


Figure 2.4: The indicator matrix representing a binary relation R (left); its Dowker complex $\mathcal{D}(X, Y, R)$, whose vertices are the row labels (middle); and the Dowker complex of the dual relation $\mathcal{D}(Y, X, R^\top)$, which has the same homotopy type.

Definition 2.2.24. Let X, Y be sets and $R \subseteq X \times Y$ a relation. The *Dowker complex* of the relation is the abstract simplicial complex $\mathcal{D}(X, Y, R)$ whose simplices are the nonempty finite subsets $\sigma \subseteq X$ that satisfy

$$\exists y \in Y : \sigma \times \{y\} \subseteq R.$$

If $\sigma \times \{y\} \subseteq R$, we say y is a *witness* of (or: *witnesses*) σ.

Definition 2.2.25. Let $R \subseteq X \times Y$ be a relation; we denote by $R^\top \subseteq (Y, X)$ its *transpose*, that is,

$$(y, x) \in R^\top \Leftrightarrow (x, y) \in R.$$

If $f = (f_X, f_Y) \colon (X, Y, R) \to (X', Y', R')$ is a map, we get a *transposed map*

$$f^\top = (f_Y, f_X) \colon (Y, X, R^\top) \to (Y', X', (R')^\top).$$

Taking the transpose of the indicator matrix of a relation R, we obtain the corresponding matrix representing the transpose relation.

The following theorem is originally due to Dowker [65], who proved it only in terms of homology equivalences; the version here pertaining to homotopy equivalences is due to Björner [23, Theorem 10.9].

Theorem 2.2.26 (Dowker duality)**.** *Let $R \subseteq X \times Y$ be a relation and denote by $R^\top \subseteq Y \times X$ its transpose. Then we have a homotopy equivalence*

$$|\mathcal{D}(X, Y, R)| \simeq |\mathcal{D}(Y, X, R^\top)|.$$

Example 2.2.27. Consider $X = \{a, b, c, d\}$, $Y = \{\alpha, \beta, \gamma, \delta, \varepsilon\}$ and $R \subseteq X \times Y$, which we equivalently represent as a binary matrix $R \in \{0, 1\}^{X \times Y}$ as indicated in Figure 2.4, left panel. The definition of the Dowker complex unfolds as follows: We build a simplicial complex on the vertex set X, in which we add a simplex $\sigma \subseteq X$ if there is a column $y \in Y$ which contains σ; for instance, we introduce the simplex $\{a, b, c\}$ as it is contained in column α. This Dowker complex $\mathcal{D}(X, Y, R)$ is shown in the middle panel of Figure 2.4. On the other hand, the Dowker complex of the transpose relation, shown on the right in Figure 2.4, has vertices Y. For instance we introduce the simplex $\{\alpha, \delta\}$ because it is contained in row b; it is also in row c, but in no other row. In other words, the rows b

and c are witnesses of the simplex $\{\alpha, \delta\}$. Observe that $\mathcal{D}(X, Y, R)$ and $\mathcal{D}(Y, X, R^\top)$ are homotopy equivalent.

Recently, motivated by persistent homology as an invariant of filtered complexes, functorial extensions of Dowker duality have been established, starting from [50], then later [156] and recently [37].

Proposition 2.2.28 ([37, Theorem 5.2])**.** *Dowker complexes and Dowker Duality are functorial in the following sense: Any morphism of relations $f = (f_X, f_Y)\colon (X, Y, R) \to (X', Y', R')$ induces a simplicial map $\mathcal{D}(f)\colon \mathcal{D}(X, Y, R) \to \mathcal{D}(X', Y', R')$; these assemble into a functor $\mathcal{D}\colon$ **Rel** $\to$ **Simp**. In addition, one can choose homotopy equivalences*

$$\Psi_R\colon |\mathcal{D}(X, Y, R)| \to |\mathcal{D}(Y, X, R^\top)|, \qquad \Psi_{R'}\colon |\mathcal{D}(X', Y', R')| \to |\mathcal{D}(Y', X', (R')^\top)|$$

such that the following diagram commutes up to homotopy:

$$\begin{array}{ccc} |\mathcal{D}(X, Y, R)| & \xrightarrow{\Psi_R} & |\mathcal{D}(Y, X, R^\top)| \\ \downarrow{\scriptstyle |\mathcal{D}(f)|} & & \downarrow{\scriptstyle |\mathcal{D}(f^\top)|} \\ |\mathcal{D}(X', Y', R')| & \xrightarrow{\Psi_{R'}} & |\mathcal{D}(Y', X', (R')^\top)| \end{array}$$

In particular, a filtration of relations gives rise to a filtration of Dowker complexes in a way compatible with Dowker duality. The focus for us will be on relations that arise as sublevel sets of some function, i.e.

$$R_r = \{(x, y)\colon \Lambda(x, y) \leqslant r\}, \text{ where } \Lambda\colon X \times Y \to [0, \infty[.$$

The most important example is X, Y being subsets of some metric space (Z, d) and $\Lambda = d|_{X\times Y}$ being the restriction of the metric. Taking $Y = Z$ in this setting recovers the Čech complex of X.

Example 2.2.29. We can view the Čech complex of some finite subset X of an ambient metric space (Z, d) as Dowker complex via $\mathcal{C}(X)_r = \mathcal{D}(X, Z, R_r)$. Indeed, let us check that they have the same simplices:

$$\begin{aligned} \sigma \in \mathcal{D}(X, Z, R_r) &\Leftrightarrow \exists z \in Z\colon (x, z) \in R_r \text{ for all } x \in \sigma \\ &\Leftrightarrow \exists z \in Z\colon d(x, z) \leqslant r \text{ for all } x \in \sigma \\ &\Leftrightarrow \exists z \in Z\colon z \in \overline{B}_r(x) \text{ for all } x \in \sigma \\ &\Leftrightarrow \bigcap_{x\in\sigma} \overline{B}_r(x) \neq \varnothing \\ &\Leftrightarrow \sigma \in \mathcal{C}(X)_r. \end{aligned}$$

Inspecting Figure 2.3a, we observe that the set of witnesses of a k-simplex is the intersection of the r-balls around the corresponding $k + 1$ points. We will impose the mass of these intersections as a second filtration parameter in chapter 7.

Similarly, for a finite metric space (X, d), the intrinsic Čech complex is the Dowker complex $\mathcal{I}(X)_r = \mathcal{D}(X, X, R_r)$, see for instance [36].

A stability result for Dowker complexes is given in [45, Lemma 4.9]; it recovers Čech stability in the previous example.

While the constructions presented thus far capture metric information, they are not robust to outliers and insensitive with respect to density of the underlying point cloud. One way to remedy this is the via the distance-to-measure (DTM) filtration [5].

Definition 2.2.30. Let $X \subseteq \mathbb{R}^d$ be a finite set with associated empirical measure μ_X; let $m \in]0, 1[$ a parameter. The *distance to measure* μ_X with parameter m is the function

$$\begin{aligned} d_{\mu_X,m} &\colon \mathbb{R}^d \to [0, \infty[, \\ d_{\mu_X,m}(z) &= \sqrt{\frac{1}{\lceil m|X| \rceil} \sum_{p \in \mathrm{NN}_C^{\lceil m|X| \rceil}(z)} \|p - z\|_2^2}, \end{aligned}$$

where $\mathrm{NN}_X^{\lceil m|X| \rceil}(z)$ denotes the $\lceil m|X| \rceil$ nearest neighbors of z in X.

In words, the distance to measure is an averaged distance to the nearest data points. We use this function to, intuitively speaking, let the balls around the data points grow faster in high density regions.

Definition 2.2.31. Let X, μ_X and m as before, let $x \in X$. Consider the function

$$r_x \colon [0, \infty[\to [0, \infty[, \qquad r_x(t) = t - d_{\mu_X,m}(x),$$

The *weighted Rips complex* $\mathcal{V}(X, m)_t$ at scale t is the clique complex of the graph $G(X, m)_t = (V_t, E_t)$, with vertices and edges given by

$$\begin{aligned} V_t &= \{x \in X \colon r_x(t) \geqslant 0\}, \\ E_t &= \{\{x, y\} \colon r_x(t) + r_y(t) \geqslant 2\|x - y\|_2\}. \end{aligned}$$

Again, increasing t turns the weighted Rips complexes into a filtration. Therein, a point only appears once t surpasses the average distance of the point to its $\lceil m|X| \rceil$ nearest neighbors. Outliers and points in low density regions thus appear only later in the filtration. Note that this construction depends on a suitable choice of the parameter m. In the spirit of persistence, the question arises what happens if we let m increase from 0 to 1 instead of fixing it.

2.2.3 Two-Parameter Filtrations

We have seen that while the one-parameter filtrations presented here are *stable* in the sense that they do not change much under small perturbations of the input data in terms of the (Gromov-)Hausdorff distance, they have certain shortcomings: They are insensitive to the density of the data, and in particular are not *robust* to the presence of outliers. Trying to account for density in one-parameter filtrations leads to the introduction of a new parameter which controls the mass. Instead of fixing such a parameter, we now want to study the situation in which this turned into a second filtration parameter. See [33] for a survey of ideas in this direction; most material of this subsection is adapted from [27] and Michael Lesnick's lecture notes [107].

Recall the example of the offset filtration $\mathcal{O}(X)$ associated to a finite set of points $X \subseteq \mathbb{R}^d$, which is given by a union of r-balls centered at elements of X (Definition 2.2.1). In high density regions,

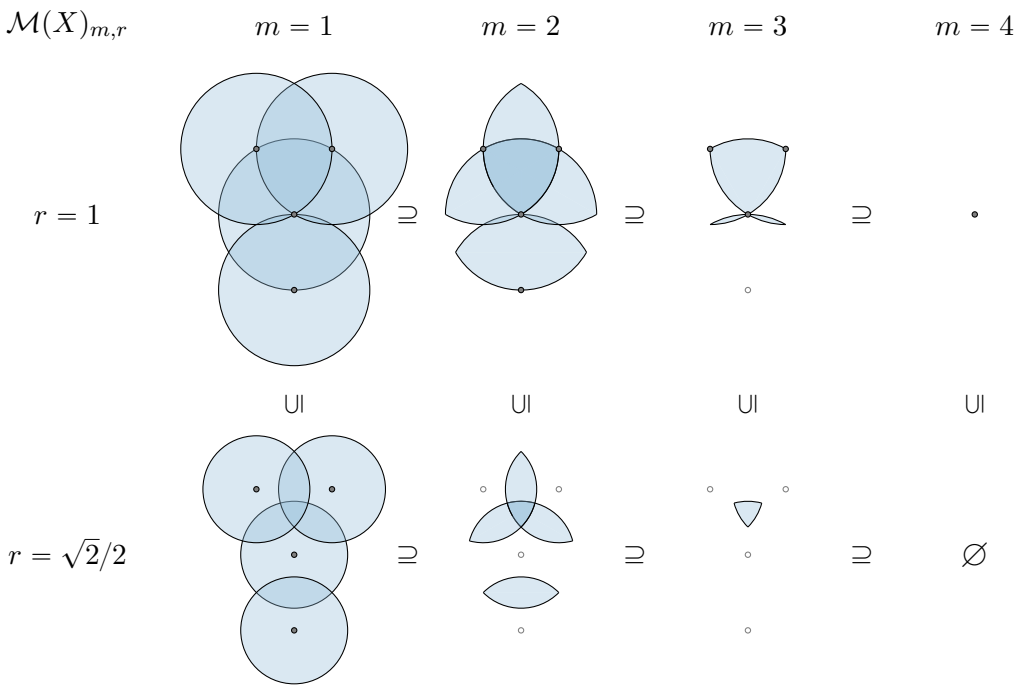


Figure 2.5: The multicover filtration of a covering by closed r-balls centered a four points in the Euclidean plane.

where many points of X are close together, we would intuitively expect a point $z \in \mathcal{O}(X)_r$ to be covered by a lot of such balls, whereas in low density regions it would perhaps only be covered by a single one. However, the homotopy type of $\mathcal{O}(X)_r$ (or, equivalently, the Čech filtration $\mathcal{C}(X)_r$) is insensitive to this multiplicity of coverings. In fact, if X is not a subset, but a multiset (repeating data points in a sample are not unheard of in practice), one would like to adapt the definition of the offset filtration to account for this fact. This motivates the following definition:

Definition 2.2.32 ([143]). Let Z be a topological space and let $\mathcal{U}$ be a cover (which may contain repeated elements), define the *multicover filtration of* $\mathcal{U}$ as

$$\begin{aligned}\mathcal{M}(\mathcal{U})\colon\,]0,\infty[^{op} &\to \mathbf{Top},\\ \mathcal{M}(\mathcal{U})_m &= \{z \in Z\colon\, z \text{ is contained in at least } m \text{ elements of the cover } \mathcal{U}\}.\end{aligned}$$

If X is a finite subset of an ambient metric space (Z, d), the *multicover bifiltration of* X is, at fixed scale r, the multicover filtration induced by the covering given by closed r-balls:

$$\begin{aligned}\mathcal{M}(X)\colon\,]0,\infty[^{op}\times[0,\infty[&\to \mathbf{Top},\\ \mathcal{M}(X)_{m,r} &= \{z \in Z : d(z,x) \leqslant r \text{ for at least } m \text{ elements } x \in X\}.\end{aligned}$$

Example 2.2.33. Consider X to be the four points $\{(0,0),(-1,0),(\frac{1}{2},\frac{\sqrt{3}}{2}),(-\frac{1}{2},\frac{\sqrt{3}}{2})\}$ in the Euclidean plane depicted in Figure 2.5 and let us look at its multicover bifiltration. In the top row, we

fix $r = 1$ and in the bottom row, we fix $r = \sqrt{2}/2$. The multicover bifiltration of X restricted to r in the second parameter is equivalently given by the multicover filtration of the covering by closed r-balls. At given m it contains those points in $\mathbb{R}^2$ within distance r of at least m points of X. The leftmost columns shows $m = 1$; this restriction of the multicover bifiltration is just the offset filtration. Then for $m = 2$ in the second column, we get all the points covered by at least two r-balls. In the third column, when $m = 3$, we obtain the threefold intersection of the balls, which consists of all points that are within distance r of three points in X. Finally, in the rightmost column, we have $m = 4$, which is empty for $r = \sqrt{2}/2$ and consists of just a single point for $r = 1$. Note that the data points themselves do not belong to the multicover bifiltration for every pair of parameters.

The key idea to account for multiplicities in X is to consider the associated counting measure $\mu_x = \sum_{x \in X} \delta_x$. If a point $x \in X$ now repeats M_x times, we can simply look at the measure $\sum_{x \in X} M_x \delta_x$. With this in mind, we recast the multicover bifiltration of a finite set of points X as

$$\begin{aligned}\mathcal{M}(X)_{m,r} &= \{z \in Z : d(z,x) \leqslant r \text{ for at least } m \text{ elements } x \in X\} \\ &= \{z \in Z : |X \cap \overline{B}_r(z)| \geqslant m\} \\ &= \{z \in Z : \mu_X(\overline{B}_r(z)) \geqslant m\}.\end{aligned}$$

Thus, the multicover bifiltration naturally generalizes to arbitrary measures as follows:

Definition 2.2.34 ([27]). Let (Z, d) be a Polish space and μ be a Borel measure on it. The *measure bifiltration* is

$$\begin{aligned}\mathcal{B}(\mu)\colon]0,\infty[^{op} \times [0,\infty[&\to \mathbf{Top}, \\ \mathcal{B}(\mu)_{m,r} &= \{z \in Z\colon \mu(\overline{B}_r(z)) \geqslant m\}.\end{aligned}$$

In particular, this bifiltration handles points with multiplicities as described above. This general formulation is useful to prove stability and robustness results (as we shall see later on), but not for computations. Just like the offset filtration admits a combinatorial model in form of the Čech filtration, the multicover filtration is weakly equivalent to the so-called subdivision Čech bifiltration [27, Theorem 3.3 (i)]. This is defined as follows:

Definition 2.2.35 ([143]). Let K be any simplicial complex and denote its barycentric subdivision by $\mathrm{Sd}(K)$. A k-simplex in $\mathrm{Sd}(K)$ is given by an ascending chain $\sigma_0 \subsetneq \ldots \subsetneq \sigma_k$ (called a *flag*) of simplices in K. The *subdivision filtration* $\mathcal{S}(K)$ at index m is given by the complex whose simplices are flags in which the minimal dimension is at least $m - 1$,

$$\begin{aligned}\mathcal{S}(K)\colon]0,\infty[^{op} &\to \mathbf{Simp}, \\ \mathcal{S}(K)_m &= \{(\sigma_0 \subsetneq \ldots \subsetneq \sigma_k)\colon \dim(\sigma_0) \geqslant m-1\} \subseteq \mathrm{Sd}(K);\end{aligned}$$

where $\mathbf{Simp}$ is the category of abstract simplicial complexes and simplicial maps.

Let (Z, d) be a Polish space and X a finite subset. The *subdivision Čech bifiltration* is

$$\begin{aligned}\mathcal{SC}(X)\colon]0,\infty[^{op} \times [0,\infty[&\to \mathbf{Simp}, \\ \mathcal{SC}(X)_{m,r} &= \mathcal{S}(\mathcal{C}(X)_r)_m.\end{aligned}$$

Let X be a finite metric space. Its *subdivision Rips bifiltration* is

$$\begin{aligned}\mathcal{SR}(X)\colon]0,\infty[^{op}\times[0,\infty[&\to \mathbf{Simp},\\ \mathcal{SR}(X)_{m,r} &= \mathcal{S}(\mathcal{R}(X)_r)_m.\end{aligned}$$

The equivalence between multicover and subdivision bifiltrations for good metric spaces is established by the following theorem, whose statement is rather technical – see [27, section 4] for a more thorough discussion. We are going to need it in Chapter 7, hence it we state it for the sake of being self-contained. Recall that T denotes a small category associated to a poset, as described at the start of Section 2.2.1.

Theorem 2.2.36 (Multicover Nerve Theorem, [143], [41],[27, Theorem 4.12 and Remark 4.13]; see also [14])**.** *Given a filtration $F\colon T \to \mathbf{Top}$ of compactly generated spaces, suppose we have a set $\mathcal{U}$ of functors $T \to \mathbf{Top}$ such that*

i) for every $t \in T$, the set $\{U_t\colon U \in \mathcal{U}\}$ is a closed cover of F_t such that every finite non-empty intersection is weakly homotopy equivalent to a point and satisfying the conditions i)-iii) of Theorem 2.2.5,

ii) for every $U \in \mathcal{U}$ and every $s \leqslant t \in T$, the map $U_{s\leqslant t}$ is the restriction of $F_{s\leqslant t}$ to U_s.

Then we have a weak equivalence of filtrations $\mathcal{M}(\mathcal{U}) \simeq |\mathcal{S}(\mathrm{Nrv}(\mathcal{U}))|$, where

$$\begin{aligned}\mathcal{M}(\mathcal{U}), |\mathcal{S}(\mathrm{Nrv}(\mathcal{U}))|\colon]0,\infty[^{op}\times T &\to \mathbf{Top},\\ \mathcal{M}(\mathcal{U})\colon (m,t) &\mapsto \mathcal{M}(\{U_t\colon U\in\mathcal{U}\})_m,\\ |\mathcal{S}(\mathrm{Nrv}(\mathcal{U}))|\colon (m,t) &\mapsto |\mathcal{S}(\mathrm{Nrv}(\{U_t\colon U\in\mathcal{U}\}))_m|.\end{aligned}$$

As an example, note that for a finite set $X \subseteq \mathbb{R}^d$, the offset filtration $\mathcal{O}(X)$ can be taken as F in the theorem. In this setting, the set of functors giving covers is indexed by the poset $[0,\infty[$ and we have $\mathcal{U} = \{\overline{B}_\bullet(x)\}_{x\in X}$, where

$$\overline{B}_\bullet(x)\colon [0,\infty[\to \mathbf{Top}, \qquad r \mapsto \overline{B}_r(x).$$

Let us illustrate this by continuing Example 2.2.33.

Example 2.2.37. Recall the setting of Example 2.2.33, in which we described the multicover filtration of a set of four points in the Euclidean plane. We repeat a part of that picture in the top half of Figure 2.6. A combintorial model is given by the subdivision Čech bifiltration, which is shown in the bottom row.

However, due to the appearance of barycentric subdivisions, these complexes are usually intractable for use in computations. Instead, one often considers the following subcomplexes of the Rips filtration, although there are other options like the rhomboid bifiltration [54, 71], which is equivalent to the multicover bifiltration in Euclidean space, but computationally much less expensive.

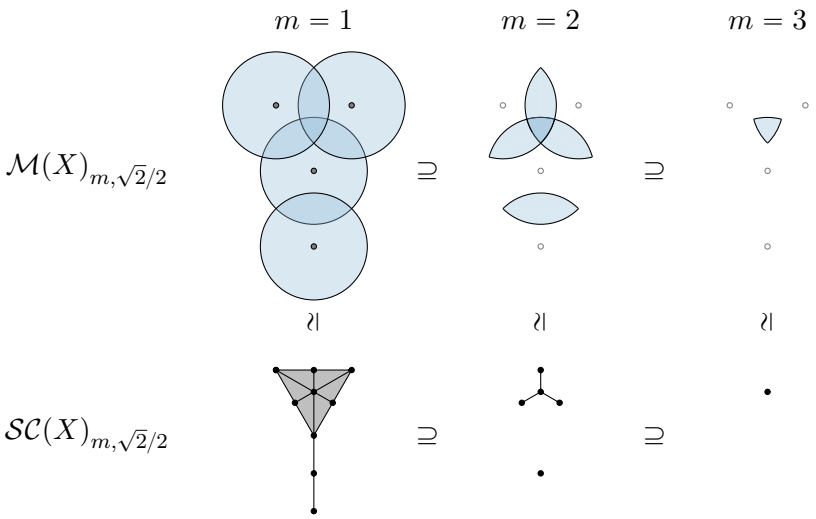


Figure 2.6: Four points X in the Euclidean plane with a portion of its multicover bifiltration and the equivalent subdivision Čech bifiltration for a fixed value of r.

Definition 2.2.38 ([109]). Let (Z, d) be a Polish space and μ be a Borel probability measure on it. The *degree Rips bifiltration* is

$$\begin{aligned} \mathcal{DR}(Z,\mu)\colon]0,\infty[^{op}\times[0,\infty[&\to \mathbf{Simp}, \\ \mathcal{DR}(Z,\mu)_{m,r} &= \mathcal{R}(\mathcal{B}(\mu)_{m,r})_r. \end{aligned}$$

That is, at each stage (m, r), we evaluate the measure bifiltration and take the resulting metric subspace as an input for the Rips complex. The name comes from the fact that for $\mu = \mu_X = \sum_{x\in X} \delta_x$ being the counting measure of a finite metric space X, the set $\mathcal{B}(\mu_X)_{m,r}$ is precisely the set of vertices in $\mathcal{R}(X)_r$ of degree $\geqslant m - 1$.

Our next aim is to describe the stability of bifiltrations in analogy to the stability of one-parameter filtrations. Recall that the Hausdorff distance is appropriate to measure distances between point clouds in the one-parameter setting. In the two-parameter setting, we are going to use the Prokhorov distance (Definition 2.1.6), which can be thought of as a density-sensitive analogue of the Hausdorff distance. Therefore, we wish to regard the input data as a probability measure. While the measure bifiltration is built to handle general measures, the other bifiltrations are defined more combinatorially and hence need to be modified. To motivate this modification, recall that the measure bifiltration of a counting measure of a finite set of points X in $\mathbb{R}^d$ is the same as its multicover bifiltration, $\mathcal{M}(X) = \mathcal{B}(\mu_X)$. Recall furthermore that the empirical probability measure associated to X is obtained by normalizing the counting measure, $\nu_X = \frac{1}{|X|}\mu_X$. Thus, for any $m > 0$ and any Borel set $A \subset \mathbb{R}^d$, we have $\nu_X(A) \geqslant m \Leftrightarrow \mu_X(A) \geqslant |X|m$ and thus $\mathcal{B}(\nu_X)_{m,r} = \mathcal{M}(X)_{|X|m,r}$ for any $r \geqslant 0$. In this vein, we introduce normalized bifiltrations.

Definition 2.2.39. For a non-empty finite metric space (X, d), define the following normalized bifiltrations

$$\mathcal{M}^n(X)\colon]0,\infty[^{op}\times[0,\infty[\to \mathbf{Top},$$
$$\mathcal{SR}^n(X)\colon]0,\infty[^{op}\times[0,\infty[\to \mathbf{Simp}.$$

- The *normalized multicover bifiltration* is given by $\mathcal{M}^n(X)_{m,r} := \mathcal{M}(X)_{|X|m,r}$,
- the *normalized subdivision Rips bifiltration* is given by $\mathcal{SR}^n(X)_{m,r} := \mathcal{SR}(X)_{|X|m,r}$,

Analogously, for a finite simplicial complex K, its *normalized subdivision filtration* is given by $\mathcal{S}^n(K)_m = \mathcal{S}(K)_{|K_0|m}$.

As an example of a stability result which unites topological and probabilistic information, let us consider the situation of the measure bifiltration (Def. 2.2.34) in detail. The result below is a generalization of the Prokhorov stability in Theorem 1.6 of [27], which is recovered by setting $\varepsilon = d_{Pr}$. An equivalent result was known to the experts, but not published before[4].

Roughly speaking, the theorem says that if we want a small deviation in distance direction, we lose control over the deviation in measure direction (and vice versa).

Theorem 2.2.40. *Let $\varepsilon > 0$, let μ, η be probability measures on a common metric space (X, d). For any $\varepsilon > 0$, consider the forward shift*

$$\alpha^\varepsilon\colon]0,\infty[^{op}\times[0,\infty[\to]0,\infty[^{op}\times[0,\infty[, \qquad (m,r) \mapsto (m - \Pi(\varepsilon), r+\varepsilon)$$

where $\Pi = \Pi_{\mu,\eta}$ denotes the Prokhorov profile (Definition 2.1.5). The measure bifiltrations $\mathcal{B}(\mu), \mathcal{B}(\eta)$ are $(\alpha^\varepsilon, \alpha^\varepsilon)$-interleaved, i.e. for all $k, r > 0$,

$$\mathcal{B}(\mu)_{m,r} \subseteq \mathcal{B}(\eta)_{m-\Pi(\varepsilon),r+\varepsilon} \text{ and } \mathcal{B}(\eta)_{m,r} \subseteq \mathcal{B}(\mu)_{m-\Pi(\varepsilon),r+\varepsilon}$$

Proof. Let γ be a coupling (Definition 2.1.4) between μ and η. Let $x \in \mathcal{B}(\mu)_{m,r}$, then

$$m \leqslant \mu(B_r(x)) = \gamma(B_r(x)\times X) = \gamma(B_r(x)\times B_{r+\varepsilon}(x)) + \gamma(B_r(x)\times (X\backslash B_{r+\varepsilon}(x))). \quad (2.2)$$

We want to show that $x \in \mathcal{B}(\eta)_{m-\Pi(\varepsilon),r+\varepsilon}$, in other words, $\eta(B_{r+\varepsilon}(x)) \geqslant m - \Pi(\varepsilon)$. First, we use additivity of the measures to rewrite and then we use the assumption on x to estimate

$$\begin{aligned}\eta(B_{r+\varepsilon}(x)) &= \gamma(X \times B_{r+\varepsilon}(x))\\ &= \gamma(B_r(x)\times B_{r+\varepsilon}(x)) + \gamma((X\backslash B_r(x))\times B_{r+\varepsilon}(x))\\ &\overset{(2.2)}{\geqslant} m - \gamma(B_r(x)\times(X\backslash B_{r+\varepsilon}(x))).\end{aligned}$$

It remains to observe that for $x_1 \in B_r(x)$ and $x_2 \in X\backslash B_{r+\varepsilon}(x)$, their distance is lower bounded as $d(x_1,x_2) > \varepsilon$. Indeed, the triangle inequality implies

$$d(x_1,x_2) \geqslant d(x,x_2) - d(x,x_1) > r + \varepsilon - r = \varepsilon.$$

[4] Michael Lesnick, personal communication

Consequently, we can estimate

$$\gamma(B_r(x) \times (X \backslash B_{r+\varepsilon}(x))) \leqslant \gamma(\{(x_1, x_2) \in X \times X : d(x_1, x_2) > \varepsilon\}).$$

As γ was arbitrary, going to the infimum yields the desired bound

$$\eta(B_{r+\varepsilon}(x)) \geqslant m - \Pi(\varepsilon).$$

A symmetric argument, interchanging the roles of μ and η, completes the proof. □

Moreover, one has homotopy interleavings of filtered simplicial complexes as follows:

Theorem 2.2.41 ([27, Theorem 1.6 iii])**.** *Let X_1, X_2 be two non-empty finite metric spaces endowed with their empirical probability measures ν_1, ν_2. Then*[5]

$$d_{HI}(\mathcal{SR}^n(X_1)_{\bullet,2\bullet}, \mathcal{SR}^n(X_2)_{\bullet,2\bullet}) \leqslant d_{GPr}(\nu_1, \nu_2).$$

Theorem 2.2.42 ([142, Theorem 6.5.1], [27, Theorem 1.7])**.** *For any metric probability spaces (X_1, d_1, μ_1), (X_2, d_2, μ_2), we have*

$$d_{HI}(\mathcal{DR}(X_1, d_1, \mu_1), \mathcal{DR}(X_2, d_2, \mu_2)) \leqslant d_{GHPr}((X_1, d_1, \mu_1), (X_2, d_2, \mu_2));$$

moreover, for any $\delta > d_{GPr}(\mu_1, \mu_2)$, we have a homotopy-interleaving with respect to the forward-shift $(m, r) \to (m - \delta, 3r + \delta)$

Since the bound in Theorem 2.2.41 is with respect to Gromov-Prokhorov, we can interpret it as ascertaining that subdivision-Rips is *robust* (cf. [27, Remark 2.16]). The degree-Rips bifiltration only satisfies a weaker robustness result (with a multiplicative factor of 3). Moreover, we can interpret the appearance of the Hausdorff distance as $\mathcal{DR}$ being more easily affected by metric perturbations. This presents a trade-off between computability and robustness; in order to make progress towards avoiding it we will propose a new bifiltration in Chapter 7.

2.3 Persistent Homology

So far, we have gathered a collection of filtered spaces and complexes, which we regard as functors from a suitable poset $T = [0, \infty[$ or $]0, \infty[^{op}$ or $]0, \infty[^{op} \times [0, \infty[$. We can compose such a functor with (ordinary) homology, one of the classical functorial invariants of algebraic topology. To this end, we fix a field k for the remainder of this thesis. Usually, in computations $k = \mathbb{Z}/p$, with $p = 2$ the most frequent choice. For the sake of simplicity, we will not consider homology with coefficients in rings which are not fields in this thesis. Let us remind the reader of the classical definitions of singular and simplicial chain complexes and their homology.

[5] Recall that the definitions of the Vietoris–Rips complex of [27] and ours differ by a factor of two.

Definition 2.3.1. A *chain complex* $C_\bullet = (\{C_i\}_{i\in\mathbb{Z}}, \partial)$ consists of a family of k-vector spaces $\{C_i\}_{i\in\mathbb{Z}}$ together with linear maps $\partial_i = \partial_i \colon C_i \to C_{i-1}$, called *boundary maps* or *differentials*, such that $\partial_i \circ \partial_{i+1} = 0$. The i^{th} *homology* of the chain complex C is the quotient vector space

$$H_i(C) = \frac{\ker(\partial_i)}{\operatorname{im}(\partial_{i+1})}.$$

A *chain map* $f \colon (C, \partial^C) \to (D, \partial^D)$ is a sequence of linear maps

$$f_i \colon C_i \to D_i \text{ such that } \partial_i^D \circ f_i = f_{i-1} \circ \partial_i^C.$$

This compatibility with boundary maps ensures that there is a well defined induced map in homology,

$$\bar{f}_i \colon H_i(C) \to H_i(D) \qquad x + \operatorname{im}(\partial_{i+1}^C) \mapsto f_i(x) + \operatorname{im}(\partial_{i+1}^D).$$

It has been the successful story of algebraic topology to associate algebraic invariants to topological spaces in a functorial way, with singular homology being a prime example.

Definition 2.3.2. Let X be a topological space. Recall that $|\Delta^d| \subset \mathbb{R}^{d+1}$ denotes the geometric d-simplex. We construct the *singular chain complex* $C_\bullet^{sing}(X)$ via taking

$$C_i^{sing}(X) = k^{\{\sigma \colon |\Delta^i| \to X \text{continuous}\}},$$

i.e. the free k-vector space generated by all continuous maps $|\Delta^i| \to X$. Next, we define the boundary maps on basis elements $\sigma \colon |\Delta^i| \to X$ by giving the values as linear combinations of basis elements $|\Delta^{i-1}| \to X$. To this end, consider the face map

$$\begin{aligned} F_i^j \colon |\Delta^{i-1}| &\to |\Delta^i|, \\ (x_1, \ldots, x_i)^\top &\mapsto (x_1, \ldots, x_{j-1}, 0, x_j, \ldots, x_i)^\top, \end{aligned}$$

for $0 \leqslant j \leqslant i$. Put differently, say σ has vertices $[v_1, \ldots, v_{i+1}]$, where v_j is the image of the j^{th} standard basis vector in $\mathbb{R}^{i+1}$ under the map σ. Then $\sigma \circ F_i^j$ has the vertices $[v_1, \ldots, v_{j-1}, v_{j+1}, \ldots, v_i]$; that is, the j^{th} face is given by leaving out the j^{th} vertex. Now, the boundary map is defined via its values on basis elements $\sigma \colon |\Delta^i| \to X$ as

$$\begin{aligned} \partial_i^{sing} \colon C_i^{sing}(X) &\to C_{i-1}^{sing}(X), \\ \partial_i^{sing}(\sigma) &= \sum_{j=0}^{i} (-1)^j \sigma \circ F_i^j. \end{aligned}$$

A crucial feature of this construction is its functoriality, a continuous map $f \colon X \to X'$ induces a chain map

$$C_i^{sing}(f) \colon C_i^{sing}(X) \to C_i^{sing}(X'), \qquad \sigma \mapsto f \circ \sigma.$$

It is straight-forward to check that this assignment commutes with the differentials. The homology of this chain complex $C_\bullet^{sing}$ is known as the *singular homology of X with coefficients in k*. We thus obtain a family of functors

$$H_i^{sing}(\cdot, k)\colon \mathbf{Top} \to k\mathbf{Vect}.$$

Intuitively, non-trivial i^{th} homology of X means, that there is a way to map the boundary of an $(i+1)$-simplex into X in such a way that it cannot be filled in. In this sense, the dimension of $H_i^{sing}(X, k)$, called the i^{th} Betti number, counts how many i-dimensional "holes" there are in X.

In the context of computations, one is faced with the problem that the singular chain complex is verly large. Similar in spirit to how we introduced combinatorial models like the Čech complex to encode the topology of an equivalent, but uncountable, topological space, one considers a simpler construction of homology for simplicial complexes.

Definition 2.3.3. Let K be a simplicial complex and fix a total order on its vertices. We construct the associated *simplicial chain complex* $C_\bullet^\Delta(K)$ by taking $C_i^\Delta(K) = k^{K^{(i)}}$, i.e. the vector space generated by all i-simplices in K. Note that if K is finite, this chain complex consists of finite-dimensional vector spaces and thus the boundary maps can be viewed as matrices. Explicitly, the boundary maps are given on basis elements as follows:

$$\partial_i^\Delta\colon C_i^\Delta(K) \to C_{i-1}^\Delta(K)$$
$$\partial_i^\Delta([v_0, \ldots, v_i]) = \sum_{j=0}^{i} (-1)^j [v_0, \ldots, v_{j-1}, v_{j+1}, \ldots, v_i];$$

where we assume that the order of vertices respects the total order we fixed, $v_0 < \ldots < v_i$. Note the similarity to singular homology, we again take an alternating sum over the the faces leaving out one vertex. Thinking in terms of matrices, ∂_1^Δ is the incidence matrix of the 1-skeleton of K as a directed graph with edges oriented according to the total order on the vertices. In addition, this construction enjoys functoriality as well; if $f\colon K \to K'$ is a simplicial map, we define

$$C_i^\Delta(K) \to C_i^\Delta(K'), \qquad \sigma \mapsto f(\sigma),$$

which commutes with differentials, as is easy to check. Thus, we get a family of functors

$$H_i^\Delta(\cdot, k)\colon \mathbf{Simp} \to k\mathbf{Vect},$$

called the *simplicial homology of K with coefficients in k*.

In the light of geometric realization (Definition 2.2.3), one might wonder about the relation between $H_i^{sing}(|K|, k)$ and $H_i^\Delta(K, k)$. It famously turns out that they agree; therefore, we shall leave out the superscript from the notation. In addition, we will usually also drop the field of coefficients since it is fixed. By functoriality of homology, we obtain the following diagram:

$$T \to \mathbf{C} \xrightarrow{H_*} k\mathbf{vect}_{fd},$$

where T is again one of the posets $[0, \infty[$ or $]0, \infty[^{op}$ or $]0, \infty[^{op} \times [0, \infty[$, the category $\mathbf{C}$ is either $\mathbf{Top}$ or $\mathbf{Simp}$, k is our chosen field and $k\mathbf{vect}_{fd}$ denotes the finite-dimensional vector spaces. Note that having finite dimensional homology is a requirement on the functor $T \to \mathbf{C}$, which will always be satisfied in practice as we deal with finite data. To simplify notation, we will suppress the field and only write H_* for homology. The goal of this section is to describe the algebraic structure of such functors.

2.3.1 Persistence Modules

Definition 2.3.4. Fix a field k. A *persistence module* (PM) M is a functor from a poset T to finite dimensional k vector spaces[6],

$$M \colon T \to k\mathbf{vect}_{fd}.$$

Here, $k\mathbf{vect}_{fd}^T$ denotes the functor category.

While it is easy to state this definition in the full generality of an arbitrary poset, we will only ever use $[0, \infty[$ or $]0, \infty[^{op}$ or $]0, \infty[^{op} \times [0, \infty[$. More explicitly, a PM M consists of a finite dimensional vector space M_t for every $t \in T$ and maps

$$M_{s \leqslant t} \colon M_s \to M_t \qquad \text{for } s \leqslant t.$$

These transition maps respect composition in the following way:

$$M_{r \leqslant t} = M_{s \leqslant t} \circ M_{r \leqslant s} \qquad \text{for } r \leqslant s \leqslant t.$$

Maps of PMs are natural transformations. That means a map $f \colon M \to N$ consists of components $f_t \colon M_t \to N_t$, which satisfy

$$f_t \circ M_{s \leqslant t} = N_{s \leqslant t} \circ f_s \qquad \text{for } s \leqslant t.$$

As a functor category whose codomain is abelian, $k\mathbf{vect}_{fd}^T$ itself is also abelian. Thus, we can talk about (co)kernels, direct sums and so on. The notion of interleavings also makes sense for persistence modules, hence we already know how to compare them. Functorialty of homology immediately turns an interleaving of spaces or simplicial complexes into an interleaving of persistence modules. This entails the following result:

Proposition 2.3.5. *Let X, Y be two filtered spaces or simplical complexes. Then*

$$d_I(H_*(X), H_*(Y)) \leqslant d_I(X, Y).$$

2.3.2 One-Parameter Modules

Let us turn to some concrete examples in the case of $T = [0, \infty[$. A particularly simple and important kind of persistence module can be defined as follows:

[6] Sometimes, this is called a *pointwise finite dimensional persistence module*, and one defines persistence modules without the finite-dimensionality condition. However, we shall not need infinite-dimensional persistence modules in this thesis.

Example 2.3.6. For an interval $I \subseteq [0, \infty[$ consider the module kI built as follows:

$$(kI)_t = \begin{cases} k & \text{if } t \in I, \\ 0 & \text{otherwise,} \end{cases}$$
$$(kI)_{s \leqslant t} = \begin{cases} \mathrm{id}_k & \text{if } s, t \in I, \\ 0 & \text{otherwise.} \end{cases}.$$

Here, k is again the fixed chosen field.

Definition 2.3.7. Modules of the shape of Example 2.3.6 are called *interval modules*.

Of course, persistence modules can be more complicated. However, interval modules form the "building blocks" of persistence modules in the following precise sense:

Theorem 2.3.8 ([55])**.** *Let M be a persistence module. Then there is a unique multiset of intervals $\mathcal{I}_M$ such that*

$$M \cong \bigoplus_{I \in \mathcal{I}_M} kI.$$

In the context of persistent homology, we think of an interval as a topological feature that is present in the filtration for some time.

Definition 2.3.9. The multiset of intervals $\mathcal{I}_M$ in the decomposition in Theorem 2.3.8 is called the *barcode* of the persistence module M. For an interval $I \in \mathcal{I}_M$, we say $b(I) := \inf(I)$ is its *birth time* and $d(I) := \sup(I)$ is its *death time*.

Such a barcode can be visualized via a persistence diagram.

Definition 2.3.10. A *persistence diagram* (PD) is multiset of points in $(\mathbb{R} \cup \{\infty\})^2$, consisting of

- points above the diagonal $(b, d), b < d$, each with finite multiplicity and
- each point on the diagonal $\Delta = \{(s, s) \in \mathbb{R}^2\}$ with countable multiplicity.

The convention to include diagonal points with infinite multiplicity will be useful for the construction of distances between persistence diagrams.

To obtain a PD from the above interval decomposition, collect the birth and death times of the intervals

$$\mathrm{Dgm}(M) = \{(b(I), d(I)) \in (\mathbb{R} \cup \{\infty\})^2 \colon I \in \mathcal{I}_M\};$$

add all the points on the diagonal with countable multiplicities. Off-diagonal points have finite multiplicities since the persistence module is pointwise finite dimensional. We will freely identify off-diagonal points in the diagram with the corresponding interval. Points close to the diagonal have a short lifetime and are often regarded as noise.

Example 2.3.11. Consider the Vietoris-Rips complex of the point cloud $\{(0,0),(0,4),(1,5),(3,4),(5,2),(3,-1)\}$ shown in Figure 2.7. We work with $([0, \infty[, \leqslant)$ as the underlying poset, but changes in homology in dimensions 0 and 1 only happen at $r \in \{0, \sqrt{2}, \sqrt{5}, 3, \sqrt{10}, \sqrt{13}, 4, 5\}$. For negative filtration values, the simplical complex is empty. At filtration index $r = 0$, the six points appear, giving rise to intervals starting at 0 in the barcode and points with horizontal coordinate equal 0 in the PD

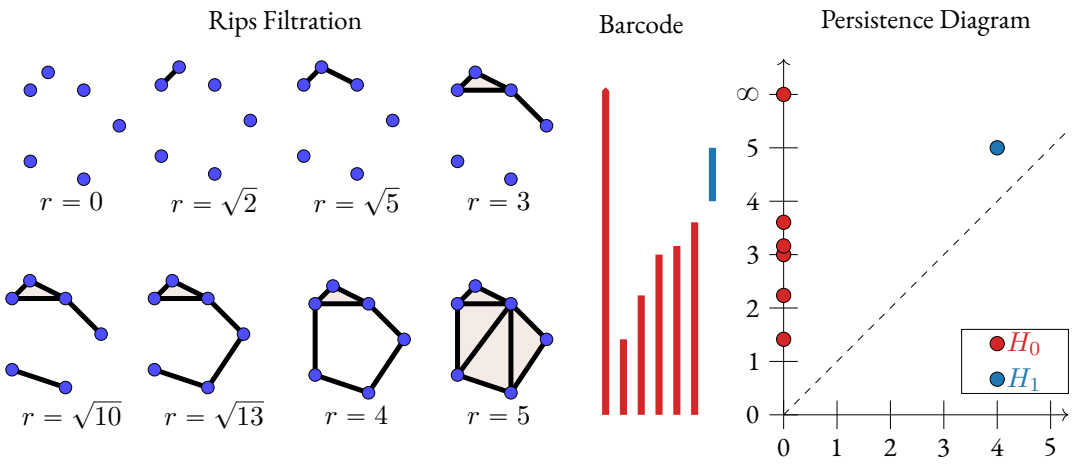


Figure 2.7: The Rips complex of a small point cloud and its persistent homology, which is represented by the barcode or, equivalently, by the persistence diagram.

of H_0. The closest distance between two points is $r = \sqrt{2}$, at this point in the filtration $(0, 4)$ and $(1, 5)$ become connected. As a consequence, we get an interval $[0, \sqrt{2}[$ in the barcode (we exclude the right endpoint because at $r = \sqrt{2}$ the component has already merged) and a point $(0, \sqrt{2})$ in the PD of H_0. Similarly, merges happen at $r \in \{\sqrt{5}, 3, \sqrt{10}, \sqrt{13}\}$, causing bars to end at these coordinates. Then at $r = 4$, a cycle is created. Thus, a new interval appears in the barcode. This cycle gets filled by triangles at $r = 5$, setting the right endpoint of the interval and giving rise to the point $(4, 5)$ in the persistence diagram. One connected component never vanishes and continues to exist forever, i.e. it corresponds to an interval $[0, \infty[$. In the persistence diagram, we draw points with vertical coordinate ∞ at the very top.

Distances and Stability

To compare persistence diagrams, we consider one-to-one correspondences between them. To take care of different cardinalities of off-diagonal points and to get rid of noisy, short-lifetime points, we allow them to be mapped to the diagonal. This explains the inclusion of the diagonal with infinite multiplicity in the above definition.

Definition 2.3.12. A *matching* η between persistence diagrams X and Y is a bijection which fixes all but finitely many diagonal points. The *cardinality* or *size* of a matching η, denoted by $|\eta|$, is the number of points which are not fixed.

Definition 2.3.13. The *bottleneck distance* between two persistence diagrams X, Y is

$$W_\infty(X, Y) = \inf_{\eta\colon X \to Y} \sup\{d(x, \eta(X))\colon x \in X\},$$

where η ranges over all matchings.

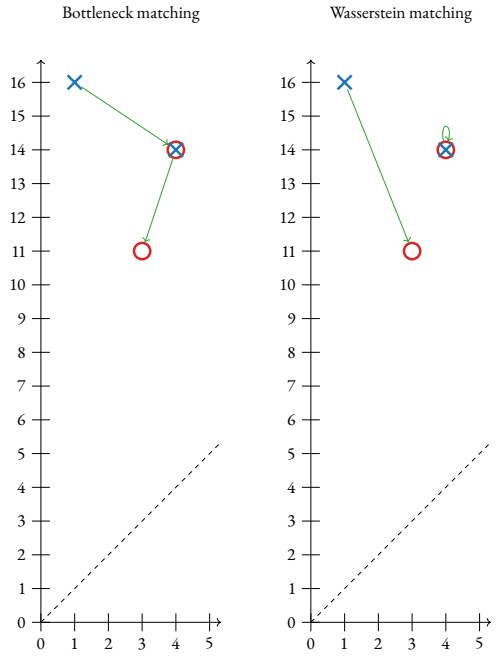


Figure 2.8: Two different matchings realizing $W_\infty(X,Y)$ and $W_1(X,Y)$, respectively.

Definition 2.3.14. Let $1 \leqslant p < \infty$. The *p-Wasserstein distance* between two persistence diagrams X, Y is

$$W_p(X,Y) = \inf_{\eta\colon X \to Y} \left(\sum_{x \in X} d(x, \eta(x))^p \right)^{\frac{1}{p}},$$

where η ranges over all matchings.

The notation of Definition 2.3.14 has the advantage of being compact, but note that the summation index ranges over an uncountable set. Although usually, only finitely many, namely $|\eta|$ for the optimal matching η, will be non-zero. Similarly, also only finitely many elements of the uncountable set of which we take the supremum in Definition 2.3.13 are non-zero. If there are points in the persistence diagram with non-finite death coordinates, they are treated according to the formula

$$d((b_1, \infty), (b_2, \infty)) := |b_1 - b_2|.$$

In particular, Wasserstein and Bottleneck distance are infinite if the number of infinite bars in the barcodes X and Y disagrees. For the metric d on $\mathbb{R}^2$, one usually chooses one induced by a q-norm $\|\cdot\|_q$, common choices include $q = 1, 2, \infty$ and in particular $q = p$ in the context of W_p.

Example 2.3.15. Consider the two persistence diagrams $X = \{(1,16), (4,14)\}$ and $Y = \{(3,11), (4,14)\}$. In Figure 2.8, we show X with blue ✕ symbols, and Y with red ○s. Using $\|\cdot\|_\infty$ on the plane $\mathbb{R}^2$,

let us work out the matchings realising Bottleneck and 1-Wasserstein distances. In the left panel, we indicate the optimal bottleneck matching $\eta_\infty \colon X \to Y$,

$$\eta_\infty \colon (1,16) \mapsto (4,14); \qquad (4,14) \mapsto (3,11).$$

The bottleneck distance is thus computed as

$$W_\infty(X,Y) = \sup\{\max\{3,2\}, \max\{1,3\}\} = 3.$$

However, this matching has 1-Wasserstein cost $3+3 = 6$, which is suboptimal. Indeed, the Wasserstein distance $W_1(X,Y) = 5$ is realized by the matching $\eta_1 \colon X \to Y$, which is shown in the right panel and maps

$$\eta_1 \colon (1,16) \mapsto (3,11); \qquad (4,14) \mapsto (4,14).$$

Definition 2.3.16. Let $p \geqslant 1$. We say a persistence diagram X has *finite pth moment*, if the p-Wasserstein distance to the empty diagram is finite: $W_p(X,\varnothing) < \infty$.

Except from section 3.3.2, the persistence diagrams in this thesis are assumed to have finitely many off-diagonal points. Therefore, the infima in definitions 2.3.13 and 2.3.14 are actually minima.

Notice the analogy between Definitions 2.1.8 and 2.3.14. We replace probability measures by counting measures and hence turn the integral into a sum. The infimum is taken over all matchings instead of all couplings. This observation will serve as a blueprint for the construction of the discrete Prokhorov metric for persistence diagrams in Section 3.3.

It turns out the bottleneck distance, defined in computational-geometric terms, is intimately related to the interleaving distance. This was established by [16], whose exposition we follow here.

Definition 2.3.17. Let $\delta > 0$. Recall the notion of the δ-*shift* endofunctor $-[\delta] \colon k\mathbf{vect}_{fd}^{[0,\infty[} \to k\mathbf{vect}_{fd}^{[0,\infty[}$ on an object M to be the PM with

$$M[\delta]_t = M_{t+\delta}, \qquad M[\delta]_{s \leqslant t} = M_{s+\delta \leqslant t+\delta},$$

and on morphisms $f \colon M \to N$ to be

$$f[\delta] \colon M[\delta] \to N[\delta], \qquad f[\delta]_t = f_{t+\delta}.$$

One readily checks that identity and composition are respected.

Definition 2.3.18. Let $\delta > 0$. A PM M is called δ-*trivial* if the canonical shift map $\varphi_M^\delta \colon M \to M[\delta]$ is the zero map.

The insight of Bauer and Lesnick [16] is the following equivalent characterisation of the interleaving distance.

Theorem 2.3.19. *Let M, N PMs, let $\delta > 0$. The modules M, N are δ-interleaved if and only if there is a map $f \colon M \to N[\delta]$ such that* $\ker(f)$ *and* $\operatorname{coker} f$ *are 2δ-trivial.*

As a consequence, they were able to re-prove the following *isometry theorem*, which states that the bottleneck distance of the diagrams agrees with the interleaving distance of the modules:

Theorem 2.3.20. $W_\infty(\mathrm{Dgm}(M), \mathrm{Dgm}(N)) = d_I(M, N)$.

Versions of this theorem were already known previously [43]. While the interleaving distance is NP-hard to compute in general [22], the bottleneck distance can be efficiently calculated by combining a binary search with a maximum cardinality unweighted bipartite matching algorithm [72, 100]. Hence, bottleneck distances are an efficient way to compare one-parameter persistence modules. The important consequence of Theorem 2.3.20 is stability of persistent homology when endowed with the bottleneck distance:

Corollary 2.3.21. *Let X, Y be finite subsets of some metric space (Z, d). Then*

$$W_\infty(\mathrm{Dgm}(H_*(\mathcal{C}(X))), \mathrm{Dgm}(H_*(\mathcal{C}(Y)))) \leqslant d_H(X, Y).$$

For Rips complexes, one also obtains a stability theorem in terms of Bottleneck distance as a consequence of Theorems 2.2.21 and 2.3.20, which reads as follows:

Theorem 2.3.22 ([44, Theorem 3.1])**.** *Let X, Y be finite metric spaces. Then*

$$W_\infty(\mathrm{Dgm}(H_*(\mathcal{R}(X))), \mathrm{Dgm}(H_*(\mathcal{R}(Y)))) \leqslant 2d_{GH}(X, Y).$$

The stability theory of p-Wasserstein distances for $p \neq \infty$ has been established much more recently and is comparatively more intricate [144].

Remark 2.3.23. The algebraic theory of two-parameter modules is intrinsically much more complicated than the one-parameter case [33]. We shall not need it in this thesis.

2.3.3 Invariants and Vectorizations

Definition 2.3.24. Let $M \in k\mathbf{vect}_{fd}^T$ be a persistence module. Its *Hilbert function* is

$$\mathrm{hf}^M \colon T \to \mathbb{N}, \qquad t \mapsto \dim(M_t).$$

In the case of $T = [0, \infty[$, the Hilbert function is also referred to as *Betti curve*. This last name is motivated by the case of persistence modules arising as homology of filtered spaces or complexes: The function hf^M assigns to a filtration parameter the Betti number (of a certain dimension) of the filtered space or complex at this filtration index. We write

$$\beta_i(\mathcal{K}) := \mathrm{hf}^{H_i(\mathcal{K})} \colon \mathbb{R} \to \mathbb{N}, \qquad t \mapsto \dim(H_i(\mathcal{K}_t)),$$

where $\mathcal{K}$ is some filtered complex or space. Recall that aggregating the Betti numbers via an alternating sum yields the Euler characteristic.

Definition 2.3.25. Let $\mathcal{K}$ be a filtered space or complex over the poset T. Its *Euler characteristic profile (ECP)* [61] is the function

$$\chi(\mathcal{K}) \colon T \to \mathbb{Z}, \qquad t \mapsto \chi(\mathcal{K}_t) = \sum_{i \geqslant 0} (-1)^i \dim(H_i(\mathcal{K}_t));$$

in the case of $T = [0, \infty[$ this is also called the *Euler characteristic curve (ECC)*.

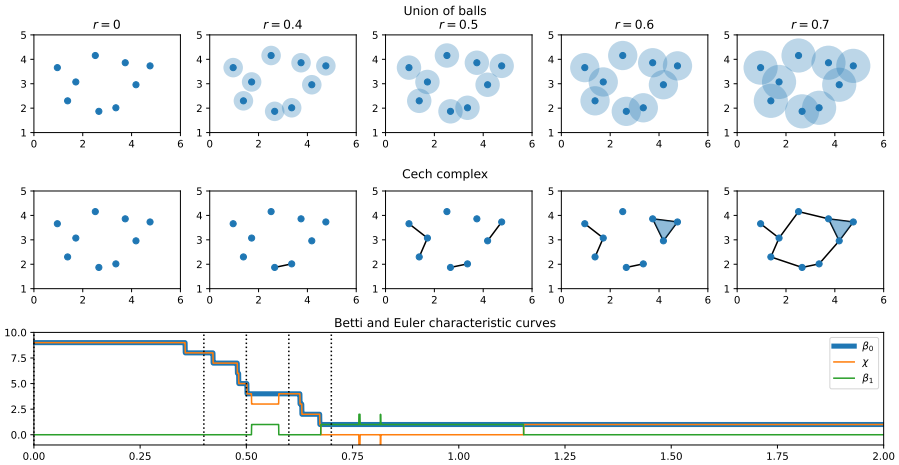


Figure 2.9: Given a point cloud, we show its offset filtration in the top panel, the corresponding Čech complex in the middle panel and the Betti and Euler characteristic curves in the bottom panel. (Note that β_0 is drawn thicker only to ease visualisation, for it would otherwise be covered most of the time by one of the other curves.) The dashed lines in the bottom panel indicate the filtration values which the top two panels are depicting.

For simplicial complexes, we get a persistent version of a classical result that is sometimes referred to as Euler-Poincaré formula:

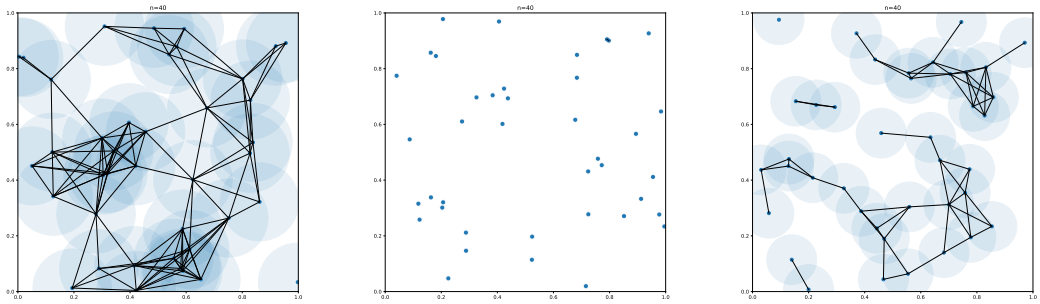
Proposition 2.3.26. *Let $\mathcal{K} \in \mathbf{Simp}^T$. Its ECP satisfies*

$$\chi(\mathcal{K})(t) = \sum_{\sigma \in \mathcal{K}_t} (-1)^{\dim(\sigma)}.$$

Example 2.3.27. Consider the point cloud drawn in Figure 2.9 with balls of increasing radius. Initially, there are nine disconnected balls, yielding $\beta_1 = 0$ and $\beta_0 = \chi = 9$. At $r = 0.4$, only the closest two points have overlapping balls, introducing an edge in the Čech complex and decreasing β_0 and χ by 1. Between $r = 0.5$ and $r = 0.6$, the topology changes more interestingly: As the three rightmost points form a triangle, β_1 becomes 1, causing $\chi = \beta_0 - 1$. Then at $r = 0.6$, the triangle is filled in as the three r-balls admit a non-empty threefold intersection. Subsequently, at $r = 0.7$, a big cycle has formed, which persists for a long time. This is reflected by a wide range of radii for which β_1 takes value 1. Finally, all the balls will admit a common intersection, in which case the Čech complex becomes the full simplex on 9 vertices. Consequently, $\chi = \beta_0 = 1$ and $\beta_1 = 0$ from that point on.

2.4 Random Complexes and their Topology

As many phenomena in the real world are random in their nature, the question arises how the introduced topological invariants behave in a stochastic context. It turns out, however, persistent homology is a complicated setting for statistics; for example, there are no unique means [149]. Instead, the focus for us is on Euler characteristic curves.



(a) Supercritical regime. (b) Subcritical regime. (c) Critical regime.

Figure 2.10: Each panel shows 40 points in the unit square sampled from a i.i.d. uniform distribution. In the supercritical case, the everything gets connected (left); in the subcritical regime, everything is disconnected (middle). In the critical regime, points are only connected locally (right), as we keep the area, which is covered by the balls, at a fixed level as $n \to \infty$.

While the Euler characteristic was studied as an intrinsic volume in stochastic geometry [49], our perspective is slightly different. Namely, the starting point for our discussion is the theory of random geometric graphs [124], which naturally generalizes to simplicial complexes [96]. In the considered setting, the vertex set from which we build simplicial complexes is sampled from some unknown distribution F on $\mathbb{R}^d$, which admits a density f. The literature distinguishes two approaches, *Poisson* and *Bernoulli* sampling; see [30] for a survey. The results of Chapter 4 are valid in both of them; we will focus on the Bernoulli sampling scheme in Chapter 5. In the Poisson setting on the one hand, the samples are assumed to be generated by a spatial Poisson process with intensity nf. In the Bernoulli setting on the other hand, we consider samples of n points sampled i.i.d. from some d-dimensional distribution. Furthermore, there are three regimes to be considered when the sample size goes to infinity [124, Section 1.4], see Figure 2.10. We consider the geometric complex at scale r_n for a sequence $r_n \to 0$ whose topology is determined by the behaviour whether

$$n \cdot {r_n}^d \to \begin{cases} \infty, \\ \lambda \in]0, \infty[, \\ 0. \end{cases}$$

Note that $n \cdot {r_n}^d$ is proportional to the total volume of the union of balls of radius r_n centered at the n sample points. In the *dense* or *supercritical regime* (Figure 2.10a), $n \cdot {r_n}^d \to \infty$, so that the domain gets covered by the union of r_n-balls and the geometric complex is highly connected. Intuitively, this regime maintains only global topological information and forgets about local density. In the *sparse* or *subcritical regime* (Figure 2.10b), $n \cdot {r_n}^d \to 0$, so that the union of balls covers a vanishing subset and the geometric complex is, informally speaking, disconnected (consult [30] for details). In Chapters 4 and 5, we focus on the *thermodynamic* or *critical regime* (Figure 2.10c), i.e. we let the quantity $n \cdot {r_n}^d \to \lambda \in]0, \infty[$ approach a finite, non-zero limit. In other words, we keep the measure of the union of r_n-balls at a controlled level. This will be the perspective in Chapter 4.

Specifically, let ω_d denote the volume of the unit ball in $\mathbb{R}^d$, so that $\omega_d r_n^d$ is the volume of a ball of radius r_n and $n\omega_d r_n^d$ is the total volume of n such balls. We will state the results in Chapter 4 in terms of $\Lambda = \lim_{n\to\infty} n\omega_d r_n^d$, which turns out to simplify notation.

In Chapter 5, we will take on a slightly different perspective: Up to a constant factor, the quantity $n \cdot r_n^d$ is the average number of points in a ball of radius r_n [30, Section 1]. Now it is straightforward to observe that a subset of our sample $\sigma \subseteq X$ forms a simplex in the Čech complex at scale r_n iff

$$\bigcap_{x\in\sigma} \overline{B}_{r_n}(x) \neq \varnothing \Leftrightarrow \bigcap_{x\in n^{1/d}\sigma} \overline{B}_{\lambda}(x) \neq \varnothing.$$

This is because for any $x \in X, x' \in \mathbb{R}^d$, we have

$$\begin{aligned} \|x' - x\| \leqslant r_n &\Leftrightarrow n^{1/d}\|x' - x\| \leqslant n^{1/d} r_n \\ &\Leftrightarrow \|n^{1/d}x' - n^{1/d}x\| \leqslant \lambda^{1/d} \end{aligned}$$

This observation motivates us to scale a sample of size n by $n^{1/d}$. In fact, this setup aligns with the approach of [104], which is the key result for the construction of our test statistic in Chapter 5, whence we decide to adopt this convention of theirs.

Due to this scaling, the average number of points in a ball of radius $r = \lambda^{1/d}$ stays the same as we increase $n \to \infty$. Therefore, it makes sense to compare ECCs at fixed radius $r = \lambda^{1/d}$ for samples of different sizes. Visually speaking, we can compare (expected) ECCs from samples of different sizes in a common coordinate system using the r-axis scaled in this way. In particular, one can study the point-wise limit of the expected ECC; that is, when the sample size approaches infinity for a fixed r. Moreover, this rescaling allows us to conduct two sample tests with samples of different sizes, cf. Section 5.1.2.

For a survey on the topology of random geometric complexes see [30]. A text book for the case of one-dimensional complexes, i.e. graphs, is [124]. The Euler characteristic of random Čech complexes has been studied in [29, 31]. Notably, in [31], the limiting expected ECC in the thermodynamic regime is described for a bounded density f on a compact closed Riemannian manifold $\mathcal{M}$.

Definition 2.4.1. Let $\overline{\chi}_F \colon [0, \infty[\to \mathbb{R}$ be the function

$$\overline{\chi}_F(\Lambda) = \begin{cases} 1 & \text{if } \Lambda = 0, \\ \lim\limits_{n\to\infty} n^{-1}\mathbb{E}[\chi(\mathcal{C}(X_n)_{r_n})] & \text{otherwise,} \end{cases}$$

where X_n consists of n points sampled i.i.d. from F and $n\omega_d r_n \to \Lambda \in]0, \infty[$ as $n \to \infty$. We call the function $\overline{\chi}_F$ the *expected Euler characteristic curve*, or *EECC* for short.

Theorem 2.4.2 ([31, Cor. 4.5])**.** *We have* $\overline{\chi}_F(\Lambda) = 1 + \sum_{i=1}^{d} \gamma_i^f(\Lambda)$. *Here,* $\Lambda = \lim_{n\to\infty} n\omega_d r_n^d$ *and*

$$\gamma_i^f(\Lambda) = \frac{\Lambda^i}{\omega_d^i (i+1)!} \int_{\mathcal{M}} \int_{(\mathbb{R}^d)^i} (f(x))^{i+1} h_1^c(0, y) e^{-\Lambda(R(0,y))^d f(x)} \, \mathrm{d}y \, \mathrm{d}x, \tag{2.3}$$

where h_1^c and R are functions of the minimal enclosing sphere of their arguments, cf. [31, equations (2.5) and (2.8)].

More recently, [148] provided a functional central limit theorem for ECCs, which was subsequently generalized by [104].

Theorem 2.4.3 ([148] and [104, Theorem 3.4])**.** *We have convergence of the centered, standardized ECC in distribution in the Skorokhod J_1-topology on $[0,T]$, for any $T \in]0,\infty[$, to a centered Gaussian process f_r,*

$$n^{-1/2}(\chi(\mathcal{C}(X)_r) - \mathbb{E}_F(\chi(\mathcal{C}(X)_r))) \xrightarrow[n\to\infty]{D} f_r. \tag{2.4}$$

Recall that the Skorokhod J_1-topology formalizes the intuitive idea of allowing for "small wiggles in the time direction" as follows [21, Section 12]:

Definition 2.4.4. Let $\mathbb{D}([0,1])$ denote the set of *cadlag* functions $f\colon [0,1] \to [0,1]$, i.e. for all $t_0 \in [0,1]$ the limit from below $\lim_{t\nearrow t_0} f(t)$ exists and the from above exists as well and equals $\lim_{t\searrow t_0} f(t) = f(t_0)$. The *Skorokhod J_1-topology* is the one induced by the metric

$$\begin{aligned} d_{J_1}\colon \mathbb{D}([0,1]) \times \mathbb{D}([0,1]) &\to [0,\infty[\\ (f,g) &\mapsto \inf_\lambda \{\max\{\|f\circ\lambda - g\|_\infty, \|\lambda - \mathrm{id}_{[0,1]}\|_\infty\}\}, \end{aligned}$$

where λ ranges over increasing bijections $[0,1] \to [0,1]$.

In order to use the ECC for statistical testing in Chapter 5, one needs to understand when it is able to distinguish different probability distributions. Very recently, Vishwanath et al. [157] provided sufficient criteria to check the injectivity of topological summary statistics including ECCs, which we will expand upon in Chapter 4.

Another topological invariant is given by Betti numbers (cf. Definition 2.3.24). In the setting of random geometric complexes, they were studied initially by [96] (although implicitly already by [131]), then limit theorems and a law of large numbers were established by [165] and later strengthened by [84]. Of course, these results imply statements about the Euler characteristic via taking the alternating sum. However, there are more tools available for the Euler characteristic than for Betti numbers, allowing for example more explicit expressions for the limit expectation like the one above (Theorem 2.4.2).

3 Bottleneck Profiles and Discrete Prokhorov Metrics for Persistence Diagrams

Abstract. In topological data analysis (TDA), persistence diagrams have been a successful tool. To compare them, Wasserstein and Bottleneck distances are commonly used. We address the shortcomings of these metrics and show a way to investigate them in a systematic way by introducing bottleneck profiles. This leads to a notion of discrete Prokhorov metrics for persistence diagrams as a generalization of the Bottleneck distance. These metrics satisfy a stability result and can be used to bound Wasserstein metrics both from above and from below. We provide algorithms to compute the newly introduced quantities and end with an discussion about experiments.

Author's contributions. This chapter contains joint work with Paweł Dłotko published as [62].The project was conceived and carried out by the author of the thesis, under supervision of P.D..

3.1 Introduction

Recall that the classical Wasserstein distance from probability theory (Definition 2.1.8) can be adapted to compare persistence diagrams (Definition 2.3.14). As the Prokhorov metric (Definition 2.1.6) is another classical optimal transport distance, it is natural to look for its adaptation to persistence diagrams in an analogous way. This is the subject of this chapter. To this end, we first introduce Bottleneck profiles (Definition 3.2.1) as a discrete analogue of Definition 2.1.5, from which then a Prokhorov distance is constructed. It turns out that the Bottleneck and the Prokhorov distance are just two instances of a whole family of Prokhorov-style metrics introduced in this chapter (Definition 3.3.1). This family is parameterised by subclass of functions $f\colon [0,\infty[\to[0,\infty[$. Not every function f gives in fact rise to a genuine metric; we examine the conditions on f in which cases it does (Definition 3.3.2, such f are called *admissible*). In particular, we show:

Theorem 3.3.7. *Fix an admissible function $f\colon [0,\infty[\to[0,\infty[$. The discrete f-Prokhorov metric is an extended pseudometric.*

In addition to theoretical development, we discuss algorithms to compute the bottleneck profile and various Prokhorov-type distances. In particular, a computational complexity analysis of those algorithms is given:

Proposition 3.3.21. *Let $f\colon [0,\infty[\to[0,\infty[$ be monotonically increasing. Assume that the values and preimages of f can be computed in $O(1)$. Then $\pi_f(X,Y)$ can be computed in $O(n^2\log(n))$.*

We provide a run-time analysis and experiments on a number of data sets. The algorithms are provided as an open source implementation.

3.2 Bottleneck Profiles

The bottleneck distance W_∞ has a major drawback: It only captures the single most extreme difference between two persistence diagrams. As a consequence, the same bottleneck distance can be realized by infinitely many different pairs of persistence diagrams. For instance, looking at Figure 3.1, one would like to say that there are four bottlenecks in the left panel and one in the right in a rigorous way. We introduce the notion of the bottleneck profile that is capable of capturing the topic of secondary, tertiary,... bottlenecks and their multiplicities.

Definition 3.2.1. Given two persistence diagrams X, Y, define their *bottleneck profile* to be

$$D_{X,Y}\colon [0,\infty[\to\mathbb{N}\cup\{\infty\},\qquad t\mapsto \inf_{\eta\colon X\to Y}|\{x\colon d(x,\eta(x))>t\}|;$$

where $|\cdot|$ denotes the cardinality of the set.

Example 3.2.2. Let $X=\{x\}$ and $Y=\{y\}$ both consist of one point each and assume that $d(x,y)<d(x,x')+d(y,y')$, where the prime denotes the projection to the diagonal. That means that $x\mapsto y$ is an optimal matching. Consequently, the bottleneck profile looks as follows:

$$D_{X,Y}(t)=\begin{cases}1 & \text{if } 0\leqslant t\leqslant d(x,y),\\ 0 & \text{if } t>d(x,y).\end{cases}$$

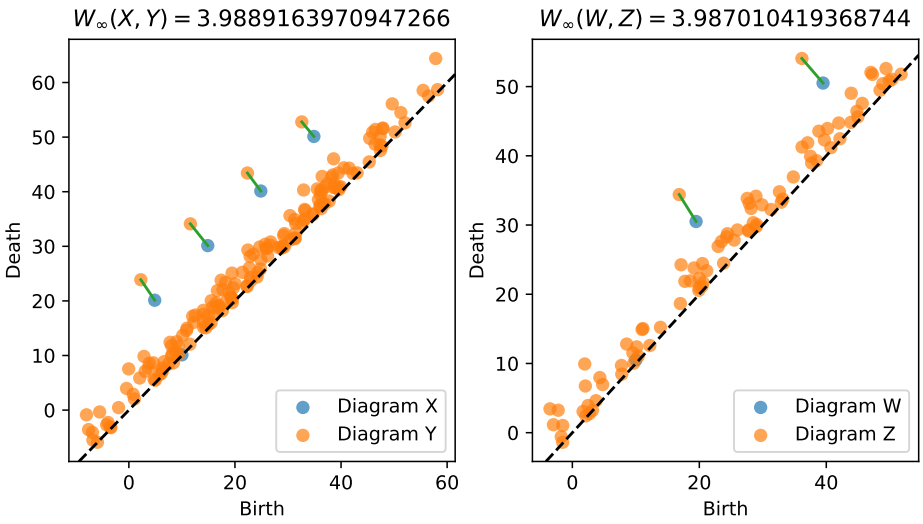


Figure 3.1: Four bottlenecks on the left, a single bottleneck on the right, realizing almost the same bottleneck distance. In green, parts of the optimal matchings are indicated; the remaining orange points are mached with the diagonal.

Example 3.2.3. If we take one of the persistence diagrams to be the empty one, there is only one choice of matching: everything is paired with the diagonal. As a consequence,

$$D_{X,\varnothing}(t) = |\{x = (x_1, x_2) \colon \frac{x_2 - x_1}{2} > t\}| = |\{x = (x_1, x_2) \colon x_1 + 2t < x_2\}|.$$

This is also known as the stable rank function corresponding to the contour $C(a, \varepsilon) = a + 2\varepsilon$, introduced in [42], which counts the bars of X of length $> 2t$.

For $d : \mathbb{R}^2 \times \mathbb{R}^2 \to [0, \infty[$ we take a p-metric $d(x, y) = \|x - y\|_p$, where the choice of p might depend on the setting. For example, when comparing with the p-Wasserstein distance, one might like to choose this same p.

Since the infimum is taken over a subset of the natural numbers, it is actually a minimum. To be consistent with the notation in definitions 2.3.13 and 2.3.14, we choose to adhere to the use of infimum.

The following observation is immediate:

Lemma 3.2.4. *The bottleneck profile $D_{X,Y}$ is monotonically non-increasing.*

Proof. Let $\eta \colon X \to Y$ be any matching realizing $D_{X,Y}(s)$ for some s. Let now $t > s$, then every distance longer than t is in particular longer than s and consequently

$$|\{x \colon d(x, \eta(x)) > t\}| \leqslant |\{x \colon d(x, \eta(x)) > s\}| = D_{X,Y}(s).$$

Taking the infimum over all matchings decreases the left hand side and yields $D_{X,Y}(t)$. □

Knowing this, it is interesting when the bottleneck profile becomes zero.

Lemma 3.2.5. $D_{X,Y}(t) = 0 \Leftrightarrow t \geqslant W_\infty(X,Y)$.

Proof. By definition, the bottleneck distance is the smallest $t > 0$ such that there is a matching mapping all points within distance t. In formulas,

$$\begin{aligned} W_\infty(X,Y) &= \inf\{t > 0 \colon \inf_{\eta \colon X \to Y} |\{x \colon d(x, \eta(x)) > t\}| = 0\} \\ &= \inf\{t > 0 \colon D_{X,Y}(t) = 0\}. \end{aligned}$$

□

Thus, we recover the bottleneck distance from the bottleneck profile. The bottleneck cost of a matching is the longest distance over which two points are matched. Minimizing the bottleneck cost over all matchings yields the bottleneck distance, which we can think of as the primary bottleneck. Similarly, the secondary bottleneck cost of a matching is the second longest distance over which two points are matched. Taking the minimum over all matchings here gives a notion of a secondary bottleneck, which equals $\inf\{t > 0 \colon D_{X,Y}(t) \leqslant 1\}$ by an argument analogous to the previous proof. This motivates the name bottleneck profile.

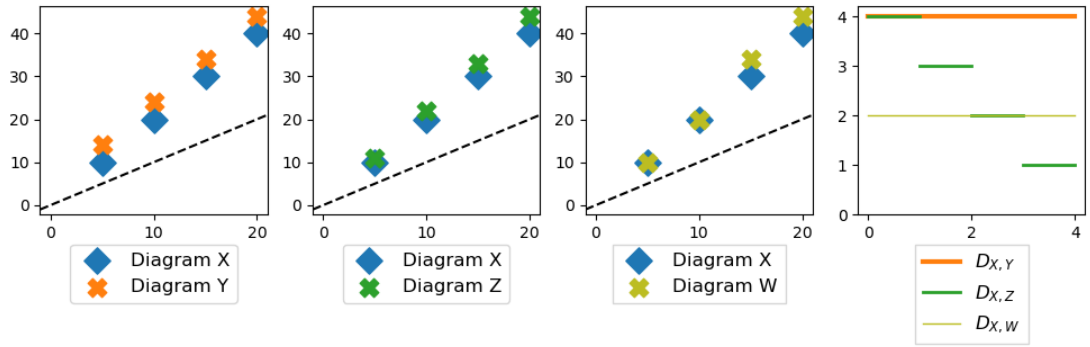


Figure 3.2: The PD X has bottleneck distance 3 to each of the PDs Y, Z, W (first three images). However, it is attained with different multiplicities, which one can read off from the bottleneck profile (right-most image)

Example 3.2.6. Consider some particular simple persistence diagrams. The first three parts of Figure 3.2 each show a base diagram ("Diagram X", in blue) with four points and perturbations of it: The orange diagram ("Diagram Y") in the first image is obtained by shifting the blue one by three. The green diagram ("Diagram Z") shifts the top point of X by three, the next point by two, the third by one and leaves the lowest point unchanged.For the yellow diagram ("Diagram W") in the third image, we only shift two points from X by three and leave the other two untouched. Clearly, the bottleneck distance between the base diagram and each of the shifted versions is three. But the amount of shifted points is reflected in the bottleneck profile: While $D_{X,Y}(t)$ is four, $D_{X,Z}(t)$ is two (i. e. the multiplicity of the bottleneck) for $0 < t < 3$. And $D_{X,W}$ displays more steps, reflecting the fact that there are secondary and tertiary bottlenecks.

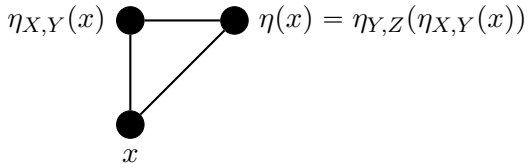


Figure 3.3: The situation in the proof of Lemma 3.2.7

Note that the function D enjoys some properties reminiscent of a metric (hence the notation D): It is obviously symmetric. The triangle inequality does not hold pointwise but in a scaled version, that is:

Lemma 3.2.7. *For all persistence diagrams X, Y, Z and all real numbers $s, t \geqslant 0$, $D_{X,Z}(s+t) \leqslant D_{X,Y}(s) + D_{Y,Z}(t)$.*

Proof. This follows from the triangle inequality on $\mathbb{R}^2$. Fix $s, t \geqslant 0$, let $\eta_{X,Y} \colon X \to Y$ and $\eta_{Y,Z} \colon Y \to Z$ denote optimal matchings realizing $D_{X,Y}(s)$ and $D_{Y,Z}(t)$, respectively. Let $\eta = \eta_{Y,Z} \circ \eta_{X,Y} \colon X \to Z$ be the matching obtained by composition. It suffices to show that

$$|\{x \colon d(x, \eta(x)) > s + t\}| \leqslant |\{x \colon d(x, \eta_{X,Y}(x)) > s\}| + |\{y \colon d(y, \eta_{Y,Z}(y)) > t\}|,$$

because the left hand side only decreases if we take the infimum over all matchings. Hence we have to investigate what happens when a point x is matched to $\eta(x)$ which is farther apart than $s + t$. Note that $\eta(x) = \eta_{Y,Z}(\eta_{X,Y}(x))$, so we compare the distances of the matched points using the triangle inequality,

$$s + t < d(x, \eta(x)) \leqslant d(x, \eta_{X,Y}(x)) + d(\eta_{X,Y}(x), \eta(x)).$$

Therefore, it cannot be that both $d(x, \eta_{X,Y}(x)) \leqslant s$ and $d(\eta_{X,Y}(x), \eta(x)) \leqslant t$ (compare Figure 3.3). That means, we have $d(x, \eta_{X,Y}(x)) > s$ or $d(\eta_{X,Y}(x), \eta(x)) > t$ or both. Using the principle of inclusion-exclusion, conclude

$$\begin{aligned} |\{x \colon d(x, \eta(x)) > s + t\}| &= |\{x \colon d(x, \eta_{X,Y}(x)) > s\}| + |\{y \colon d(y, \eta_{Y,Z}(y)) > t\}| \\ &\quad - |\{x \in \colon d(x, \eta_{X,Y}(x)) > s \text{ and } d(\eta_{X,Y}(x), \eta_{Y,Z}(\eta_{X,Y}(x))) > t\}| \\ &\leqslant |\{x \colon d(x, \eta_{X,Y}(x)) > s\}| + |\{y \colon d(y, \eta_{Y,Z}(y)) > t\}|. \end{aligned}$$

□

Note that $D_{X,Y}(t) = 0$ for all $t > 0$ implies $X = Y$ only under some finiteness assumptions. For example, consider a converging sequence $(a_n)_{n \in \mathbb{N}} \subset \mathbb{R}^2$ above the diagonal with limit $a \notin (a_n)$, which is also above the diagonal. Set X to consist of all elements of the sequence $\{a_n \colon n \in \mathbb{N}\}$. Set Y to be $X \cup \{a\}$. Then for all $\varepsilon > 0$ there exists $\eta \colon X \to Y$ such that $d(x, \eta(x)) < \varepsilon$ for all $x \in X$. Therefore, $D_{X,Y}(t) = 0$ for every $t > 0$, but $X \neq Y$.

Following [26], we denote by $\bar{\mathcal{B}}$ the set of persistence diagrams such that for each $\varepsilon > 0$ there are finitely many points of persistence $> \varepsilon$. The next lemma is an immediate consequence of [26, Lemma 3.4].

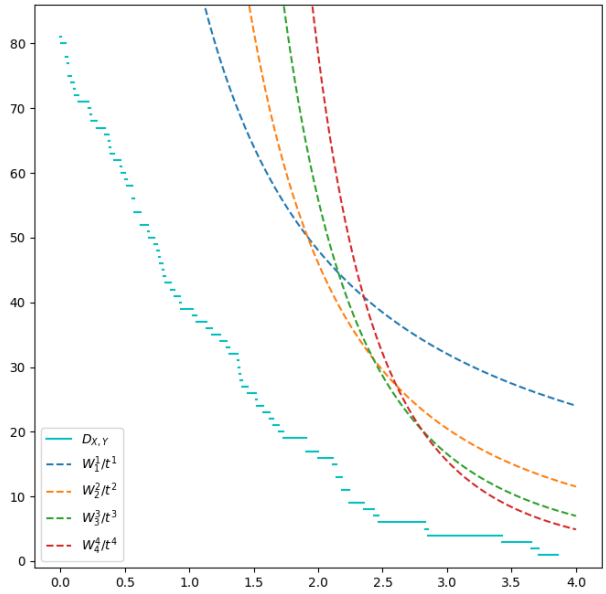


Figure 3.4: An example for the relation between $D_{X,Y}$ and the Wasserstein distance for the persistence diagrams X, Y shown on the left in Figure 3.1. Recall our intuitive statement that there were four bottlenecks in that picture – this is now evidenced by the bottleneck profile taking value 4 for a lang range of t-values.

Lemma 3.2.8. *The bottleneck profile satisfies $D_{X,X}(t) = 0$ for all Persistence diagrams X and $t > 0$. Moreover, $D_{X,Y}(t) = 0$ for all $t > 0$ implies $X = Y$ for $X, Y \in \bar{\mathcal{B}}$.*

Proof. If $D_{X,Y}(t) = 0$ for all $t > 0$, then $W_\infty(X, Y) = 0$ by Lemma 3.2.5. Now for $X, Y \in \bar{\mathcal{B}}$, this only happens if $X = Y$ by [26, Lemma 3.4]. □

3.2.1 Relation to Wasserstein distances

We have already seen how the bottleneck profile is related to the bottleneck distance. This is actually part of a more general result comparing it to p-Wasserstein metrics.

Lemma 3.2.9. *Let X, Y be two persistence diagrams, and let $p > 0$. Then*

$$D_{X,Y}(t) \leqslant \frac{1}{t^p} W_p(X, Y)^p.$$

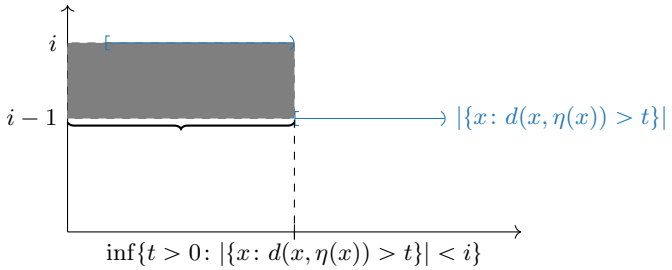


Figure 3.5: Illustrating the proof of Lemma 3.2.10: Decomposing the area under the graph into rectangles.

Proof. This follows from the Chebychev inequality for counting measures. To spell out the details, estimate that for every bijection η

$$\begin{aligned}|\{x\colon d(x,\eta(x)) > t\}| &= \sum_{\{x\colon d(x,\eta(x))>t\}} 1 \\ &\leqslant \sum_{\{x\colon d(x,\eta(x))>t\}} \frac{d(x,\eta(x))^p}{t^p} \\ &\leqslant \sum_{x\in X} \frac{d(x,\eta(x))^p}{t^p} \\ &= \frac{1}{t^p}\sum_{x\in X} d(x,\eta(x))^p.\end{aligned}$$

Now choosing η to minimize the right hand side, we have by definition of the Wasserstein distance an estimate for $D_{X,Y}$:

$$D_{X,Y}(t) \leqslant |\{x\colon d(x,\eta(x)) > t\}| \leqslant \frac{1}{t^p} W_p(X,Y)^p.$$

□

This is illustrated by Figure 3.4. Note that we recover Lemma 3.2.5 in the limit for $p \to \infty$:

$$\left(\frac{W_p}{t}\right)^p \longrightarrow \begin{cases}\infty & \text{if } t < W_\infty \\ 1 & \text{if } t = W_\infty \\ 0 & \text{if } t > W_\infty.\end{cases}$$

For 1-Wasserstein, we have a further estimate:

Lemma 3.2.10. $\int_0^\infty D_{X,Y}(t)dt \leqslant W_1(X,Y)$.

Proof. Let $\eta\colon X \to Y$ be the matching realizing $W_1(X,Y)$. We compute the area under the graph of the function $t \mapsto |\{x\colon d(x,\eta(x)) > t\}|$, which is piece-wise constant. Decomposing it into rectangles of height one yields a width of $\inf\{t > 0\colon |\{x\colon d(x,\eta(x)) > t\}| < i\}$ for $i \geqslant 1$, cf.

Figure 3.5. The width of the ith rectangle is the length of the ith longest edge in the matching. Summing over all i is therefore the same as summing the distances over which points are matched. In formulas:

$$\begin{aligned} W_1(X,Y) &= \sum_{x\in X} d(x,\eta(x)) \\ &= \sum_{i\geqslant 1} \inf\{t>0\colon |\{x\colon d(x,\eta(x))>t\}|<i\} \\ &= \int_0^\infty |\{x\colon d(x,\eta(x))>t\}|dt \\ &\geqslant \int_0^\infty D_{X,Y}(t)dt. \end{aligned}$$

□

Proposition 3.2.11. *If the bottleneck profile $D_{X,Y}(t)$ can be realized by the same matching η for all $t>0$, then η realizes $W_1(X,Y)$.*

Proof. If η realizes $D_{X,Y}(t)$ for all $t>0$, then the inequality in the proof of the previous lemma becomes an equality

$$\int_0^\infty D_{X,Y}(t)dt = \int_0^\infty |\{x\colon d(x,\eta(x))>t\}|dt = \sum_{x\in X} d(x,\eta(x)) \overset{(*)}{\geqslant} W_1(X,Y).$$

Combining this with Lemma 3.2.10, we obtain

$$\int_0^\infty D_{X,Y}(t)dt = W_1(X,Y).$$

Consequently, the inequality $(*)$ is actually an equality, which is what we wanted to prove. □

3.2.2 Algorithms

Recall the definition

$$D_{X,Y}(t) = \inf_\eta |\{x\colon d(x,\eta(x))>t\}|,$$

and let η be the matching realizing the infimum. Then η also realizes the following supremum:

$$\sup_\eta |\{x\colon d(x,\eta(x))\leqslant t\}|,$$

and consequently

$$D_{X,Y}(t) = |\eta| - \sup_\eta |\{x\colon d(x,\eta(x))\leqslant t\}|.$$

Here, $|\eta|$ denotes the number of matched pairs which involve at least one off-diagonal point. The computation of $\sup_{\eta} |\{x\colon d(x,\eta(x)) \leqslant t\}|$ is a version of the unweighted maximum cardinality bipartite matching problem. First, set up the following notation (following [69, chapter VIII.4]). Denote by X_0 the off-diagonal points of X and by X_0' their projections to the diagonal (and analogously for Y). Set $U = X_0 \cup Y_0'$ and $V = Y_0 \cup X_0'$ and consider the bipartite graph $G = (U \cup V, E)$ with $e = \{u, v\} \in E$ if either of the following holds:

- $u \in X_0, v \in Y_0$ and $d(u,v) \leqslant t$,
- $u \in X_0, v \in X_0'$ is its projection to the diagonal and $d(u,v) \leqslant t$,
- $v \in Y_0, u \in Y_0'$ is its projection to the diagonal and $d(u,v) \leqslant t$,
- $u \in Y_0'$ and $v \in X_0'$.

Let $M \subset E$ be a matching of maximal cardinality. Observe that such a matching corresponds to a bijection $\eta\colon X \to Y$ maximizing $|\{x\colon d(x,\eta(x)) \leqslant t\}|$.

To estimate the run-time of this algorithm, let $n = |X| + |Y|$. We solve the unweighted maximum cardinality bipartite matching problem using the Hopcroft-Karp algorithm [92]. Let us briefly recall this classical algorithm. The algorithm extends a partial matching M until it reaches a maximum one. It achieves this by augmenting paths: A path p that starts at an unmatched vertex in U and ending at an unmatched vertex in V such that edges from U to V are not in M but edges from V to U are. Removing edges from $p \cap M$ from the matching and instead inserting edges from $p \cap (E \backslash M)$ increases the size of M by one. The Hopcroft-Karp algorithm finds vertex-disjoint augmenting paths in $O(n^2)$ via the so-called *layer subgraph*, which is constructed via a depth-first search in $O(n^2)$. After extending the matching using all these augmenting paths, the algorithm starts over. The algorithm terminates after $O(\sqrt{n})$ of these iterations.

While this consequently takes $O(n^{2.5})$ in the worst case, we perform a variant which exploits the geometric nature of the setting, as suggested in [72]. Instead of building the layer graph explicitly, one can use a geometric data structure that allows for querying neighbors within a given distance, as well as removing points. Following [100], k-d trees achieve this requiring $O(\sqrt{n})$ for either of the two operations. Consequently, as noted by [100] and [72], our variant of the Hopcroft-Karp algorithm runs in $O(n^2)$. Summarizing, we find the following:

Proposition 3.2.12. *Let X, Y be finite persistence diagrams and denote $n = |X| + |Y|$. The value of the bottleneck profile at t, $D_{X,Y}(t)$, can be computed in $O(n^2)$.*

Remark 3.2.13. Using k-d trees is useful in practice, but does not yield optimal theoretical run-times. Indeed, the more sophisticated data structure from [72], Section 5.1, can be constructed in $O(n \log(n))$. The two relevant operations on it require $O(\log(n))$, so that the bottleneck profile could be evaluated in $O(n^{1.5} \log(n))$ using this method.

Remark 3.2.14. Instead of using Hopcroft-Karp, one can regard the matching problem as a linear program. For each $x \in X$ and $y \in Y$, we have a binary variable f_{xy} indicating whether the edge from x to y is in the matching. The coefficients (the cost of the edge) are given by

$$c_{xy} = \begin{cases} 1 & \text{if } d(x,y) > t, \\ 0 & \text{otherwise.} \end{cases}$$

The objective is

$$\begin{aligned}&\text{minimize } \sum_{x,y} c_{xy} f_{xy}\\ &\text{subject to } \forall x \in X\colon \sum_y f_{xy} = 1,\ \forall y \in Y\colon \sum_x f_{xy} = 1.\end{aligned}$$

3.3 Discrete Prokhorov Metrics for Persistence Diagrams

A straight-forward discretization of the coupling characterization of the probabilistic Prokhorov metric (Definition 2.1.6) gives the main notion of this section.

Definition 3.3.1. Given two persistence diagrams X, Y, consider matchings $\eta\colon X \to Y$ to define their *Prokhorov distance* as

$$\begin{aligned}\pi(X,Y) &= \inf\{t > 0\colon D_{X,Y}(t) < t\}\\ &= \inf\{t > 0\colon \inf_{\eta\colon X \to Y} |\{x\colon d(x, \eta(x)) > t\}| < t\}.\end{aligned}$$

Informally, we look at the intersection of the bottleneck profile with the diagonal. Similarly, we have already seen that the bottleneck distance arises as the intersection of $D_{X,Y}$ with the horizontal axis. This motivates the the question, what functions we can intersect the bottleneck profile with to obtain a sensible notion of distance.

Definition 3.3.2. Consider a function $f\colon [0,\infty[\to [0,\infty[$. We say f is *superadditive* if for any $s, t \geqslant 0$ we have $f(s+t) \geqslant f(s) + f(t)$. A superadditive function f is called *admissible* if $\lim_{t \searrow 0} f(t) = 0$. Furthermore, the function $f \equiv 1$ is also said to be admissible.

Notice that such superadditive functions are monotonically non-decreasing. For example, any linear function $f(t) = m \cdot t$ with non-negative slope $m \geqslant 0$ is admissible. Moreover, increasing convex functions f with $\lim_{t \searrow 0} f(t) = 0$ are admissible For instance, polynomials with non-negative coefficients and absolute term zero fulfill this criterion.

Definition 3.3.3. Given a fixed admissible function $f\colon [0,\infty[\to [0,\infty[$, define for any two PDs X, Y their *f-Prokhorov distance* to be

$$\begin{aligned}\pi_f(X,Y) &= \inf\{t > 0\colon D_{X,Y}(t) < f(t)\}\\ &= \inf\{t > 0\colon \inf_{\eta} |\{x\colon d(x, \eta(x)) > t\}| < f(t)\}.\end{aligned}$$

Plugging in $f = id$ gives the Prokhorov distance, plugging in $f \equiv 1$ recovers the bottleneck distance (this is why this function is admissible even though it is not superadditive).

Intuitively, for $n \in \mathbb{N}$, plugging in $f \equiv n$ (although this is *not* an admissible function) gives the nth bottleneck.

For two Prokhorov-close PDs, we require the number (=counting measure) of unmatched points to be small. Points with small persistence get matched to the diagonal and thus do not blow up the Prokhorov distance. Hence it is robust with respect to noise.

Example 3.3.4. Assume f is invertible. Recall the situation of Example 3.2.2: $X = \{x\}$ and $Y = \{y\}$ both consist of one point each and we assume that $d(x,y) < d(x,x') + d(y,y')$, where the prime denotes the projection to the diagonal. We saw that the bottleneck profile looks as follows:

$$D_{X,Y}(t) = \begin{cases} 1 & \text{if } 0 \leqslant t \leqslant d(x,y), \\ 0 & \text{if } t > d(x,y). \end{cases}$$

It follows that

$$\pi_f(X,Y) = \min(f^{-1}(1), d(x,y)).$$

Lemma 3.3.5. *For f admissible, $D_{X,Y}(\pi_f(X,Y)) \leqslant f(\pi_f(X,Y))$.*

Proof. Note that $D_{X,Y}$ is right-continuous by construction. □

The triangle inequality follows from Lemma 3.2.7.

Lemma 3.3.6. *Fix an admissible function $f\colon [0,\infty[\to [0,\infty[$. For any three persistence diagrams X, Y, Z, we have*

$$\pi_f(X,Z) \leqslant \pi_f(X,Y) + \pi_f(Y,Z).$$

Proof. We make the following estimates:

$$\begin{aligned} D_{X,Z}(\pi_f(X,Y) + \pi_f(Y,Z)) &\leqslant D_{X,Y}(\pi_f(X,Y)) + D_{Y,Z}(\pi_f(Y,Z)) \\ &\leqslant f(\pi_f(X,Y)) + f(\pi_f(Y,Z)) \\ &\leqslant f(\pi_f(X,Y) + \pi_f(Y,Z)). \end{aligned}$$

Here we used Lemma 3.2.7 for the first inequality, Lemma 3.3.5 for the second and superadditivity of f for the final one. Therefore,

$$\inf\{t > 0 \colon D_{X,Z}(t) < f(t)\} \leqslant \pi_f(X,Y) + \pi_f(Y,Z);$$

the left hand side is the definition of $\pi_f(X,Z)$, as desired. □

As the symmetry is clear, we have shown:

Theorem 3.3.7. *Fix an admissible function $f\colon [0,\infty[\to [0,\infty[$. The discrete f-Prokhorov metric is an extended pseudometric.*

Just like for the bottleneck distance, we need some finiteness property for the π_f to be a genuine metric. Let $\bar{\mathcal{B}}$ denote the persistence diagrams which for every $\varepsilon > 0$ have only finitely many points of persistence $> \varepsilon$. Then Lemma 3.2.8 implies:

Lemma 3.3.8. *Let $f\colon [0,\infty[\to [0,\infty[$ be admissible. For $X, Y \in \bar{\mathcal{B}}$, we have $\pi_f(X,Y) = 0$ only if $X = Y$.*

Proof. If $\pi_f(X,Y) = 0$, then $D_{X,Y}(t) < f(t)$ for all $t > 0$. As the bottleneck profile is monotonically decreasing and $\lim_{t \searrow 0} f(t) = 0$, this implies $D_{X,Y}(t) = 0$ for all $t > 0$. By Lemma 3.2.8, this happens only if $X = Y$. □

Our next task is to investigate how π_f depends on the function f. While from a metric point of view, we need to fix f, the context of data science suggests a different perspective: For given training data (a fixed set of persistence diagrams) adjust f to obtain a metric that performs well on it (e.g. in a classification problem, cf. section 3.4).

Lemma 3.3.9. *Let $f, g\colon [0,\infty[\to [0,\infty[$ such that $f(t) \leqslant g(t)$ for all $t \geqslant 0$. Then for any two persistence diagrams X, Y, we have $\pi_g(X,Y) \leqslant \pi_f(X,Y)$.*

Proof. If $t > 0$ satisfies $D_{X,Y}(t) < f(t)$, then also $D_{X,Y}(t) < g(t)$. Therefore,

$$\inf\{t > 0\colon D_{X,Y}(t) < g(t)\} \leqslant \inf\{t > 0\colon D_{X,Y}(t) < f(t)\}$$

and by definition $\pi_g(X,Y) \leqslant \pi_f(X,Y)$. □

For fixed persistence diagrams, the Prokhorov metric is continuous with respect to the functions in supremum metric.

Proposition 3.3.10. *Fix two persistence diagrams X, Y. Let $f\colon [0,\infty[\to [0,\infty[$ be admissible. Then for all $\varepsilon > 0$ there is $\delta > 0$ such that for each admissible $g\colon [0,\infty[\to [0,\infty[$, we have*

$$\|f - g\|_\infty < \delta \Rightarrow |\pi_f(X,Y) - \pi_g(X,Y)| < \varepsilon.$$

Proof. Without loss of generality, assume that $f(\pi_f(X,Y)) \leqslant g(\pi_g(X,Y))$ (otherwise exchange f and g below). This implies $\pi_f(X,Y) \geqslant \pi_g(X,Y)$ by monotonicity of $D_{X,Y}$. We choose $\delta < f(\varepsilon)$ and estimate

$$\begin{aligned}
f(\pi_g(X,Y) + \varepsilon) &\geqslant f(\pi_g(X,Y)) + f(\varepsilon) && \text{by superadditivity}\\
&> f(\pi_g(X,Y)) + \delta && \text{by choice of } \delta\\
&> f(\pi_g(X,Y)) + \|f - g\|_\infty && \text{by choice of } g\\
&\geqslant g(\pi_g(X,Y)) && \text{by definition of the sup-norm}\\
&\geqslant f(\pi_f(X,Y)) && \text{by assumption.}
\end{aligned}$$

By monotonicity of f we find that

$$\pi_f(X,Y) - \pi_g(X,Y) = |\pi_f(X,Y) - \pi_g(X,Y)| < \varepsilon.$$

□

From a data science perspective, the preceding Lemma allows us to tune the parameter function f on a fixed training set of persistence diagrams.

3.3.1 Comparison with Wasserstein

Fix a persistence diagram X and consider Wasserstein metrics and Prokhorov distances to some other diagram Y. We can perturb Y by adding more "noise". More precisely, we add k points whose distance to the diagonal is less than $\pi_f(X,Y)$ and denote this diagram by Y_k. This does not affect

the Prokhorov metric at all, while for all $p \in [1, \infty[$, the value of $W_p(X, Y_k)$ goes to infinity when k does. This is what we mean when we say that the Prokhorov metric is more robust with respect to noise compared to the Wasserstein metric. In other (more mathematical) words, the identity map $\mathrm{id}\colon (\mathrm{Dgm}, \pi_f) \to (\mathrm{Dgm}, W_p)$, where Dgm is the set of all persistence diagrams, is nowhere continuous for $p \in [1, \infty[$[1]. In this section, we further explore the relation between Prokhorov and Wasserstein distances.

Similarly to the proofs in [82] for the measure-theoretic variants, we can bound our metric in terms of the Wasserstein distance. As we will explain, the metrics $\pi_{t \mapsto t^q}$ are of special interest.

Proposition 3.3.11. *Let $p \geqslant 1, q \geqslant 0, c > 0$ and $f(t) = c \cdot t^q$. For two persistence diagrams X, Y we have*

$$\pi_f(X, Y) \leqslant W_p(X, Y)^{\frac{p}{p+q}} \cdot c^{\frac{-1}{p+q}}.$$

Proof. Recall from Lemma 3.2.9 that

$$D_{X,Y}(t) = \inf_{\eta} |\{x\colon d(x, \eta(x)) > t\}| \leqslant \frac{1}{t^p} W_p(X, Y)^p.$$

We now want to find a suitable value of t such that $D_{X,Y}(t) < c \cdot t^q$ to infer that $\pi_f(X, Y) \leqslant t$. Plugging in $t = W_p(X, Y)^{\frac{p}{p+q}} \cdot c^{\frac{-1}{p+q}}$, one obtains

$$\inf_{\eta} |\{x\colon d(x, \eta(x)) > W_p(X, Y)^{\frac{p}{p+q}} \cdot c^{\frac{-1}{p+q}}\}| \leqslant \frac{W_p(X, Y)^p}{W_p(X, Y)^{\frac{p^2}{p+q}} \cdot c^{\frac{-p}{p+q}}}$$

Now if $q = 0$, the right hand side simplifies to $c = f(W_p(X, Y) \cdot c^{-1/p})$. If $q > 0$, we compute

$$\begin{aligned} \frac{W_p(X, Y)^p}{W_p(X, Y)^{\frac{p^2}{p+q}} \cdot c^{\frac{-p}{p+q}}} &= W_p(X, Y)^{\frac{p^2+pq}{p+q} - \frac{p^2}{p+q}} \cdot c^{\frac{p+q-q}{p+q}} \\ &= c \cdot \left(W_p(X, Y)^{\frac{p}{p+q}} \cdot c^{\frac{-1}{p+q}} \right)^q. \end{aligned}$$

Therefore,

$$\begin{aligned} \inf_{\eta} |\{x\colon d(x, \eta(x)) > W_p(X, Y)^{\frac{p}{p+q}} \cdot c^{\frac{-1}{p+q}}\}| &\leqslant c \cdot \left(W_p(X, Y)^{\frac{p}{p+q}} \cdot c^{\frac{-1}{p+q}} \right)^q \\ &= f\left(W_p(X, Y)^{\frac{p}{p+q}} \cdot c^{\frac{-1}{p+q}} \right) \end{aligned}$$

and we conclude $\pi_f(X, Y) \leqslant W_p(X, Y)^{\frac{p}{p+q}} \cdot c^{\frac{-1}{p+q}}$ as desired. □

Corollary 3.3.12. *Let $p \geqslant 1$, $q \geqslant 0$ and $c > 0$. The map $\mathrm{id}\colon (\mathrm{Dgm}, W_p) \to (\mathrm{Dgm}, \pi_{c \cdot t^q})$ is continuous.*

When comparing with the bottleneck distance, i.e. $p = \infty$ in the above setting, we can say even more:

[1] To avoid such problems, one usually restricts to a subset of Dgm of diagrams with "finite pth moment" [119] when using p-Wasserstein distances.

Proposition 3.3.13. *For all admissible f and all persistence diagrams we have $\pi_f(X,Y) \leqslant W_\infty(X,Y)$.*

Proof. We recall by Lemma 3.2.5,

$$\inf_\eta \{x \colon d(x,\eta(x)) > W_\infty(X,Y)\} = 0 \leqslant f(W_\infty(X,Y)),$$

and therefore $\pi_f(X,Y) \leqslant W_\infty(X,Y)$. □

Specializing to $c = 1$ and $p \in \{1,\infty\}$ or $q \in \{0,1\}$, we obtain:

Corollary 3.3.14. *The following inequalities hold:*

p \ q	0	1	q
1	$d_B \leqslant W_1$	$\pi \leqslant \sqrt{W_1}$	$\pi_{t^q} \leqslant W_1^{\frac{1}{1+q}}$
∞	$d_B \leqslant d_B$	$\pi \leqslant d_B$	$\pi_{t^q} \leqslant d_B$
p	$d_B \leqslant W_p$	$\pi \leqslant W_p^{\frac{p}{p+1}}$	$\pi_{t^q} \leqslant W_p^{\frac{p}{p+q}}$

In particular, the Bottleneck Stability Theorem 2.3.22 implies stability for the new metrics by Proposition 3.3.13:

Theorem 3.3.15. *Let X, Y be finite metric spaces, fix some admissible function f and $k \in \mathbb{N}$. Then we have*

$$\pi_f(\mathrm{Dgm}(H_k(\mathcal{R}(X))), \mathrm{Dgm}(H_k(\mathcal{R}(Y)))) \leqslant 2d_{GH}(X,Y),$$

where d_{GH} is the Gromov-Hausdorff distance (Definition 2.1.10).

We can provide not only lower but also upper bounds for Wasserstein distances in terms of the Prokhorov distance.

Proposition 3.3.16. *$W_q(X,Y)^q \leqslant \pi_{t^q}(X,Y)^q(\max(d(x,\eta(x)))^q + |\eta|)$, where $\eta \colon X \to Y$ is any matching realizing $\pi_{t^q}(X,Y)$.*

Proof. For an arbitrary bijection $\eta \colon X \to Y$, consider any $t > 0$ such that $|\{d(x,\eta(x)) > t\}| \leqslant t^q$. We estimate:

$$\begin{aligned}
W_q(X,Y)^q &\leqslant \sum_x d(x,\eta(x))^q \\
&= \sum_{d(x,\eta(x))>t} d(x,\eta(x))^q + \sum_{d(x,\eta(x))\leqslant t} d(x,\eta(x))^q \\
&\leqslant |\{d(x,\eta(x)) > t\}| \max(d(x,\eta(x)))^q + t^q |\{d(x,\eta(x)) \leqslant t\}| \\
&= |\{d(x,\eta(x)) > t\}| \max(d(x,\eta(x)))^q + t^q(|\eta| - |\{d(x,\eta(x)) > t\}|) \\
&= |\{d(x,\eta(x)) > t\}|(\max(d(x,\eta(x)))^q - t^q) + t^q|\eta| \\
&\leqslant t^q \max(d(x,\eta(x)))^q - t^{2q} + t^q|\eta|
\end{aligned}$$

Taking the infimum over all matchings and all such t we obtain the desired inequality

$$W_q(X,Y)^q \leqslant \pi_{t^q}(X,Y)^q(\max(d(x,\eta(x)))^q + |\eta|).$$

□

Combining the two inequalities from Propositions 3.3.11 and 3.3.16, we obtain a comparison for different Wasserstein metrics.

Corollary 3.3.17. $W_q(X,Y)^q \leqslant W_p(X,Y)^{\frac{pq}{p+q}}(\max(d(x,\eta(x)))^q + |\eta|)$.

Remark 3.3.18. Another inequality relating Wasserstein distances for different p and q originates from the Hölder inequality, given in [11, Lemma 3.5]: For finite persistence diagrams X, Y and real numbers $1 \leqslant q < p < \infty$, we have

$$W_q(X,Y) \leqslant |\eta|^{\frac{1}{q}-\frac{1}{p}} W_p(X,Y),$$

where η is the matching realizing $W_p(X,Y)$. Our inequality above yields a lower exponent for $W_p(X,Y)$ at the cost of multiplying with the largest distance in the matching. In particular, for $q = 1, p = 2$, our formula reads

$$W_1(X,Y) \leqslant W_2(X,Y)^{\frac{2}{3}}(\max(d(x,\eta(x))) + |\eta|),$$

with η realizing $\pi_{t^q}(X,Y)$, whereas the one of [11] reads (with η realizing $W_2(X,Y)$)

$$W_1(X,Y) \leqslant W_2(X,Y)|\eta|^{\frac{1}{2}}.$$

Depending on the size of $W_p(X,Y)$ relative to the size of X and Y, our inequality can provide sharper bounds than the one of [11]. To investigate the size of $\max(d(x,\eta(x)))$ remains an interesting question for future work. One possible application of such inequalities is that they allow to infer stability results for vectorizations with respect to W_p for $p > 1$ from the stability with respect to W_1. Another use of Propositions 3.3.11 and 3.3.16 is that the bounds they provide for Wasserstein distances are easily computed, as we will see in Section 3.3.3 below.

3.3.2 Metric and Topological Properties

Using the comparison with Wasserstein (Section 3.3.1) and the results from [119], we address questions of convergence and separability. We run into similar issues as [38, Theorems 4.20, 4.24, 4.25] and [26, section 3]. In this section, we explicitly allow diagrams with a countably infite number of off-diagonal points under certain finiteness assumptions specified below.

Theorem 3.3.19. *Let $p \geqslant 1$. The space of persistence diagrams with finite pth moment endowed with the $c \cdot t^q$-Prokhorov metric is separable.*

Proof. Let $\varepsilon > 0$, X a persistence diagram and $p \geqslant 1$. Let S be a countable dense subset for the p-Wasserstein metric; this exists by [119, Theorem 12]. In fact they show that we can take S to be the set

of finite diagrams whose points have rational coordinates. Let $X_S \in S$ be a persistence diagram such that $W_p(X, X_S) < \varepsilon^{\frac{p+q}{p}} \cdot c^{\frac{1}{p}}$. Then by Proposition 3.3.11, we have

$$\pi_{c\cdot t^q}(X, X_S) \leqslant W_p(X, X_S)^{\frac{p}{p+q}} \cdot c^{\frac{-1}{p+q}} < \varepsilon^{\frac{p}{p+q}\frac{p+q}{p}} \cdot c^{\frac{-1}{p+q}} \cdot c^{\frac{1}{p+q}} = \varepsilon.$$

□

Note that the assumptions in the previous Theorem are weaker than the ones usually considered for the bottleneck distance, compare [38, Theorem 4.18].

Recall that $\bar{\mathcal{B}}$ denotes the persistence diagrams which for all $\varepsilon > 0$ have finitely many points of persistence $> \varepsilon$. The next Theorem is a consequence of [26, Theorem 3.5], which asserts that the bottleneck distance makes $\bar{\mathcal{B}}$ into a Polish space.

Theorem 3.3.20. *The space $\bar{\mathcal{B}}$ endowed with the Prokhorov metric π_f is Polish for all admissible f.*

Proof. Let $(X_n) \subset \bar{\mathcal{B}}$ be a Cauchy sequence with respect to the Prokhorov metric π_f. Let $\varepsilon > 0$ such that $f(\varepsilon) \leqslant 1$. Then the inequality $\pi_f(X_m, X_n) < \varepsilon$ implies by definition of π_f that

$$D_{X_m, X_n}(\varepsilon) < f(\varepsilon) \leqslant 1.$$

As the bottleneck profile takes values in the integers, we conclude that $D_{X_m,X_n}(\varepsilon) = 0$ and hence, by Lemma 3.2.5, we have $\varepsilon \geqslant W_\infty(X_m, X_n)$. In particular, X_n is a Cauchy sequence with respect to the bottleneck distance. By completeness of $\bar{\mathcal{B}}$ with the bottleneck distance, there is a limit diagram $X \in \bar{\mathcal{B}}$ to which the sequence converges. Finally by Lemma 3.3.13, convergence in bottleneck implies convergence in Prokhorov.

Now for separability, consider a subset $A \subset \bar{\mathcal{B}}$ which is dense with respect to the bottleneck distance. Let $X \in \bar{\mathcal{B}}$ and $\varepsilon > 0$. Then by assumption, there is $Y \in A$ with $W_\infty(X, Y) < \varepsilon$. Then, since by Proposition 3.3.13 $\pi_f(X, Y) \leqslant W_\infty(X, Y)$, we also have $\pi_f(X, Y) < \varepsilon$. Therefore, A is dense in $\bar{\mathcal{B}}$ with respect to π_f as well. □

3.3.3 Algorithms

In this section, all persistence diagrams are finite. Now we will provide an algorithm to compute $\pi_f(X, Y)$ for continuous monotonically increasing functions f. In this case, there is always a single value $t_0 \in [0, \infty[$ such that $D_{X,Y}(t) < f(t)$ for $t > t_0$ and $D_{X,Y}(t) > f(t)$ for $t < t_0$. We can find its location by bisection. Recall that we set $n = |X| + |Y|$.

Proposition 3.3.21. *Let $f\colon [0, \infty[\to [0, \infty[$ be monotonically increasing. Assume that the values and preimages of f can be computed in $O(1)$. Then $\pi_f(X, Y)$ can be computed in $O(n^2 \log(n))$.*

Proof. First, observe that the Prokohorv distance takes its value among the pairwise distances of points in the persistence diagrams (if f crosses the bottleneck profile at one of its vertical gaps) or among preimages of integers under f (if f crosses the bottleneck profiles at one of its constant pieces), in formulas

$$\pi_f(X, Y) \in \{d(x, y)\colon x \in X, y \in Y\} \cup f^{-1}(\mathbb{N}_{\leqslant |X|+|Y|}) =: T_1.$$

To perform a binary search, we sort the elements in T_1 as a preprocess, which has runtime complexity $O(n^2 \log(n))$. In each iteration of the binary search we pick the median $t \in T_i$. Next we compute the value of the bottleneck profile $D_{X,Y}(t)$ using Proposition 3.2.12, taking $O(n^2)$. Then we compute $f(t)$, which by assumption takes $O(1)$. Now if $D_{X,Y}(t) > f(t)$ set T_{i+1} to be the right half, if $D_{X,Y}(t) \leqslant f(t)$ set T_{i+1} to be the left half of T_i. Hence we obtain a runtime of $O(n^2 \log n)$ for the binary search as well.

Algorithm 3.1: The binary search to compute $\pi_f(X, Y)$

Input: Persistence diagrams X, Y; function f
Output: $\pi_f(X, Y)$
$T = \{d(x, y) : x \in X, y \in Y\} \cup f^{-1}(\mathbb{N}_{\leqslant |X|+|Y|})$
sort T
$L = 0$; $R = \text{length}(T)$
while $L < R$ **do**
  $m = \lfloor \frac{R+L}{2} \rfloor$
  $t = T[m]$
  if $D_{X,Y}(t) > f(t)$ **then**
    $L = m + 1$
  else
    $R = m$
  end
end
return $T[L]$

□

In particular, if one uses a more efficient geometric data structure to improve the runtime of the matching algorithm, the sorting preprocessing dominates the runtime. Compare [72], Theorem 3.2 and the preceding discussion therein for more details and possible improvements of the runtime complexity. Please refer to Section 3.5 for details about our implementation and its availability.

There is an easy modification to the above algorithm to approximate π_f up to an additive error of ε. Instead of performing the binary search on the indicated discrete set (which needs to be sorted or otherwise pre-processed in a costly way, as noted), one can run it on an interval $[0, M]$. Here, M is some upper bound, for example the sum of the longest lifespans of points in X and Y respectively (which is computed in $O(n)$). We bisect the interval until we arrive at one of length less than 2ε. Its midpoint is guaranteed to be less than ε away from the true value of $\pi_f(X, Y)$.

3.4 Experiments

A simple application of the bottleneck profile, based on simple synthetic persistence diagrams, was already presented in Example 3.2.6.

3.4.1 Highlighting Geometric Intuition

This experiment is a toy example, showing how the Prokhorov distance can capture our geometric intuition more accurately than bottleneck or Wasserstein. Consider three different shapes in $\mathbb{R}^2$: a) a big circle ($r = 6$), b) a big ($r = 6$) and a medium circle ($r = 4$), c) a big ($r = 6$), a medium ($r = 4$) and small circle ($r = 2$). We take five samples with noise from each shape according to Table 3.1.

shape	number of circles	radii	samples	noise	colour in the figures
a	1	6	120	uniform from $[-0.2, 0.2]^2$	blue
b	2	6, 4	300	uniform from $[-0.23, 0.23]^2$	red
c	3	6, 4, 2	120	uniform from $[-0.2, 0.2]^2$	green

Table 3.1: The three shapes: one two and three circles.

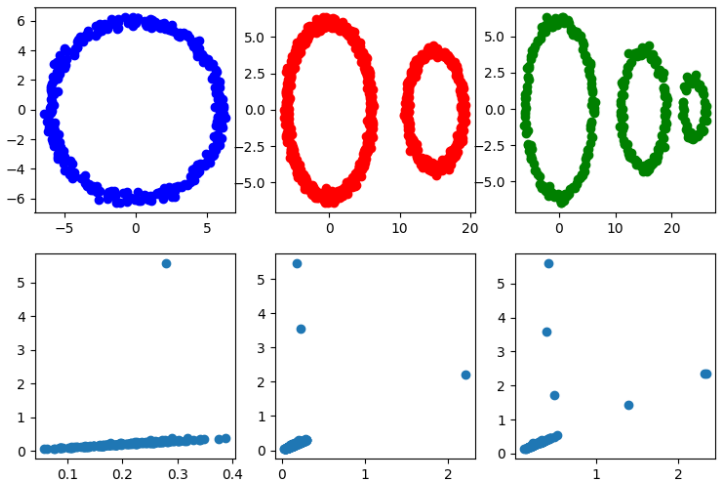


Figure 3.6: One two and three noisy circles and their PDs for the first persistent homology.

For each point cloud we compute the first persistent homology modules of it alpha complex filtration and represent them as PDs (see Figure 3.6). We can look at the averaged D-function for each pair of shapes (Figure 3.7). After careful inspection of this figure and some trial and error, we come up with the choice of $f(t) = t^3 \cdot 20^t$ to separate three bottleneck profiles in a most efficient way: Between around 0.55 and 0.65, the averaged bottleneck profiles involving shape c) with the small circle decrease, while the one comparing a) and b) stays constant. Intersecting with a function in this interval will provide a good choice for the Prokhorov distance: It puts the two and three circles closest to each other and one and three circles the farthest apart. In data science tasks, we will of course need an automated way to find a good parameter function f, we will discuss this in more detail below.

Now we want to compare the Bottleneck, Prokhorov and Wasserstein distances.

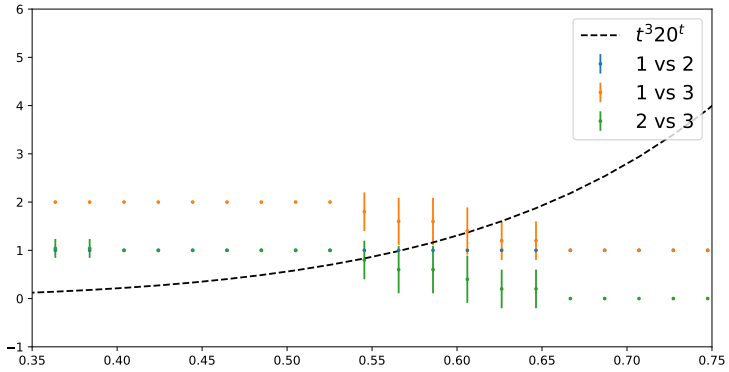


Figure 3.7: The averaged bottleneck profile for the three circles.

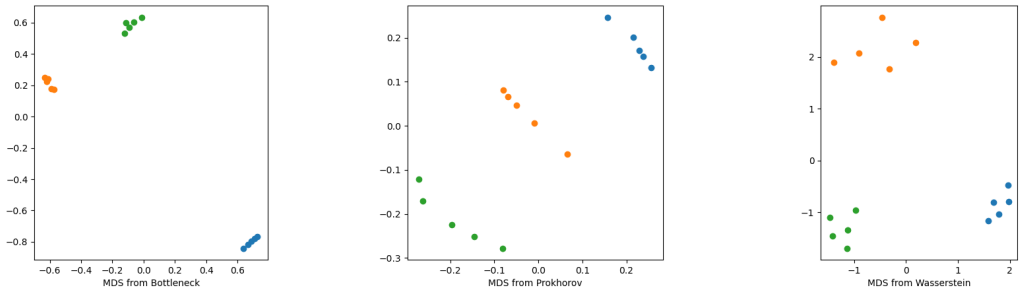


Figure 3.8: MDS plots of the dataset in Section 3.4.1.

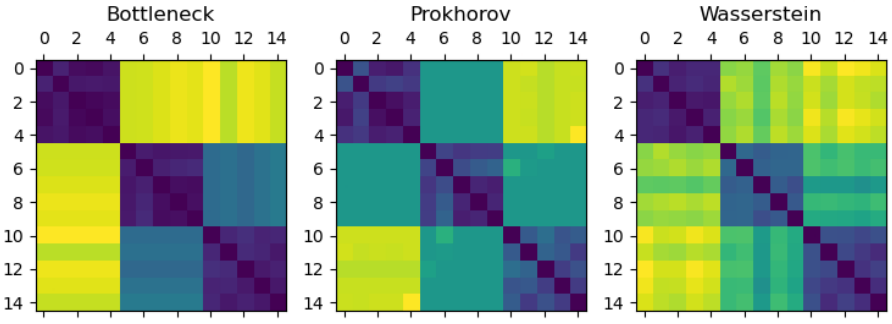


Figure 3.9: Distance matrices of the dataset in Section 3.4.1.

The bottleneck distance between shapes a) and both b) and c) is roughly the same. This distance does not take the presence of the additional small circle in shape c). By blowing up the sample size and the noise in shape b), the Wasserstein distance from a) and c) to it are artificially blown up (Figures 3.8 and 3.8). The Prokhorov distance is built to avoid these pitfalls and nicely captures the geometry of the setting. The MDS plot for Prokhorov agrees with our intuition and places b) between a) and c) (Figures 3.8).

3.4.2 Classification Experiments

We now turn to more sophisticated data sets to illustrate the usage and advantages of the Prokhorov distance. In particular, we consider persistence diagrams that actually arise in applications of TDA. We use the library [123] for standard machine learning algorithms (in particular K-Neighbors). For the Bottleneck and Wasserstein metrics we use the Gudhi library [83] and [60]. To score the different metrics, we use K-neighbors classification accuracy as well as classification accuracy based on K-Medoids clustering with the "build" initialization [140], [141]. In the latter case, points are assigned to the class of the medoid of their cluster. We split the data sets into training and testing with 50% of the points each. All computations were carried out on a laptop with an Intel i5-8265U CPU with 1.60 GHz and 8 GB memory. The code to reproduce the experiments is available online[2].

Parameter Tuning – Choosing f

One needs to specify an admissible function f as a parameter for the Prokhorov distance π_f. The set off all such functions is vast, therefore it is sensible to restrict to a smaller subset. In the experiments below, we choose f from linear functions with integer slope $\in [10, 100]$. We do this by performing a grid search over the parameters and evaluating them by five-fold cross-validation. By selecting this subset of parameters, we reduce the risk of overfitting and are able to run the parameter selection in reasonable time. We leave it as a problem for further investigation to find better means to run the parameter selection, but note that the fact that the bottleneck profile is piecewise constant obstructs the use of gradient descent.

Prokhorov Distance for Cubical Complexes with Outlier Pixels

We generate[3] 100×100 pixel greyscale images according to the following procedure, cf. Figure 3.10. Initializing every pixel with 0, we choose n points at random, at which we add a Gaußian with $\sigma = 3$. We normalize the values to $[0, 2]$ and then shift them up by 64. The goal is to distinguish images with $n = 15$ from images with $n = 20$. The obstacle is that we superimpose a particular kind noise, similar to salt-and-pepper noise. We choose k pixels randomly at which we set the value to a random integer from $[1, 128]$; the eight surrounding pixels are set to zero. For each of the four combinations $n \in \{15, 20\}$ and $k \in \{3, 5\}$ we sample 50 greyscale images. We then create a cubical complex from each using the pixels as top-dimensional cells (lower-star filtration) and compute persistent homology in dimensions 0 and 1. We proceed as indicated at the beginning of this section to assess the accuracy of the different metrics. The results are summarized in Table 3.2. Both in dimmension 0 and 1, the

[2] https://github.com/nihell/ProkhorovExamples/blob/master/Experiments.ipynb

[3] Code available at https://github.com/nihell/ProkhorovExamples/blob/master/GenerateCubicalNoise.ipynb

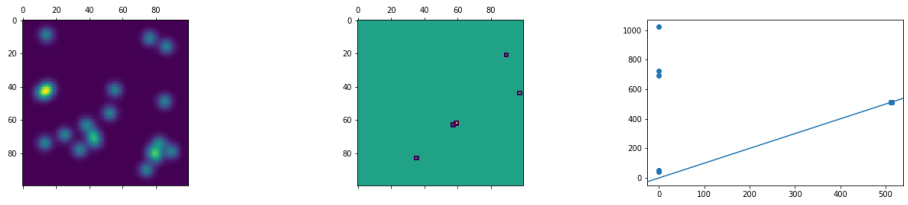


Figure 3.10: The underlying Gaußians, the superimposed noise and the resulting persistence diagram

K-Neighbors classifier is inconclusive in the setting of Bottleneck and Wasserstein. With a suitable Prokhorov metric, we are able to achieve an accuracy of more than 80%. In the K-Medoids approach, the story is similar but less pronounced: Bottleneck and Wasserstein are inconclusive, but Prokhorov achieves around 60% accuracy.

	dim	$f(t)$	Prokhorov	Bottleneck	1-Wasserstein	2-Wasserstein
K neighbors training score	0	$49t$	0.8425	0.515	0.58	0.525
K neighbors test score			0.8575	0.485	0.535	0.4925
computation time $[s]$			26.69	36.61	44.56	125.2
parameter tuning time $[s]$		1059				
K medoids training score	0	$18t$	0.5975	0.5325	0.515	0.51
K medoids test score			0.62	0.51	0.485	0.5125
computation time $[s]$			70.14	108.6	127.6	379.4
parameter tuning time $[s]$		1082				
K neighbors training score	1	$92t$	0.8625	0.5025	0.545	0.4825
K neighbors test score			0.825	0.5375	0.575	0.495
computation time $[s]$			26.57	36.68	44.43	125.9
parameter tuning time $[s]$		1025				
K medoids training score	1	$16t$	0.62	0.4925	0.49	0.485
K medoids test score			0.5975	0.5125	0.4975	0.515
computation time $[s]$			77.08	113.9	132.6	401.9
parameter tuning time $[s]$		1098				

Table 3.2: Classification scores for the synthetic dataset.

3D Segmentation

We adapt an example from [39] and [60], which is based on the dataset [46]. The task is to classify 3D-meshes based on the persistence diagrams of certain functions defined on them. The shapes are for example airplanes, hands, chairs ... The results of classification are presented in the Tables 3.3. All the considered metrics yield a similar accuracy. Prokhorov is the fastest, however at the cost of first having to find the suitable parameter, which took more than ten hours in this case.

	$f(t)$	Prokhorov	Bottleneck	1-Wasserstein	2-Wasserstein
K neighbors training score		0.9101	0.9098	0.9042	0.9059
K neighbors test score	$10t$	0.9270	0.9245	0.9312	0.9298
computation time $[s]$		795.2	1252	838.5	1740
parameter tuning time $[s]$	40440				
K medoids training score		0.4792	0.4905	0.4021	0.4592
K medoids test score	$13t$	0.4985	0.4891	0.4126	0.5125
computation time $[s]$		1946	3467	2057	5009
parameter tuning time $[s]$	41715				

Table 3.3: Classification scores for the 3d segmentation dataset.

Synthetic Dataset

Finally, we consider the dataset introduced by [3, Section 6.1]. It contains six shape classes: A sphere, a torus, clusters, clusters within clusters, a circle and the unit cube. From each class take 25 samples of 500 points. Then add two levels of Gaussian noise ($\eta = 0.05, 0.1$) and the zeroth and first persistent homology of the Vietoris-Rips filtration are computed. We compute the distance matrices and evaluate them based on the K-neighbors and K-medoids classifiers. The results are displayed in Table 3.4. We find that Prokhorov performs better Bottleneck and only slightly worse than Wasserstein. Prokhorov takes at most similarly long as 1-Wasserstein; Bottleneck is faster and 2-Wasserstein is slower.

3.4.3 Discussion

First and foremost, we found that Prokhorov is able to produce good results in situations where the classical tools of Bottleneck and Wasserstein fail. In particular, the Prokhorov distance is more robust against outliers in the persistennce diagram. Moreover, it can serve as an third option if Bottleneck is not accurate enough but Wasserstein computations are to costly. In order to explain the differences in the computation time, we note the size of the persistence diagrams in the various settings:

By inspecting Table 3.5 wee see that the 3D segmentation dataset contains way smaller diagrams, on which the Prokhorov metric seems to perform well, both in terms of runtime and score. On the bigger diagrams from the synthetic dataset, the Wasserstein metrics yield the highest scores. Prokhorov outperforms Bottleneck in the scores at the cost of higher runtimes. The difference in the computation time is caused by the evaluation of $f(t)$, which is the only difference between the Bottleneck and Prokhorov implementations.

Bottleneck – and to some extent also Prokhorov – work less well on zero-dimensional PDs. There, every class is born at time zero, hence the PD is intrinsically one-dimensional and points are matched in linear order. The bottleneck distance is less meaningful in this setting. Moreover, the Prokhorov (and even more the Bottleneck) distance do not take points matched over a small distance into account. This is a consequence of being designed to be robust against noise. However, this data can actually contain meaningful information, which is picked up by the Wasserstein distances. This is a possible explanation for the fact that Wasserstein yields better scores in the synthetic dataset.

	dim	noise	$f(t)$	Prokhorov	Bottleneck	1-Wasserstein	2-Wasserstein
K neighbors training score	0	0.05	$42t$	0.9067	0.8133	1.0	0.9867
K neighbors test score				0.84	0.7867	0.96	0.9467
computation time $[s]$				144.2	45.39	252.8	1063
parameter tuning time $[s]$			4218				
K medoids training score	0	0.05	$93t$	0.8	0.68	0.9733	0.8933
K medoids test score				0.9067	0.6	0.88	0.8933
computation time $[s]$				465.3	156.1	801.3	3207
parameter tuning time $[s]$			4507				
K neighbors training score	0	0.1	$87t$	0.9733	0.7867	0.9867	0.9867
K neighbors test score				1.0	0.7467	1.0	1.0
computation time $[s]$				145.8	44.22	267.0	1081
parameter tuning time $[s]$			4267				
K medoids training score	0	0.1	$95t$	0.8	0.6	0.9867	0.96
K medoids test score				0.9067	0.56	0.96	0.9733
computation time $[s]$				465.3	161.0	791.4	3195
parameter tuning time $[s]$			4850				
K neighbors training score	1	0.05	$51t$	0.9733	0.92	1.0	1.0
K neighbors test score				0.96	0.9333	1.0	1.0
computation time $[s]$				24.97	22.82	23.77	118.5
parameter tuning time $[s]$			736.2				
K medoids training score	1	0.05	$98t$	0.8	0.7867	1.0	1.0
K medoids test score				0.8667	0.8267	1.0	1.0
computation time $[s]$				77.63	76.08	72.20	366.7
parameter tuning time $[s]$			779.6				
K neighbors training score	1	0.1	$61t$	0.9333	0.92	0.92	0.9333
K neighbors test score				0.9467	0.93333	0.9867	0.9867
computation time $[s]$				28.17	22.28	26.98	138.4
parameter tuning time $[s]$			809.1				
K medoids training score	1	0.1	$50t$	0.88	0.6933	0.8133	0.8133
K medoids test score				0.8133	0.7067	0.8533	0.8533
computation time $[s]$				88.91	75.50	80.01	413.8
parameter tuning time $[s]$			832.2				

Table 3.4: Classification scores for the synthetic dataset from [3].

	3D-Segmentation	Synthetic data $H_0, \eta = 0.05$	Synthetic data $H_0, \eta = 0.1$	Synthetic data $H_1, \eta = 0.05$	Synthetic data $H_1, \eta = 0.1$
Mean size	11.84	500	500	177.7	189.9
standard deviation	4.893	0	0	40.53	38.84

Table 3.5: Cardinalities of the persistence diagrams for the considered experiments.

Hence, the Prokhorov metric works best on rather small diagrams and runs fastest with simple (e. g. linear) parameter functions f. Even then, one needs to take the additional time for tuning the parameter f into account.

3.5 Discussion and Outlook

Summarizing the results from the previous section, we find that the Prokhorov metric is well-suited for small persistence diagrams. Large scale computations can be improved by the technique of entropic regularization from the theory of optimal transport [105]. As the classical Prokhorov metric admits an optimal transport characterization, our discrete variant might be tractable using similar techniques.

A major aspect of the importance of the Bottleneck distance is its algebraic formulation in terms of interleavings. This theory generalizes to incorporate the family of Prokhorov metrics. An algebraic formulation would also provide a perspective on generalizations to multiparameter persistence.

Our results in section 3.3.2 establish that our construction yields a Polish space. This makes it suitable for statistical inference. In a similar vein, one can also investigate bottleneck profiles persistence diagrams arising from random geometric complexes. What kind of limit objects appear in this context? Can they be used to perform statistical testing?

Morally, stability theorems should involve related metrics on the input point cloud and on the persistence diagram side. This motivates to investigate Prokhorov-type distances for point clouds in $\mathbb{R}^n$. Such distances might be useful throughout data science.

Declarations

Code Availability

We provide an implementation as a part of a custom gudhi fork at `https://github.com/nihell/persistence-prokhorov`. It is a modification of the GUDHI implementation of the Bottleneck distance [83]. Let us first illustrate how to use it before we come to runtime considerations. The algorithm is implemented in C++ and comes with Python bindings.

```
prokhorov_distance(diagram_1: numpy.ndarray[numpy.float64],
                   diagram_2: numpy.ndarray[numpy.float64],
                   coef: numpy.ndarray[numpy.float64]) -> float
```

It asks for three inputs: `diagram_1`, `diagram_2` and `coef`. The two diagrams need to be presented as 2D numpy arrays. The third parameter is a 1D numpy array representing the coefficients of a polynomial to be used as f. Note that the zeroth entry needs to be zero in order to obtain a metric, compare Lemma 3.3.8. However, setting the polynomial to be a constant integer one recovers the values of $D_{X,Y}$, which is a feature. In the technical details, our approach follows [83], which follows [100].

In addition, we also add the Prokhorov metric to [60], allowing for parallel computations of distance matrices and integration with `sklearn`.

4 When Do Two Distributions Yield the Same Expected Euler Characteristic Curve in the Thermodynamic Limit?

Abstract. Let F be a probability distribution on $\mathbb{R}^d$ which admits a bounded density. We investigate the Euler characteristic of the Čech complex on n points sampled from F i.i.d. as $n \to \infty$ in the thermodynamic limit regime. As a main result, we identify a condition for two probability distributions to yield the same expected Euler characteristic under this construction. Namely, this happens if and only if their densities admit the same excess mass transform. Building on work of Bobrowski, we establish a connection between the limiting expected Euler characteristic of any such probability distribution F and the one of the uniform distribution on $[0,1]^d$ through an integral transform. Our approach relies on constructive proofs, offering explicit calculations of expected Euler characteristics in lower dimensions as well as reconstruction of a distribution from its limiting Euler characteristic. This research sheds light on the relationship between a probability distribution and topological properties of the Čech complex on its samples in the thermodynamic limit.

Author's contributions. This chapter contains joint work with Tobias Fleckenstein [77], submitted to *Advances in Applied Mathematics*, for which T.F. and N.H. share co-lead authorship. T.F. suggested that results to the extent of what became Theorems 4.2.1 and 4.3.1 should be true. N.H. proved Theorem 4.3.1; T.F. and N.H. jointly proved Theorem 4.2.1.

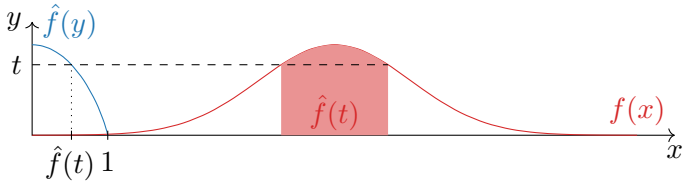


Figure 4.1: Illustration of a density (whose domain is the horizontal axis) and its excess mass, which is defined on the vertical axis and takes values on the horizontal axis.

The subject of the present chapter is to show the perhaps surprising implication that the expected Euler characteristic curve in the thermodynamic limit completely determines the excess mass. This is Theorem 4.3.1 below. Together with recent results of Vishwanath [157], who established the opposite implication, we conclude the following theorem:

Theorem 4.0.1. *Let F, G be probability distributions on $\mathbb{R}^d$ with densities with respect to the Lebesgue measure f, g which are bounded. The following are equivalent:*

i) The excess mass transforms agree $\int_{\mathbb{R}^d} \mathbb{1}_{[t,\infty[}(f(x))\, f(x)\, \mathrm{d}x = \int_{\mathbb{R}^d} \mathbb{1}_{[t,\infty[}(g(x))\, g(x)\, \mathrm{d}x$ for all $t > 0$,

ii) in the thermodynamic limit, the expected Euler characteristic curves agree: $\overline{\chi}_F(\Lambda) = \overline{\chi}_G(\Lambda)$ for all $\Lambda > 0$.

We will need this as a crucial ingredient in the next chapter, cf. Equation (5.8).

Remark 4.0.2. Note that Vishwanath et al. actually showed that admitting the same excess mass("$\mathcal{F}$-equivalence" in their terminology) is a sufficient condition for topological summaries of a wide variety to be indiscriminative in the thermodynamic limit, both in expectation and in distribution. Our Theorem 4.3.1 now says that if the expected ECCs of two distributions agree, then those other topological summaries necessarily agree as well (in the thermodynamic limit).

4.1 Background

Let F be a probability distribution on $\mathbb{R}^d$ which admits a density $f\colon \mathbb{R}^d \to \mathbb{R}$ with respect to the Lebesgue measure. Throughout, we assume it is bounded, i.e. $\|f\|_\infty < \infty$.

Definition 4.1.1. We define the *excess mass transform* of a probability density $f\colon \mathbb{R}^d \to [0, \infty[$ as

$$\hat{f}(t) = \int_{\mathbb{R}^d} \mathbb{1}_{[t,\infty[}(f(x))\, f(x)\, \mathrm{d}x. \tag{4.1}$$

It is easy to see that the function $1 - \hat{f}$ is the distribution function of the random variable $f(X)$ where $X \sim F$. Note that our definition is slightly different from Müller & Sawitzki [120] and Polonik [126]. See Figure 4.1 for an illustration.

We are interested in sampling more and more points from F, recall this can be done in the *Bernoulli* or in the *Poisson* setting. We then study the asymptotics of the expected Euler characteristic curve

$\overline{\chi}_F$ (cf. Definition 2.4.1); more specifically, we identify the fibre of the map $F \mapsto \overline{\chi}_F$. Recall (cf. Theorem 2.4.2) that $\overline{\chi}_F = 1 + \sum_{k=1}^{d} (-1)^k \gamma_k^f(\Lambda)$. Bobrowski and Mukherjee provide explicit formulas for γ_k for uniform distributions in dimension up to 3. In general, the EECC of a uniform distribution is of the form $\overline{\chi}_{\mathcal{U}^d} = e^{-\Lambda} P(\Lambda)$, for a certain polynomial $P(\Lambda) = \sum_{i=0}^{d} p_i \Lambda^i$ with $p_0 = 1$ [32, Corollary 6.2]. For $d = 1, 2, 3$, they are known explicitly (see, for instance, [131]):

$$\begin{aligned}
\overline{\chi}_{\mathcal{U}^1}(\Lambda) &= e^{-\Lambda} \\
\overline{\chi}_{\mathcal{U}^2}(\Lambda) &= e^{-\Lambda}(1 - \Lambda) \\
\overline{\chi}_{\mathcal{U}^3}(\Lambda) &= e^{-\Lambda}\left(1 - 3\Lambda + \frac{3\pi^2}{32}\Lambda^2\right).
\end{aligned}$$

If one replaces Euclidean by a more general p-distance, analogous results to Theorem 2.4.2 were established in [147, Theorem 4.3.1]. Formulas of the limit expectation for the uniform distribution are provided only for $p = \infty$ in terms of Touchard polynomials [147, Corollary 4.3.3].

4.2 An Integral Transform Formula

Throughout, we let F be a probability distribution on $\mathbb{R}^d$ which admits a density f with respect to the Lebesgue measure. Before we state our theorem, we give some intuitive heuristic motivating it. Consider a small volume element A around a point $x \in \mathbb{R}^d$. For a sample of sufficiently large size n, the relative amount of points falling into A is roughly $vol(A)f(x)$. If we choose A small enough, we can replace f by its average value on A. We expect $vol(A)f(x)$ times as many points as from a uniform sample in A. Therefore, also the total volume of the union of balls gets scaled by $f(x)$. In the thermodynamic limit, we can ignore the effects of points outside A. Then the local contribution of our small region to the EECC $\overline{\chi}_F(\Lambda)$ is consequently $f(x)\overline{\chi}_{\mathcal{U}^d}(\Lambda f(x))vol(A)$. Letting A become infinitesimally small and integrating over all local contributions now recovers the EECC:

Theorem 4.2.1. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a bounded probability density. Then we have the following formula for the expected ECC in the thermodynamic limit:*

$$\overline{\chi}_F = \int_{\mathbb{R}^d} f(x)\overline{\chi}_{\mathcal{U}^d}(\Lambda f(x))\, \mathrm{d}x. \tag{4.2}$$

In addition, we have

$$\overline{\chi}_F = -\int_0^{\|f\|_\infty} \hat{f}'(y)\, \overline{\chi}_{\mathcal{U}^d}(\Lambda y)\, \mathrm{d}y, \tag{4.3}$$

where $\hat{f}'$ is the derivative of the excess mass function, which can be understood in a distributional sense.

Proof. We simply rearrange the formula 2.3 and introduce an integral over $[0,1]^d$ of a constant function, which is just a multiplication by one:

$$\begin{aligned}
\gamma_k^f(\Lambda) &= \frac{\Lambda^k}{\omega_d^k(k+1)!}\int_{\mathbb{R}^d}\int_{(\mathbb{R}^d)^k} f^{k+1}(x)h_1^c(0,y)e^{-\Lambda R^d(0,y)f(x)}\,\mathrm{d}y\,\mathrm{d}x\\
&= \int_{\mathbb{R}^d}\frac{\Lambda^k}{\omega_d^k(k+1)!}(f(x))^{k+1}\int_{(\mathbb{R}^d)^k} h_1^c(0,y)e^{-\Lambda R^d(0,y)f(x)}\,\mathrm{d}y\,\mathrm{d}x\\
&= \int_{\mathbb{R}^d} f(x)\frac{(\Lambda f(x))^k}{\omega_d^k(k+1)!}\int_{(\mathbb{R}^d)^k} h_1^c(0,y)e^{-(\Lambda f(x))R^d(0,y)}\,\mathrm{d}y\,\mathrm{d}x\\
&= \int_{\mathbb{R}^d} f(x)\frac{(\Lambda f(x))^k}{\omega_d^k(k+1)!}\int_{[0,1]^d}\int_{(\mathbb{R}^d)^k} h_1^c(0,y)e^{-(\Lambda f(x))R^d(0,y)}\,\mathrm{d}y\,\mathrm{d}z\,\mathrm{d}x\\
&= \int_{\mathbb{R}^d} f(x)\gamma_k^{\mathcal{U}^d}(\Lambda f(x))\,\mathrm{d}x.
\end{aligned}$$

The first formula of the theorem then follows by taking an alternating sum as in Theorem 2.4.2.

The second formula follows from the first via the integration by parts. Namely, we have

$$\begin{aligned}
\int_{\mathbb{R}^d} f(x)\overline{\chi}_{\mathcal{U}^d}(\Lambda f(x))\,\mathrm{d}x &= \int_{\mathbb{R}^d} f(x)\overline{\chi}_{\mathcal{U}^d}(\Lambda f(x))\,\mathrm{d}x - \overline{\chi}_{\mathcal{U}^d}(0) + 1\\
&= 1 + \int_{\mathbb{R}^d} f(x)[\overline{\chi}_{\mathcal{U}^d}(\Lambda y)]_{y=0}^{y=f(x)}\,\mathrm{d}x\\
&= 1 + \int_{\mathbb{R}^d} f(x)\int_0^{f(x)} \Lambda\overline{\chi}'_{\mathcal{U}^d}(\Lambda y)\,\mathrm{d}y\,\mathrm{d}x\\
&= 1 + \int_0^{\|f\|_\infty}\int_{\mathbb{R}^d} f(x)\mathbb{1}_{f(x)\geqslant y}\Lambda\overline{\chi}'_{\mathcal{U}^d}(\Lambda y)\,\mathrm{d}x\,\mathrm{d}y\\
&= 1 + \int_0^{\|f\|_\infty} \Lambda\overline{\chi}'_{\mathcal{U}^d}(\Lambda y)\hat{f}(y)\,\mathrm{d}y\\
&= 1 + \Big[\hat{f}(y)\overline{\chi}_{\mathcal{U}^d}(\Lambda y)\Big]_{y=0}^{y=\|f\|_\infty} - \int_0^{\|f\|_\infty}\hat{f}'(y)\overline{\chi}_{\mathcal{U}^d}(\Lambda y)\,\mathrm{d}y.
\end{aligned}$$

Now, we use $\hat{f}(\|f\|_\infty) = 0$ and $\hat{f}(0)\overline{\chi}_{\mathcal{U}^d}(0) = 1\cdot 1 = 1$ to complete the proof. □

Remark 4.2.2. This result can be thought of as a law or large numbers similar to the statement about Betti numbers in [84, Theorem 1.1].

Remark 4.2.3. If we replace Euclidean balls by more general ones with respect to some p-distance, Thomas's thesis [147, Theorem 4.3.1] provides an analogous result to Theorem 2.4.2, but with an infinite series $\overline{\chi}(t) = \sum_{k=0}^{\infty}(-1)^k\psi_k(t)$, where t in the setting of that work relates to ours via $\Lambda = \omega_d t^d$. Now from parts (i) and (ii) Lemma 4.2.1 of [147], one can infer that $\sum_{k=0}^{\infty}\psi_k(t) \leqslant \exp((ct)^d\cdot\omega_d\|f\|_\infty) < \infty$. Thus, one can apply Fubini's theorem to obtain Theorem 4.2.1 in this more general setting as well.

Example 4.2.4. As a sanity check, we evaluate the integral transform formula for $F = \mathcal{U}^d$. Then, $\hat{f}(y) = \mathbb{1}_{[0,1]}(y)$ and thus $\hat{f}'(y) = \delta(y-1)$. Consequently, our formula reads as

$$\overline{\chi}_{\mathcal{U}^d}(\Lambda) = \int_0^1 \delta(y-1)\overline{\chi}_{\mathcal{U}^d}(\Lambda y)\,\mathrm{d}y = \overline{\chi}_{\mathcal{U}^d}(\Lambda),$$

which is of course tautological.

Expressing the EECC of an arbitrary density as an integral transform of the EECC of a uniform density has important implications for computations and theory. First, let us state an estimate which is a stability theorem similar to [104, Theorem 3.1] .

Corollary 4.2.5. *Let F, G be probability distributions on $\mathbb{R}^d$ admitting densities f and g, respectively. Then we have $\|\overline{\chi}_F - \overline{\chi}_G\|_\infty \leqslant \|\hat{f}' - \hat{g}'\|_1$.*

Proof. We use that $|\overline{\chi}_{\mathcal{U}^d}(\Lambda y)| \leqslant 1$ and estimate $\|\overline{\chi}_F - \overline{\chi}_G\|_\infty$ as

$$\left\| \int_0^\infty (\hat{g}'(y) - \hat{f}'(y))\,\overline{\chi}_{\mathcal{U}^d}(\Lambda y)\,\mathrm{d}y \right\|_\infty \leqslant \sup_\Lambda \int_0^\infty |\hat{g}'(y) - \hat{f}'(y)|\,|\overline{\chi}_{\mathcal{U}^d}(\Lambda y)|\,\mathrm{d}y \leqslant \int_0^\infty |\hat{g}'(y) - \hat{f}'(y)|\,\mathrm{d}y.$$

□

As a second consequence, we can find formulas for the EECC of probability densities which were previously intractable.

Example 4.2.6. Consider the standard normal distribution in two dimensions $\mathcal{N}^2$ with density $f(x_1, x_2) = \frac{1}{2\pi}\exp\left(-\frac{x_1^2+x_2^2}{2}\right)$. Due to the rotational symmetry of f, an easy application of polar coordinates shows that its excess mass is given by $\hat{f}(y)\colon [0, \frac{1}{2\pi}] \to \mathbb{R},\ y \mapsto 1 - 2\pi y$ with derivative $\hat{f}'(y) = -2\pi$. Plugging this into our formula yields

$$\overline{\chi}_{\mathcal{N}^2}(\Lambda) = -\int_0^{\frac{1}{2\pi}} -2\pi\,\overline{\chi}_{\mathcal{U}^2}(\Lambda y)\,\mathrm{d}y = \int_0^{\frac{1}{2\pi}} 2\pi\exp(-\Lambda y)(1-\Lambda y)\,\mathrm{d}y = \exp\left(-\frac{\Lambda}{2\pi}\right).$$

See Table 4.1 for more results and Figure 4.2 for corresponding plots; we omit the tedious, but straight forward calculus arguments deriving them. Observe that the EECC of a two-dimensional standard normal distribution coincides with that of a one-dimensional uniform distribution on $[0, 1/2\pi]$. However, the excess masses are different. If we fix the dimension d this cannot happen, as we shall see next.

4.3 Uniqueness of Excess Mass

In this section we establish a third consequence of Theorem 4.2.1, namely that the dependence on the excess mass is injective. This is to say, for fixed amient dimension d, the excess mass is uniquely determined by the expected ECC in the thermodynamic limit. In fact, we can use Theorem 4.2.1 to show:

f	$\overline{\chi}_F(\Lambda)$
e^{-x}	$\frac{1-e^{-\Lambda}}{\Lambda}$
$\frac{1}{\sqrt{2\pi}}\exp(-x^2/2)$	$\int_0^{\frac{1}{\sqrt{2\pi}}} \frac{2\exp(-\Lambda y)}{\sqrt{-\log(2\pi y^2)}}\,\mathrm{d}y$
$\frac{1}{2\pi}\exp\left(-\frac{x_1^2+x_2^2}{2}\right)$	$\exp\left(-\frac{\Lambda}{2\pi}\right)$
$\frac{1}{2\pi}\left(1+\frac{x_1^2+x_1^2}{n}\right)^{-\frac{n+2}{2}}$	$-\left(\frac{2\pi}{\Lambda}\right)^{\frac{n}{n+2}}\frac{n}{n+2}\left(\gamma\left(1+\frac{n}{n+2},\frac{\Lambda}{2\pi}\right)-\gamma\left(\frac{n}{n+2},\frac{\Lambda}{2\pi}\right)\right)$
$\frac{1}{4\pi}\exp\left(-\frac{(x_1^2+x_2^2+x_3^2)^{3/2}}{3}\right)$	$\frac{e^{-\Lambda/(4\pi)}(-3\Lambda^2\pi-24\Lambda(-16+\pi^2)+32(-1+e^{\Lambda/(4\pi)})\pi(-32+3\pi^2))}{128\Lambda}$

Table 4.1: Probability densities and their expected ECCs. Here, $\gamma(a,x) = \int_0^x t^{a-1}e^{-t}\,\mathrm{d}t$ is the lower incomplete gamma function. For plots, see Figure 4.2. For the one-dimensional standard normal distribution, there is no solution in terms of elementary functions.

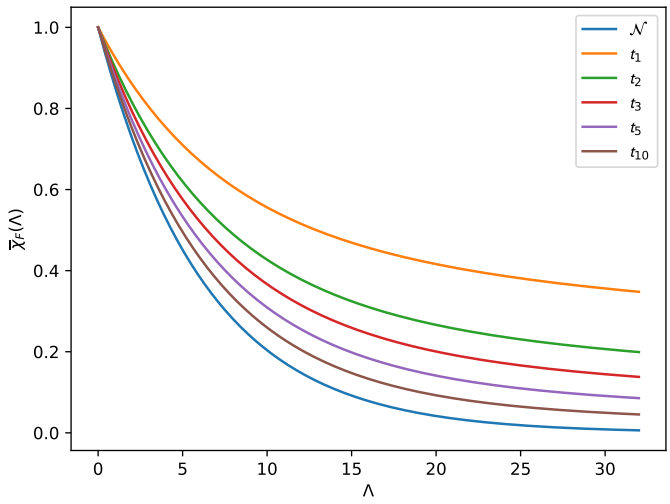


Figure 4.2: Expected ECCs in the thermodynamic limit of two-dimensional normal and t-Student distributions of various degrees of freedom. For the formulas, see Table 4.1.

Theorem 4.3.1. *Let F, G be probability distributions on $\mathbb{R}^d$ which admit densities $f, g\colon \mathbb{R}^d \to \mathbb{R}$ that are bounded. Suppose $\overline{\chi}_F(\Lambda) = \overline{\chi}_G(\Lambda)$ for all $\Lambda \geqslant 0$ and is d times differentiable in 0. Then $\hat{f} = \hat{g}$.*

Our strategy is to rewrite equation 4.3 as an ODE which both Laplace transforms $\mathcal{L}\left\{\hat{f}'\right\}$ and $\mathcal{L}\{\hat{g}'\}$ solve. Indeed, as $\overline{\chi}_{\mathcal{U}^d} = e^{-\Lambda}P(\Lambda)$ for a certain polynomial $P(\Lambda) = \sum_{i=0}^{d} p_i\Lambda^i$, formula 4.3 can be rewritten as

$$\begin{aligned}-\overline{\chi}_F(\Lambda) &= \sum_{i=0}^{d} p_i\Lambda^i\mathcal{L}\left\{\hat{f}'(y)y^i\right\}(\Lambda)\\ &= \sum_{i=0}^{d}(-1)^i p_i\Lambda^i\frac{d^i}{d\Lambda^i}\mathcal{L}\left\{\hat{f}'\right\}(\Lambda),\end{aligned}$$

using properties of the Laplace transform; see [17, chapter 7] for a textbook introduction. Then, we will infer that $\hat{f} = \hat{g}$ from the uniqueness of the solution. In order to carry this idea out, we now derive initial values which only depend on $\overline{\chi}_F = \overline{\chi}_G$ and the ambient dimension.

Lemma 4.3.2.

$$\frac{d^k}{d\Lambda^k}\mathcal{L}\left\{\hat{f}'(y)\right\}(0) = (-1)^{k-1}\frac{\overline{\chi}_F^{(k)}(0)}{\sum_{i=0}^{k}\binom{k}{i}(-1)^i P^{(k-i)}(0)} \tag{4.4}$$

Proof. First, we note that the integrand in equation 4.3 is continuously differentiable with respect to Λ, whence an application of differentiation under the integral sign yields

$$\overline{\chi}_F^{(k)}(\Lambda) = -\int_{\mathbb{R}} y^k\hat{f}'(y)e^{-\Lambda y}\sum_{i=0}^{k}\binom{k}{i}(-1)^i P^{(k-i)}(\Lambda y)\,\mathrm{d}y.$$

Here, we used the general product formula for

$$\frac{d^k}{d\Lambda^k}(P(\Lambda y)e^{-\Lambda y}) = \sum_{i=0}^{k}\binom{k}{i}y^{k-i}P^{(k-i)}(\Lambda y)(-y)^i r^{-\Lambda y} = y^k e^{-\Lambda y}\sum_{i=0}^{k}\binom{k}{i}(-1)^i P^{(k-i)}(\Lambda y).$$

On the other hand, derivatives of the Laplace transform have the following form:

$$\frac{d^k}{d\Lambda^k}\mathcal{L}\left\{\hat{f}'(y)\right\}(\Lambda) = (-1)^k\mathcal{L}\left\{y^k\hat{f}'(y)\right\}(\Lambda) = (-1)^k\int_{\mathbb{R}} y^k\hat{f}'(y)e^{-\Lambda y}\,\mathrm{d}y.$$

Our desired assertion now follows from plugging in $\Lambda = 0$:

$$\begin{aligned}\overline{\chi}_F^{(k)}(0) &= -\sum_{i=0}^{k}\binom{k}{i}(-1)^i P^{(k-i)}(0)\int_{\mathbb{R}} y^k\hat{f}'(y)\,\mathrm{d}y\\ &= (-1)^{k-1}\sum_{i=0}^{k}\binom{k}{i}(-1)^i P^{(k-i)}(0)\frac{d^k}{d\Lambda^k}\mathcal{L}\left\{\hat{f}'(y)\right\}(0).\end{aligned}$$

Note that we can do this although $\overline{\chi}$ is only defined for $\Lambda \geqslant 0$ (which means that the derivative is only right-sided) because the right-hand side of equation 4.3 is also defined for $\Lambda < 0$ and continuously differentiable in 0. □

Remark 4.3.3. It is not hard (employing integration by parts like before) to compute the expression arising in the proof: $\int_{\mathbb{R}} y^k \hat{f}'(y)\ \mathrm{d}y = \|f\|_{k+1}^{k+1}$. This can be used to derive the bounds $|\frac{d^k}{d\Lambda^k}\mathcal{L}\left\{\hat{f}'(y)\right\}(\Lambda)| \leqslant \|f\|_{k+1}^{k+1}$ and evaluate $\frac{d^k}{d\Lambda^k}\mathcal{L}\left\{\hat{f}'(y)\right\}(0) = \|f\|_{k+1}^{k+1}$, but we shall not need this result here.

Proof of Theorem 4.3.1. Recall that the we can rewrite equation (4.3) from Theorem 4.2.1 in terms of the Laplace transform as the following linear ODE:

$$-\overline{\chi}_F = \sum_{i=0}^{d} (-1)^i p_i \Lambda^i \frac{d^i}{d\Lambda^i} \mathcal{L}\left\{\hat{f}'\right\}. \tag{4.5}$$

Here, $P(\Lambda) = \sum_{i=0}^{d} p_i \Lambda^i$ is the polynomial defined by $\overline{\chi}_{\mathcal{U}^d}(\Lambda) = e^{-\Lambda} P(\Lambda)$.

Moreover, Lemma 4.3.2 provides initial values in Equation (4.4). As d is fixed, so are the coefficients p_i and because $p_0 = 1$, they are not all zero. Therefore, on every compact interval, Picard-Lindelöf guarantees that $\mathcal{L}\left\{\hat{f}'\right\}$ is the unique solution.

Finally, if $\overline{\chi}_F(\Lambda) = \overline{\chi}_G(\Lambda)$ for all $\Lambda > 0$ as in the assumption of Theorem 4.3.1, $\mathcal{L}\left\{\hat{f}'\right\}$ and $\mathcal{L}\{\hat{g}'\}$ both satisfy the ODE 4.5. In addition, they have the same initial values, given in Equation 4.4, which only depend on $\overline{\chi}_F(\Lambda) = \overline{\chi}_G(\Lambda)$ and the ambient dimension. Consequently, we infer that $\mathcal{L}\left\{\hat{f}'\right\} = \mathcal{L}\{\hat{g}'\}$. By injectivity of the Laplace transform, this means $\hat{f}' = \hat{g}'$. Now, since $\hat{f}(0) = 1 = \hat{g}(0)$ because f and g are probability densities, we conclude that $\hat{f} = \hat{g}$, as desired. □

For $d = 1, 2$, one can write down quite explicit solutions: In the one-dimensional case, $-\overline{\chi}_F = \mathcal{L}\left\{\hat{f}'\right\}$, so that $\hat{f}(y) = 1 - \int_0^y \mathcal{L}^{-1}\{\overline{\chi}_F\}(t)\ \mathrm{d}t$. In the two-dimensional case, our differential equation simplifies to

$$-\overline{\chi}_F = \frac{d}{d\Lambda}\left(\Lambda \mathcal{L}\left\{\hat{f}'\right\}(\Lambda)\right),$$

and therefore,

$$\hat{f} = -\mathcal{L}^{-1}\left\{\frac{1}{s} - \frac{1}{s^2}\int_0^s \overline{\chi}_F(\Lambda)\ \mathrm{d}\Lambda\right\}.$$

While one might like to use these ideas to estimate $\hat{f}$ from empirical estimates of the EECC, this is unfortunately impossible in practice. The usual Fixed Talbot algorithm [1] for numerically computing inverse Laplace transforms is numerically quite unstable and cannot handle noisy input data one encounters in empirical EECCs.

Remark 4.3.4. If one replaces the Euclidean metric by the supremum distance for the collection of balls, [147, Corollary 4.3.3] presents the following expression for the EECC of the uniform distribution [147, eqn. (4.11)]:

$$\overline{\chi}_{\mathcal{U}^d}(\Lambda) = -\frac{e^{-\Lambda/\omega_d}}{\Lambda/\omega^d} T_d(-\Lambda/\omega_d).$$

Here, T_p is the Touchard polynomial of degree p. Now, using the variable $\lambda = \Lambda/\omega_d$, one can argue with the Laplace transform again to establish an analogue to Theorem 4.3.1.

4.4 Outlook

To conclude this chapter, we outline two major directions for future research.

First, having established a necessary condition for the expected ECCs to coincide raises the question whether this condition is also necessary in order for the centered ECCs to coincide in distribution (Vishwanath et al. [157] showed it to be sufficient). To this end, it is tempting to try a similar approach for higher moments, starting from variance. While an analogue of Theorem 4.2.1 is readily established using the description of $\lim_{n\to\infty} n^{-1}\operatorname{Var}(\chi_F(\Lambda))$ of [28], the strategy to prove Theorem 4.3.1 cannot be replicated. This is because, unfortunately, there is no analogous expression to $\overline{\chi}_{\mathcal{U}^d} = e^{-\Lambda}P(\Lambda)$, for a certain polynomial $P(\Lambda) = \sum_{i=0}^{d} p_i\Lambda^i$.

Second, it would be interesting to have a quantitative version of Theorem 4.3.1 in the following sense: Is it possible to compute (or at least bound) the supremum distance $\|\hat{f} - \hat{g}\|_\infty$ in terms of expected ECCs? Recall that $1 - \hat{f}$ is the cumulative distribution function of the random variable $f(X)$ where $X \sim F$. Thus, $\|\hat{f} - \hat{g}\|_\infty$ is a Kolmogorov-Smirnov test statistic for the null hypothesis $f(X) \overset{D}{=} g(Y)$, where $X \sim F$, $Y \sim G$. This could pave the way towards a distribution-free multivariate two sample test using computational topology. Moreover, such a result would imply that the injective continuous map $\hat{f}' \mapsto \int_0^\infty \hat{f}'(y)\overline{\chi}_{\mathcal{U}}(\Lambda y)\,\mathrm{d}y$ is in addition a homeomorphism onto its image.

5 Topology-Driven Goodness-of-Fit Tests in Arbitrary Dimensions

Abstract. This chapter adopts a tool from computational topology, the Euler characteristic curve (ECC) of a sample, to perform one- and two-sample goodness of fit tests. We call our procedure TopoTests. The presented tests work for samples of arbitrary dimension, having comparable power to the state-of-the-art tests in the one-dimensional case. It is demonstrated that the type I error of TopoTests can be controlled and their type II error vanishes exponentially with increasing sample size. Extensive numerical simulations of TopoTests are conducted to demonstrate their power for samples of various sizes.

Author's contributions. This chapter contains joint work with Paweł Dłotko, Łukasz Stettner and Rafał Topolnicki published as [63]; the exposition is slightly revised according to the results of Chapter 4, which were originally obtained after the results of this chapter. The project was conceived by P.D, N.H. and R.T. and carried out by N.H. and R.T. as co-lead authors, under supervision of P.D., with Lemma 5.1.2 contributed by Ł.S..

In the paper corresponding to this chapter, to the best of our knowledge, we present the first mathematically rigorous approach using Euler characteristic curves to perform general goodness-of-fit testing. Specifically, we consider the Čech (or, equivalently, Alpha) complex of a sample, which we scale in such a way that asymptotically, we can employ the theoretical results about the thermodynamic regime. Our procedure is theoretically justified by Theorem 5.1.4. The concentration inequality for Gaussian processes (Lemma 5.1.2) might be of independent interest.

Simulations conducted in Section 5.3 and 5.4 indicate that TopoTest outperforms the Kolmogorov-Smirnov test we used as a baseline in arbitrary dimension both in terms of the test power but also in terms of computational time for moderate sample sizes and dimensions, and for a wide variety of null and alternative distributions.

The implementation of TopoTest is publicly available at `https://github.com/dioscuri-tda/topotests`.

5.1 Method

5.1.1 One-sample test

Consider the following setup: In ambient Euclidean space $\mathbb{R}^d$, we are given a fixed null distribution F and a sample X following an unknown distribution G, we assume that they admit bounded densities f and g, respectively. Then, in the light of the result of the preceding chapter, we aim to test the following hypothesis, in which $\hat{f}$ again denotes the excess mass (Definition 4.1.1):

$$H_0 \colon \hat{g} = \hat{f} \qquad \text{vs.} \qquad H_1 \colon \hat{g} \neq \hat{f}. \tag{5.1}$$

Compare this formulation to the problem stated in (1.2). The perhaps surprising feature of our approach is that while the hypothesis is phrased in an *analytic* language (equality of certain integrals), we test it using a *topological* method. As the ECC of the Alpha and Čech complexes are equal, they can be used interchangeably. We will phrase the theory in terms of Čech and algorithms and computational results in terms of Alpha.

Remark 5.1.1. More generally, the test works for any filtered simplicial complex $\mathcal{K}$ built on the sample points as vertices, for instance Vietoris-Rips, as long as its Euler characteristic curve satisfies a functional central limit theorem analogous to Theorem 2.4.3. However, as the results from the previous chapter only pertain to Čech complexes (or equivalent), it is unclear against which kinds of distribution a test using $\chi(\mathcal{K})$ would have power. To account for this, say that two probability distributions F, G are *Euler equivalent with respect to* $\mathcal{K}$, if

$$\lim_{n\to\infty} n^{-1}\mathbb{E}[\chi(\mathcal{K}(X_n)_{r_n})] = \lim_{n\to\infty} n^{-1}\mathbb{E}[\chi(\mathcal{K}(Y_n)_{r_n})],$$

where X_n and Y_n consist of n i.i.d. points sampled from F and G, respectively, and $n \cdot r_n \to \lambda \in]0, \infty[$. The null hypothesis will then be phrased as

$$H_0 : F, G \text{ are Euler equivalent with respect to } \mathcal{K}.$$

We conjecture that for a large class of simplicial constructions including Vietoris-Rips, Euler equivalence is tantamount to distributions admitting the same excess mass.

We write

$$\chi(n, r) = \chi(\mathcal{C}(X)_r),$$

where n is the cardinality of X. Given some distribution F on $\mathbb{R}^d$ against which we want to test, we are interested in the expected ECC of the Čech complex of scale r of n i.i.d. points drawn according to F, scaled by $n^{1/d}$, denoted as $\mathbb{E}_F(\chi(n, r))$. The TopoTest employs the supremum distance between the ECC computed based on sample points scaled by $n^{1/d}$, that is $\chi(\mathcal{C}(X)_r)$, and the expected ECC, $\mathbb{E}_F(\chi(n, r))$, under H_0, i.e. the test statistic is

$$\Delta_n := n^{-1/2} \sup_{r\in[0,T]} |\chi(\mathcal{C}(X)_r) - \mathbb{E}_F(\chi(n, r))|, \tag{5.2}$$

where $T \in \mathbb{R}^+$. (The restriction to a compact interval is needed for the functional central limit theorem (Theorem 2.4.3).) Therefore, by using ECC as topological summary of the dataset we reduce the initial d-dimensional problem to a one-dimensional setting. If Δ_n defined in (5.2) is large enough the null hypothesis is rejected, while for small values of Δ_n the test fails to reject the H_0. More precisely: given the significance level α we consider a rejection region $R_\alpha = [t_\alpha, \infty[$ such that

$$\begin{aligned} &\mathbb{P}(\Delta_n \in R_\alpha | H_0) \\ &= \mathbb{P}\left(n^{-1/2} \sup_{r\in[0,T]} |\chi(\mathcal{C}(X)_r) - \mathbb{E}_F(\chi(n, r))| > t_\alpha \middle| H_0 \right) \\ &= \alpha. \end{aligned} \tag{5.3}$$

The threshold value t_α depends on the significance level α and F (and hence also on dimension d), however the dependence on F is dropped in the notation. We prove that this test is consistent below in Section 5.1.3.

5.1.2 Two-sample test

A test statistic based on Euler characteristic curves can also be adapted to the two-sample problem. Given two samples $X, Y \subset \mathbb{R}^d$ of possibly different sizes, following unknown distributions $X \sim F$ and $Y \sim G$, we are testing the null hypothesis $H_0 \colon \hat{g} = \hat{f}$, where f, g are the densities of F, G, respectively. The test statistic in this setting is the supremum distance between the normalized ECCs

$$\Delta(\chi(X), \chi(Y)) = \sup_{r\in[0,T]} \left| \frac{1}{|X|}\chi(\mathcal{C}(X)_r) - \frac{1}{|Y|}\chi(\mathcal{C}(Y)_r) \right|.$$

Moreover, recall that we rescale the samples to have a fixed average number of points in a ball of radius r, independently of the sample size. Since the null distribution is unknown, we fall back on a permutation test [10, Section 16.3] to compute the p-value, see Algorithm 5.2 for the details.

As for any permutation test, the procedure is computationally expensive as it requires computing ECCs for a variety of point sets resampled from the union of the two input datasets. The application of this approach is therefore limited to rather small sizes of input data sets. See Section 5.4 for results

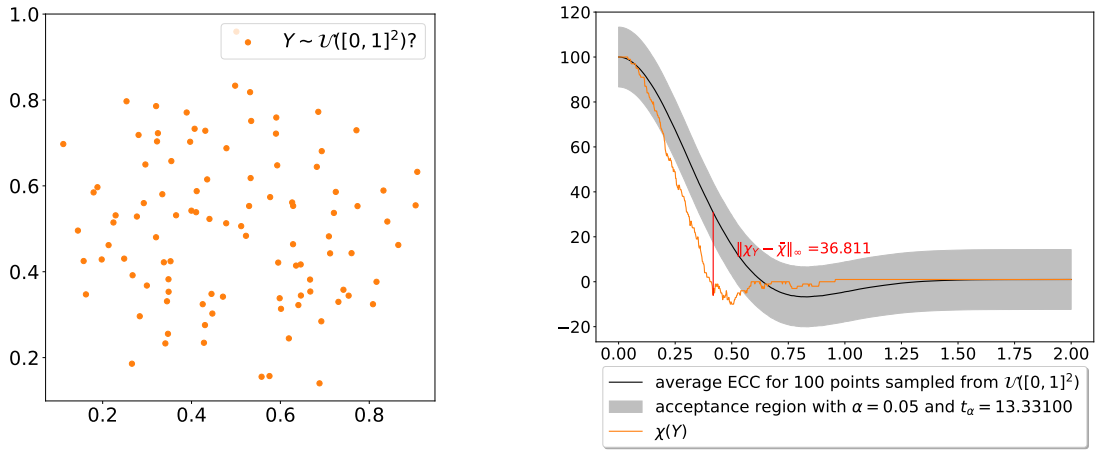


Figure 5.1: Example of the one-sample setting, testing the hypothesis that the 100 points on the left are following a uniform distribution. The expected ECC for 100 points following a uniform distribution is shown in black on the right, and 95% of samples from the null distribution yield an ECC falling within the grey acceptance region. Since the empirical ECC of our sample, drawn in orange, goes outside this region, we have evidence to reject the null hypothesis.

of a simulation study in which the performance of this approach is compared with the two-sample Kolmogorov-Smirnov test.

5.1.3 Power of the One-Sample Test

Overview

The TopoTest relies on the Functional Central Limit Theorem of Krebs et al. [104, Theorem 3.4], hence it works under the following, rather technical, assumption

Assumption 1. *The null distribution has compact convex support inside* $[0,1]^d$*. It admits a bounded density* κ *that can be uniformly approximated by blocked functions* κ_n*.*

Recall from [104, equation 3.8], that the approximation by blocked functions means $\lim_{n\to\infty} \|\kappa - \kappa_n\| = 0$, where each κ_n is constant on grid elements of a partition of the unit hypercube $[0,1]^d$ into an equidistant grid of m^d subcubes. In particular, bounded measurable functions satisfy this assumption.

We will show, for a fixed significance level α, that the mean of the test statistic Δ_n does not grow with n under the null hypothesis, while it grows at least like $\sqrt{n}$ under the alternative hypothesis. Moreover, in both cases Δ_n is concentrated around its mean allowing to control the type II error of the TopoTests.

Case H_0 true

Under the null hypothesis $\hat{f} = \hat{g}$, we have $\lim_{n\to\infty} \mathbb{E}_F(\chi(n,r)) = \lim_{n\to\infty} \mathbb{E}_G(\chi(n,r))$ due to [157]. Thus, by [148] and [104, Theorem 3.4]) (cf. Theorem 2.4.3), we have the following convergence in distribution in the Skorokhod J_1-topology to a centered Gaussian process f_r,

$$n^{-1/2}(\chi(\mathcal{C}(X)_r) - \mathbb{E}_F(\chi(n,r))) \xrightarrow[n\to\infty]{D} f_r. \tag{5.4}$$

Here it is assumed that the sample is drawn from a distribution satisfying Assumption 1 and scaled by $n^{1/d}$ (so that asymptotically, the distribution is governed by the thermodynamic regime; recall the discussion in section 2.4). Note that by [157], the distribution and covariance structure of the limiting Gaussian process only depend on the excess mass of the distribution from which the points are sampled. Let us denote

$$Z_T = \sup_{r\in[0,T]} |f_r|.$$

In the following we will approximate the finite-sample distribution of $n^{-1/2}(\chi(\mathcal{C}(X)_r) - \mathbb{E}_F(\chi(n,r)))$ by the limiting Gaussian process f_r. Therefore, for sufficiently large n we assume that

$$\Delta_n \overset{D}{=} Z_T. \tag{5.5}$$

The quality of this approximation was studied numerically – please refer to Figure 5.2. Intuitively speaking, the distribution of the tests statistic for finite samples is similar to the theoretical limit distribution of the supremum of the Gaussian process.

For Z_T we have the Borell-TIS inequality[1] [4, Section 2.1],

$$\begin{aligned} \mathbb{P}(Z_T > t) &= \mathbb{P}\left(\sup_{r\in[0,T]} |f_r| > t\right) \\ &\leqslant \exp\left(-\left[t - \mathbb{E}\left(\sup_{r\in[0,T]} |f_r|\right)\right]^2 / 2\sigma_T^2\right), \end{aligned} \tag{5.6}$$

where $\sigma_T^2 = \sup_{r\in[0,T]} \mathbb{E}(f_r^2)$.

[1] The abbreviation stands for Tsirelson, Ibragimov, and Sudakov, who discovered the inequality independently of Borell.

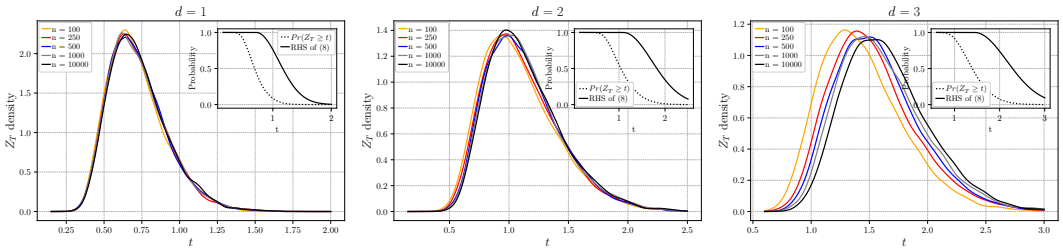


Figure 5.2: Numerical inspection of the quality of finite sample approximation (5.5). The empirical distribution of Z_T converges with the increasing sample size. Even for the three-dimensional case, the distribution obtained for $n = 100$ is a reasonable approximation for large-sample empirical distribution. An inset in each plot shows left- and right-hand side of the inequality (5.6) – this provides another justification for approximation (5.5).

Therefore, for n large enough,

$$\begin{aligned}
&\mathbb{P}(\Delta_n > t | H_0) \\
&= \mathbb{P}\left(\sup_{r\in[0,T]} \left| \frac{\chi(\mathcal{C}(X)_r) - \mathbb{E}_F(\chi(n,r))}{\sqrt{n}} \right| > t \right) \\
&\leqslant \exp\left(-\left[t - \mathbb{E}\left(\sup_{r\in[0,T]} \frac{1}{\sqrt{n}} |\chi(\mathcal{C}(X)_r) \right.\right.\right. \\
&\qquad\qquad \left.\left.\left. - \mathbb{E}_F(\chi(n,r))| \right) \right]^2 / 2\sigma_T^2 \right).
\end{aligned}$$

Plugging in (5.3) yields

$$\begin{aligned}
\alpha \leqslant \exp\left(-\left[t_\alpha - \mathbb{E}\left(\sup_{r\in[0,T]} \frac{1}{\sqrt{n}} |\chi(\mathcal{C}(X)_r) \right.\right.\right. \\
\left.\left.\left. - \mathbb{E}_F(\chi(n,r))| \right) \right]^2 / 2\sigma_T^2 \right)
\end{aligned}$$

which leads to

$$\begin{aligned}
t_\alpha \leqslant & \sqrt{-2\sigma_T^2 \ln(\alpha)} \\
& + \mathbb{E}\left(\sup_{r\in[0,T]} \frac{1}{\sqrt{n}} |\chi(\mathcal{C}(X)_r) - \mathbb{E}_F(\chi(n,r))| \right),
\end{aligned} \tag{5.7}$$

i.e. $t_\alpha = O(1)$.

Case H_0 false

Now let us study the asymptotic size of

$$\sup_{r\in[0,T]} |\chi(\mathcal{C}(Y)_r) - \mathbb{E}_F(\chi(n,r))|$$

as $n \to \infty$ when $Y \sim G$, and $\hat{g} \neq \hat{f}$.

We have

$$\begin{aligned}
&\mathbb{E}\left(\sup_{r\in[0,T]} |\chi(\mathcal{C}(Y)_r) - \mathbb{E}_F(\chi(n,r))|\right) \\
&\geqslant \sup_{r\in[0,T]} \mathbb{E}|\chi(\mathcal{C}(Y)_r) - \mathbb{E}_F(\chi(n,r))| \\
&\geqslant \sup_{r\in[0,T]} |\mathbb{E}_G(\chi(n,r)) - \mathbb{E}_F(\chi(n,r))|.
\end{aligned}$$

Because the limiting expectations of the ECCs are different under the alternative hypothesis (this is where we need Theorem 4.3.1 from the previous chapter), this last expression diverges. Due to [31], Corollary 4.5, $E_F(\chi(n,r)) \sim n$ with constant depending on F and d. In our setting, we obtain

$$\mathbb{E}\left(\sup_{r\in[0,T]} n^{-1/2}|\chi(\mathcal{C}(Y)_r) - \mathbb{E}_F(\chi(n,r))|\right) = \Omega(\sqrt{n}). \tag{5.8}$$

To complete the discussion, it is required to show that in the case of H_0 false, one also has a concentration around the mean, i.e. one needs to control

$$\begin{aligned}
C_{F,G}(t) = \mathbb{P}\Bigg(n^{-1/2}\Bigg|&\sup_{r\in[0,T]} |\chi(\mathcal{C}(Y)_r) - \mathbb{E}_F(\chi(n,r))| \\
&-\mathbb{E}\left(\sup_{r\in[0,T]} |\chi(\mathcal{C}(Y)_r) - \mathbb{E}_F(\chi(n,r))|\right)\Bigg| > t\Bigg).
\end{aligned} \tag{5.9}$$

The lemma below provides a generalization of the Borell-TIS inequality to the case of non-centred Gaussian process.

Lemma 5.1.2. *Let f_r be a centred Gaussian process and $g(r)$ some deterministic function. We have*

$$\begin{aligned}
\mathbb{P}\left(\left|\sup_{r\in[0,T]} |f_r + g(r)| - \mathbb{E}\left(\sup_{r\in[0,T]} |f_r + g(r)|\right)\right| > t\right) \\
\leqslant 2e^{-t^2/2\sigma^2},
\end{aligned} \tag{5.10}$$

where $\sigma = \sup_{r\in[0,T]}\left(\mathbb{E}[f_r^2]\right)^{1/2}$.

Proof. We follow the strategy of Ledoux [106, Section 7.1]. Argument (2.35) in Ledoux [106] yields that if γ is a standard Gaussian measure on $\mathbb{R}^n$ then for every 1-Lipschitz function F on $\mathbb{R}^n$ and $t \geqslant 0$ we have

$$\gamma\left(\left\{F \geqslant \int F d\gamma + t\right\}\right) \leqslant e^{-t^2/2}. \tag{5.11}$$

Let $r_1, \ldots, r_n$ be fixed in $[0, T]$ and consider centered Gaussian random vector $(f_{r_1}, \ldots, f_{r_n})$ in $\mathbb{R}^n$ with covariance matrix $\Gamma = B^T B$. Consequently, the law of $(f_{r_1}, \ldots, f_{r_n})$ is the same as the law of $B\mathcal{N}$ where $\mathcal{N} = (N_1, \ldots, N_n)^T$ is distributed according to the standard Gaussian measure γ on $\mathbb{R}^n$. Let $F : \mathbb{R}^n \to \mathbb{R}$ be defined as

$$F(x) = \max_{1 \leqslant i \leqslant n} |(Bx)_i + g(r_i)|, x \in \mathbb{R}^n.$$

Although we have a different F in our setting than [106], we can still bound the Lipschitz norm of F to be at most the operator norm of $B \colon (\mathbb{R}^n, \|\cdot\|_2) \to (\mathbb{R}^n, \|\cdot\|_\infty)$. Indeed, consider any $c > 0$ such that $\|Bx\|_\infty \leqslant c\|x\|_2$ for all $x \neq 0$. Using the triangle inequality, we estimate that for any $x \neq y \in \mathbb{R}^n$,

$$\begin{aligned}
|F(x) - F(y)| &= \Big| \max_{1 \leqslant i \leqslant n} |(Bx)_i + g(r_i)| \\
&\qquad - \max_{1 \leqslant i \leqslant n} |(By)_i + g(r_i)| \Big| \\
&\leqslant \max_{1 \leqslant i \leqslant n} |(Bx)_i + g(r_i) - (By)_i - g(r_i)| \\
&= \max_{1 \leqslant i \leqslant n} |(B(x-y)_i| \\
&\leqslant c\|x - y\|_2.
\end{aligned}$$

Notice that $f_{r_i} = \sum_{j=1}^n B_{ij} N_j$ and by independence of $\{N_j\}_{1 \leqslant j \leqslant n}$ we have $\mathbb{E} f_{(r_i)}^2 = \sum_{j=1}^n B_{ij}^2$. This allows us to bound the operator norm of B as follows:

$$\begin{aligned}
\|B\|_{op} &= \max_{1 \leqslant i \leqslant n} \left(\sum_{j=1}^n B_{ij}^2\right)^{1/2} = \max_{1 \leqslant i \leqslant n} \left(\mathbb{E}(f_{(r_i)}^2)\right)^{1/2} \\
&\leqslant \sup_{r \in [0,T]} \left((\mathbb{E}(f_{(r_i)}^2)\right)^{1/2} = \sigma.
\end{aligned}$$

Consequently, F/σ is 1-Lipschitz and by (5.11) we have

$$\mathbb{P}\left(\frac{1}{\sigma} F(\mathcal{N}) - \mathbb{E}[\frac{1}{\sigma} F(\mathcal{N})] \geqslant \tilde{t}\right) \leqslant e^{-\tilde{t}^2/2}$$

Letting $t = \sigma\tilde{t}$ and by symmetry argument we obtain

$$\mathbb{P}(|F(\mathcal{N}) - \mathbb{E}(F(\mathcal{N}))| \geqslant t) \leqslant 2e^{-t^2/2\sigma^2}$$

and

$$\mathbb{P}\left(\left|\sup_{1\leqslant i\leqslant n}|f_{r_i}+g(r_i)|-\mathbb{E}\left(\sup_{1\leqslant i\leqslant n}|f_{r_i}+g(r_i)|\right)\right|\geqslant t\right)\leqslant 2e^{-t^2/2\sigma^2}.$$

The right hand side does not depend on $f(r_i)$, hence letting $n\to\infty$, inequality (5.10) is obtained. □

Using the Lemma 5.1.2 we obtain following theorem

Theorem 5.1.3. *Concentration around the mean $C_{F,G}(t)$, defined in (5.9), is exponentially bounded*

$$C_{F,G}(t)\leqslant 2e^{-t^2/2\sigma_G^2}. \tag{5.12}$$

Proof. Subtracting and adding $\mathbb{E}_G(\chi(n,r))$ in (5.9) yields

$$\begin{aligned}
&C_{F,G}(t)\\
&=\mathbb{P}\Bigg(n^{-1/2}\Bigg|\sup_{r\in[0,T]}|\chi(\mathcal{C}(Y)_r)-\mathbb{E}_F(\chi(n,r))|\\
&\qquad-\mathbb{E}\left(\sup_{r\in[0,T]}|\chi(\mathcal{C}(Y)_r)-\mathbb{E}_F(\chi(n,r))|\right)\Bigg|>t\Bigg)\\
&=\mathbb{P}\Bigg(n^{-1/2}\Bigg|\sup_{r\in[0,T]}|\chi(\mathcal{C}(Y)_r)-\mathbb{E}_G(\chi(n,r))\\
&\qquad\qquad+\mathbb{E}_G(\chi(n,r))-\mathbb{E}_F(\chi(n,r))|\\
&\qquad-\mathbb{E}\Bigg(\sup_{r\in[0,T]}|\chi(\mathcal{C}(Y)_r)-\mathbb{E}_G(\chi(n,r))\\
&\qquad\qquad+\mathbb{E}_G(\chi(n,r))-\mathbb{E}_F(\chi(n,r))|\Bigg)\Bigg|>t\Bigg)\\
&=\mathbb{P}\Bigg(\Bigg|\sup_{r\in[0,T]}|g_r+h(r)|\\
&\qquad-\mathbb{E}\left(\sup_{r\in[0,T]}|g_r+h(r)|\right)\Bigg|>t\Bigg),
\end{aligned}$$

where the notation

$$\begin{aligned}
g_r&=(\chi(\mathcal{C}(Y)_r)-\mathbb{E}_G(\chi(n,r)))/\sqrt{n},\\
h(r)&=(\mathbb{E}_G(\chi(n,r))-\mathbb{E}_F(\chi(n,r)))/\sqrt{n}
\end{aligned}$$

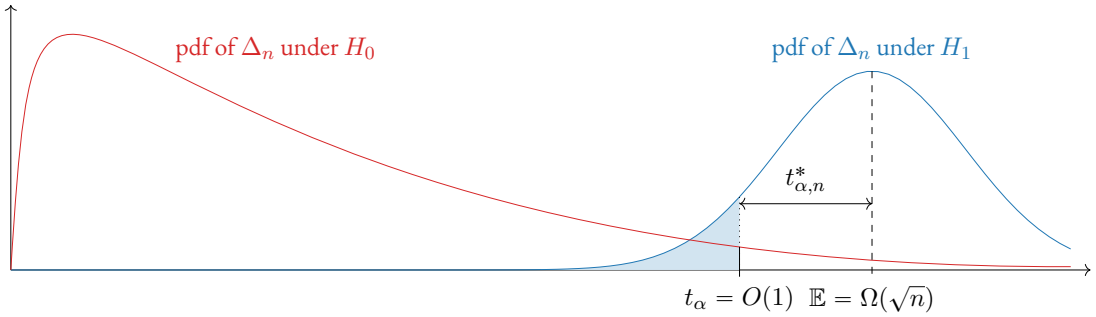


Figure 5.3: The area of shaded blue region is the probability of a type II error occuring. As $n \to \infty$, it goes to zero.

was introduced. Note that by (5.4) applied for distribution G the g_r converges to a centred Gaussian process, whereas $h(r)$ is a deterministic function. Let $\sigma_G^2 = \sup_{r\in[0,T]} \mathbb{E}(g_r^2)$. Therefore, the bound (5.12) is obtained by Lemma 5.1.2 using the same assumption as in (5.5). □

The rate of type I error is controlled by the significance level α. An asymptotic upper bound for type II error is given by the following theorem.

Theorem 5.1.4. *For fixed α, the probability of a type II error goes to 0 exponentially as $n \to \infty$.*

Proof. We will use the threshold t_α defined in (5.3) and the concentration inequality of Theorem 5.1.3. The idea is illustrated in Figure 5.3. Introduce

$$t^*_{\alpha,n} = \mathbb{E}\left(\sup_{r\in[0,T]} n^{-1/2} |\chi(\mathcal{C}(Y)_r) - \mathbb{E}_F(\chi(n,r))| \right) - t_\alpha.$$

Due to equation (5.8), the first term above is of order $\Omega(\sqrt{n})$ while second term is of order $O(1)$, therefore $t^*_{\alpha,n} = \Omega(\sqrt{n})$ and is positive for sufficiently large n. Hence we can estimate

$$\begin{aligned}
&\mathbb{P}(\text{type II error}) \\
&\leqslant \mathbb{P}\left(\sup_{r\in[0,T]} n^{-1/2}|\chi(\mathcal{C}(Y)_r) - \mathbb{E}_F(\chi(n,r))| < t_\alpha\right) \\
&= \frac{1}{2}\mathbb{P}\left[n^{-1/2}\left|\mathbb{E}\left(\sup_{r\in[0,T]}|\chi(\mathcal{C}(Y)_r) - \mathbb{E}_F(\chi(n,r))|\right)\right.\right. \\
&\qquad\qquad \left.\left. - \sup_{r\in[0,T]}|\chi(\mathcal{C}(Y)_r) - \mathbb{E}_F(\chi(n,r))|\right| > t^*_{\alpha,n}\right] \\
&\leqslant \exp\left(\frac{-{t^*_{\alpha,n}}^2}{2\sigma^2}\right) \sim e^{-n} \to 0.
\end{aligned}$$

□

5.1.4 Properties of the TopoTests

TopoTests rely on the Euler characteristic curve which is computed based on the Alpha complex of the input sample. The Alpha complex captures distance patterns between all data points in the samples. Therefore, TopoTest is not capable to discriminate distributions that differ only by translation, reflection or rotation, or, more generally, admit the same excess mass (cf. Chapter 4). As a consequence TopoTest, contrary to Kolmogorov-Smirnov, is not able to distinguish between e.g. $\mathcal{N}\left((0,0), \begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}\right)$ and $\mathcal{N}\left((\mu_1,\mu_2), \begin{bmatrix}1 & \alpha\\ \alpha & 1\end{bmatrix}\right)$, $\alpha \in [-1,0)\cup(0,1]$ as those distributions are equivalent up to translation and rotation. As a consequence, the alternative hypotheses in Kolmogorov-Smirnov and TopoTest are in fact slightly different: in the former we have $H_1 : G \neq F$ while in later the inequality is phrased only in terms of excess mass, cf. Equation (5.1). The same discussion also applies to the null hypothesis. Hence, such pairs of distributions were excluded from the forthcoming numerical study.

5.1.5 Non-Compactly Supported Distributions

The results on the asymptotic convergence presented in Section 5.1.3 work for compactly supported distributions. However, most of the distributions considered in practice, starting from normal distributions, are defined on non–compact support and the presented results do not apply to them directly. There are a number of ways we can adjust such a distribution so that the presented methodology applies. In what follows we discuss three possible strategies, starting from the one we consider the most practical one

1. *Restricting a distribution to a compact subset;*
   In this case, the given distribution is restricted to a compact rectangle. In our case we choose a symmetric rectangle $[-a,a]^d$ for a being the maximal representable double precision number.

This ensures that every sample that can be analyzed in a computer is automatically coming from such a restricted distribution. We note that, formally, such a restricted distribution need to be rescaled to become a probability distribution. However, in all practically relevant cases we are aware of, such a restricted distribution will be infinitesimally close, on its domain, to the original one, defined on an unbounded domain. Therefore, we argue that in practice, the presented methods can be applied even to distributions with no compact support. Additionally, the simulations performed provide strong evidence for this claim.

2. *Rescaling a distribution to a compact subset;*
   Here a transformation, $\arctan(\gamma x) : \mathbb{R} \to [-\frac{\pi}{2}, \frac{\pi}{2}]$ is applied separately to each coordinate to map the unbounded domain to a compact region.

   We observe that for $x \in [-2, 2]$, or for any similar interval centered around zero, $\arctan(x)$ is close to a linear function, hence the distance between points before and after applying the map, should be proportional to each other regardless of the points. To keep such a distortion of distances between points before and after rescaling, the scaling parameter γ is used. For instance, we may choose it in the way that 10 standard deviations in our data, after divided by γ, have values in the interval $[-2, 2]$. For multivariate distributions the scaling can be applied separately in each dimension. Such a rescaling does not have any major impact on the powers of the tests as discussed in Sections 5.3 and 5.4. At the same time, it allows to map any unbounded distribution to a compact domain. One should note, however, that a distribution with bounded density, transformed by $\arctan$ may have, in some pathological cases, unbounded density and thus violate Assumption 1. Hence, before using this transformation, the boundedness of the output density needs to be verified.

3. *Transforming into copula;*
   The marginals $F_1, \ldots, F_d$ of the distribution F are continuous, hence one can apply the probability integral transform [40] to each component of the random vector X sampled form a distribution F. Then the random vector

$$(U_1, \ldots, U_d) = (F_1(X_1), \ldots, F_d(X_d)) \tag{5.13}$$

   is supported on a unit cube $[0, 1]^d$ and has uniformly distributed marginals. The joint distribution function of $(U_1, \ldots, U_d)$ forms a copula. Since the null distribution F is given, the marginal distributions $F_1, \ldots, F_d$ can be derived. The transformation (5.13) must be applied to both the sample and null distribution F. Transformation (5.13) preserves the correlation structure and transforms the initial distribution F onto a compact support fulfilling the Assumption 1. Although such transformation is easy to compute and quite general, simulation studies showed that the power of resulting test is significantly reduced.

5.2 Algorithms

5.2.1 One-Sample Test

The test statistic for one-sample TopoTest, Δ defined in (5.2), involves $\mathbb{E}_F(\chi(n, r))$ being the ECC expected under H_0. There is no compact formula that can be applied to compute $\mathbb{E}_F(\chi(n, r))$ for an

arbitrary distribution function F in arbitrary dimension d although some some asymptotic formulas are available in Table 4.1. However one can use the empirical approximation of $\mathbb{E}_F(\chi(n,r))$ based on average ECC computed on a collection of randomly generated ECCs. Notice that $\chi(\mathcal{A}(X)_r)$ can only take on finitely many values because the underlying sample is finite. Therefore, $\mathbb{E}_F(\chi(n,r))$ is finite. The strong law of large numbers applies and we can approximate this expectation empirically, i.e. let $Y_1, \dots Y_M$ be i.i.d. samples each consisting of n points drawn i.i.d. from F, then

$$\widehat{\mathbb{E}}_F(\chi(n,r)) := \sum_{i=1}^{M} \frac{\chi(\mathcal{A}(Y_i)_r)}{M} \xrightarrow[M\to\infty]{a.s.} \mathbb{E}_F(\chi(n,r)). \tag{5.14}$$

Due to the continuous mapping theorem, the above point-wise convergence result allows us to use an empirical estimate $\widehat{\mathbb{E}}_F(\chi(n,r))$ instead of $\mathbb{E}_F(\chi(n,r))$ in practice when computing the statistic Δ_n leading to statistic

$$\begin{aligned} \widehat{\Delta}_n &:= \widehat{\Delta}(\chi(\mathcal{A}(X)), \widehat{\mathbb{E}}_F(\chi(n,r))) \\ &:= \sup_{r\in[0,T]} \frac{1}{\sqrt{n}} |\chi(\mathcal{A}(X)_r) - \widehat{\mathbb{E}}_F(\chi(n,r))|, \end{aligned} \tag{5.15}$$

that was actually used in simulations. It should be mentioned that the estimator $\widehat{\mathbb{E}}_F(\chi(n,r))$ does not depend on the sample being tested and by increasing M can be arbitrary close to $\mathbb{E}_F(\chi(n,r))$.

The algorithm for computing the TopoTest for one sample can be divided into two steps. Firstly, in the *preparation step* an average ECC for given null distribution F is computed. Then the critical value of the test statistic is estimated empirically by drawing a set of random samples from F and computing the distance between ECCs corresponding to those samples and the average ECC computed previously. Secondly, in the *testing step*, the distance of the ECC of the given sample to the averaged ECC for the considered distribution is computed and compared to the critical values obtained in the first step. This procedure is provided in details by Algorithm 5.1.

Remark 5.2.1. The *preparation step* in Algorithm 5.1 depends only on the sample size n and the null distribution F but is independent of the actual sample X. Hence, it needs to be performed only once if several data samples of size n are considered.

Remark 5.2.2. The threshold value t_α used in the TopoTest is obtained from a numerical Monte Carlo simulation performed for a family of finite samples of a size n and does not explicitly employ asymptotic bounds from Section 5.1.

Remark 5.2.3. The Monte Carlo parameters M and m should be sufficiently large to obtain an accurate resulting test. For the distributions considered in this chapter, values $M = m = 1000$ were selected.

Remark 5.2.4. The need to utilize the Monte Carlo approach to determine threshold value t_α stems from the fact that the distribution of the test statistic (5.2) depends on the distribution of F and the size of the samples for which TopoTest was built. In general, this distribution is unknown. The simulations showed that employing an asymptotic distribution, approximated numerically by using a large sample size n in the preparation step, provided incorrect empirical significance levels in case of samples much smaller than n.

Algorithm 5.1: Algorithm for one-sample testing

Input: Point sample $X \in \mathbb{R}^d$, null distribution F, significance level α, M: number of samples draw from F to estimate average ECC, m: number of samples draw from F to estimate the threshold value.

Output: Rejecting or failure to reject of null hypothesis, p-value

Let $n = |X|$

`/* "Preparation", i.e. determine the threshold` t_α `for rejecting the null hypothesis */`

for $i \leftarrow 1, \ldots, M$ **do**

  $Y_i \leftarrow$ i.i.d. sample of n points from F

  Compute the ECC $\chi(\mathcal{A}(Y_i))$

end

Compute the average ECC $\overline{\chi}(r) \leftarrow \frac{1}{M} \sum_{i=1}^{M} \chi(\mathcal{A}(Y_i)_r)$

for $i \leftarrow 1, \ldots, m$ **do**

  $Y_i' \leftarrow$ i.i.d. sample of n points from F

  Compute the ECC $\chi(\mathcal{A}(Y_i'))$

  Compute the deviation from average $\Delta_i \leftarrow \sup_t \frac{1}{\sqrt{n}} |\chi(\mathcal{A}(Y_i')_r) - \overline{\chi}(r)|$

end

Let $t_\alpha \in \mathbb{R}$ such that $\#\{\Delta_i > t_\alpha\} < \alpha m$

`/* "Testing", i.e. compare the threshold value with sample distance */`

Compute the ECC $\chi(\mathcal{A}(X))$

$\Delta(\chi(\mathcal{A}(X)), \overline{\chi}) \leftarrow \sup_r \frac{1}{\sqrt{n}} |\chi(\mathcal{A}(X)_r) - \overline{\chi}(r)|$

$pv \leftarrow \frac{1}{M} \#\{\Delta_i > \Delta(\chi(\mathcal{A}(X)), \overline{\chi})\}$

return $\Delta(\chi(\mathcal{A}(X)), \overline{\chi}) < t_\alpha, pv$

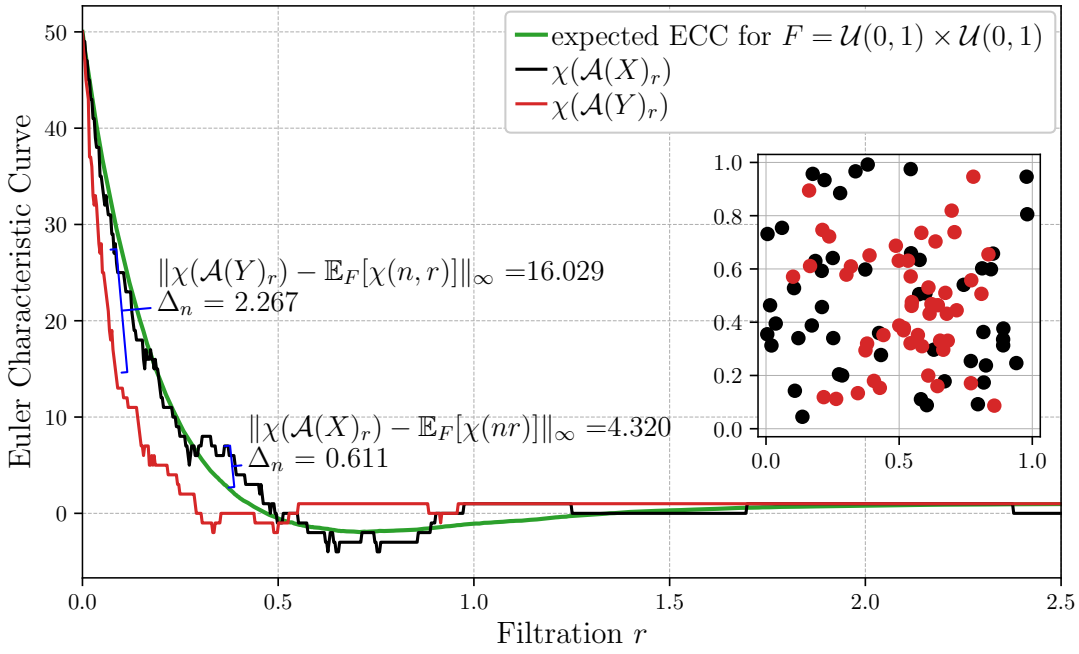


Figure 5.4: Euler characteristic curves of two samples of a size 50; $X \sim \mathcal{U}(0,1) \times \mathcal{U}(0,1)$ (in black) and $Y \sim \beta(3,3) \times \beta(3,3)$ (in red). The green curve represents the expected ECC for $\mathcal{U}(0,1) \times \mathcal{U}(0,1)$. Samples are shown in the inset.

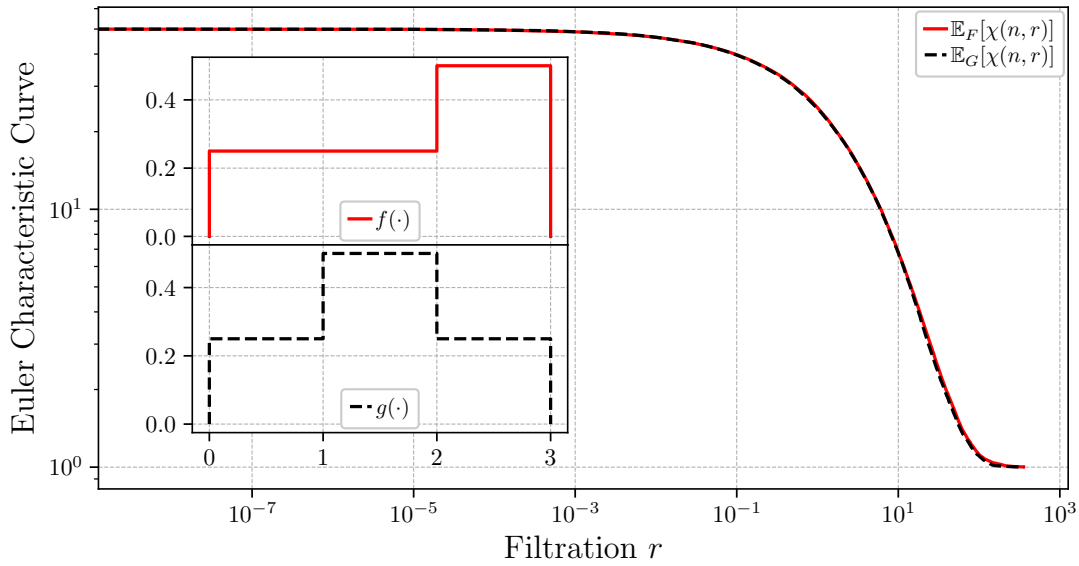


Figure 5.5: The situation of Example 5.2.6: Expected ECCs of distributions F and G with $n = 50$ are drawn with logarithmic scale. The inset shows the corresponding densities f and g.

Example 5.2.5. Consider the samples $X, Y \subseteq [0,1]^2$ consisting of the 50 black and 50 red points as shown in the inset in Figure 5.4. Let us look at the two samples separately, for each of them we perform the one-sample test against the uniform distribution. We want to test, at significance level $\alpha = 0.05$, whether they follow (up to equal excess mass) the uniform distribution. The ECC of X is shown in black and the one of Y in red in Figure 5.4. The green curve represents the expected ECC under the null hypothesis, estimated via $M = 1000$ Monte Carlo iterations using (5.14). We find the test statistic (5.15) computed between the $\chi(\mathcal{A}(X)_r)$ and the average curve is $\widehat{\Delta}_n = 0.611$. Comparing this with the computed threshold of $t_\alpha = 1.318$, we conclude that we do not have evidence to reject the null hypothesis. The p-value is 0.916. In contrast, test statistics computed for $\chi(\mathcal{A}(Y)_r)$ is much larger and equals $\widehat{\Delta}_n = 2.267$. Again using $\alpha = 0.05$, the test provides evidence to reject the null hypothesis with p-value computed to be < 0.001. And indeed, we generated X from the bivariate uniform distribution (i.e. null distribution) whereas Y was sampled from $\beta(3,3) \times \beta(3,3)$, i.e. Cartesian product of two independent univariate $\beta(3,3)$ distributions.

Example 5.2.6. Consider the real-valued distributions F and G with densities

$$f(x) = \frac{1}{4}\mathbb{1}_{(0,2)}(x) + \frac{1}{2}\mathbb{1}_{(2,3)}(x),$$
$$g(x) = \frac{1}{4}\mathbb{1}_{(0,1)}(x) + \frac{1}{2}\mathbb{1}_{(1,2)}(x) + \frac{1}{4}\mathbb{1}_{(2,3)}(x).$$

Observe that for each $t > 0$,

$$\int_{f \geqslant t} f(x)\mathrm{d}x = \int_{g \geqslant t} g(x)\mathrm{d}x = \begin{cases} 1 & \text{if } t \leqslant 1/4, \\ 1/2 & \text{if } 1/4 < t \leqslant 1/2, \\ 0 & \text{if } t > 1/2. \end{cases} \tag{5.16}$$

In other words, they admit the same excess mass. Hence by Lemma 5.1 of [157], the ECCs of F and G in the thermodynamic limit follow the same distribution. The expected ECCs for 50 samples from F and G are shown in Figure 5.5. Note that even though we have a modest sample size, rather far away from the asymptotic regime, the graphs are already almost identical. Therefore, F and G form an example of distributions that are indistinguishable by TopoTest. Indeed, the power of one-sample Kolmogorov-Smirnov test, when F is used as a null distribution and 50 elements samples are drawn from G, is 0.91 and only 0.05, i.e. α, for TopoTest.

5.2.2 Two-Sample Test

In Section 5.1.2 a related approach to the two-sample problem was presented. This idea is formally provided by the Algorithm 5.2 while a particular realization is presented in the example below. Let us

Algorithm 5.2: Two-sample testing

Input: two sample points $X = \{x_1, \ldots, x_m\}, Y = \{y_1, \ldots, y_n\}$ both in $\mathbb{R}^d$, number K of Monte Carlo iterations, significance level α.

Output: Rejecting or failure to reject of null hypothesis, p-value

Compute the distance D between normalized ECCs build on top of X and Y

$$D \leftarrow \sup_r \left| \frac{1}{m}\chi(\mathcal{A}(X)_r) - \frac{1}{n}\chi(\mathcal{A}(Y)_r) \right|$$

Pool the data points $Z \leftarrow X \cup Y$

for $p \leftarrow 1, \ldots, K$ **do**

- $Z^{\#}_{(p)} \leftarrow permute(Z)$
- Split $Z_{(p)}$ into two samples of size m and n
- $X_{(p)} \leftarrow \{Z_{(p),1}, Z_{(p),2}, \ldots, Z_{(p),m}\}$
- $Y_{(p)} \leftarrow \{Z_{(p),m+1}, Z_{(p),m+2}, \ldots, Z_{(p),m+n}\}$
- Compute the distance between ECCs build on top of $X_{(p)}$ and $Y_{(p)}$
- $d_{(p)} \leftarrow \sup_r \left| \frac{1}{m}\chi\left(\mathcal{A}\left(X_{(p)}\right)_r\right) - \frac{1}{n}\chi\left(\mathcal{A}\left(Y_{(p)}\right)_r\right) \right|$

end

$pv \leftarrow \frac{1}{K}\#\{d_{(p)} > D\}$

return $pv < \alpha$, pv

begin with the situation in which the null hypothesis is not rejected.

Example 5.2.7. Consider both X and Y sampled from $\mathcal{U}(0,1)^2$ with $|X| = 30$, $|Y| = 50$, shown in the inset of Figure 5.6. We compute the supremum distance between the normalized ECCs to be

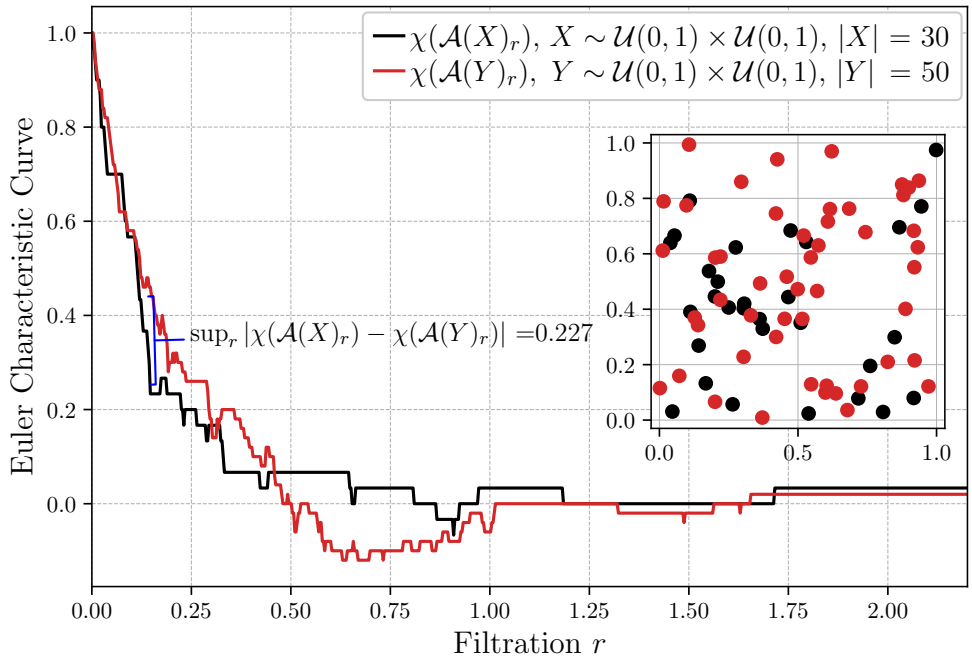


Figure 5.6: Normalized Euler Characteristic Curves of two samples of size 30 and 50 drawn from bivariate uniform distribution, $\mathcal{U}(0,1) \times \mathcal{U}(0,1)$. Samples are shown in the inset.

$D = 0.227$, as illustrated in Figure 5.6. Using $K = 1000$ Monte Carlo iterations we find that a distance between ECCs at least as extreme as D happens roughly 73% of the time. We conclude that we do not have evidence to reject the null hypothesis at significance level $\alpha = 0.05$.

Now let us turn to an example in which the null hypothesis is rejected.

Example 5.2.8. In the Figure 5.7, we have sampled X as 30 points from the bivariate uniform distribution on the unit square $\mathcal{U}(0,1)^2$, whereas Y consists of 50 points sampled from $\beta(3,3) \times \mathcal{U}(0,1)$. We compute the distance between corresponding normalized ECCs to be $D = 0.453$. In $K = 1000$ Monte Carlo iterations, we find that an ECC distance at least as extreme as D never happens, hence using $\alpha = 0.05$ this establishes evidence to reject the null hypothesis.

5.3 Numerical Experiments, One-Sample Problem

In this study, Monte Carlo simulations were used to evaluate the power of TopoTests and compare it with the power of corresponding Kolmogorov-Smirnov tests. In case of univariate distributions, Cramér-von Mises was considered as well for completeness. To obtain more detailed insight into performance of TopoTests, samples of various sizes ranging from $n = 30$ up to $n = 1000$, were examined. In the following subsections three types of experiments are presented:

1. Fixing the null distribution to be standard normal and test samples drawn from a vast variety of alternative distributions with different parameters; Laplace, uniform, t-distribution, as well

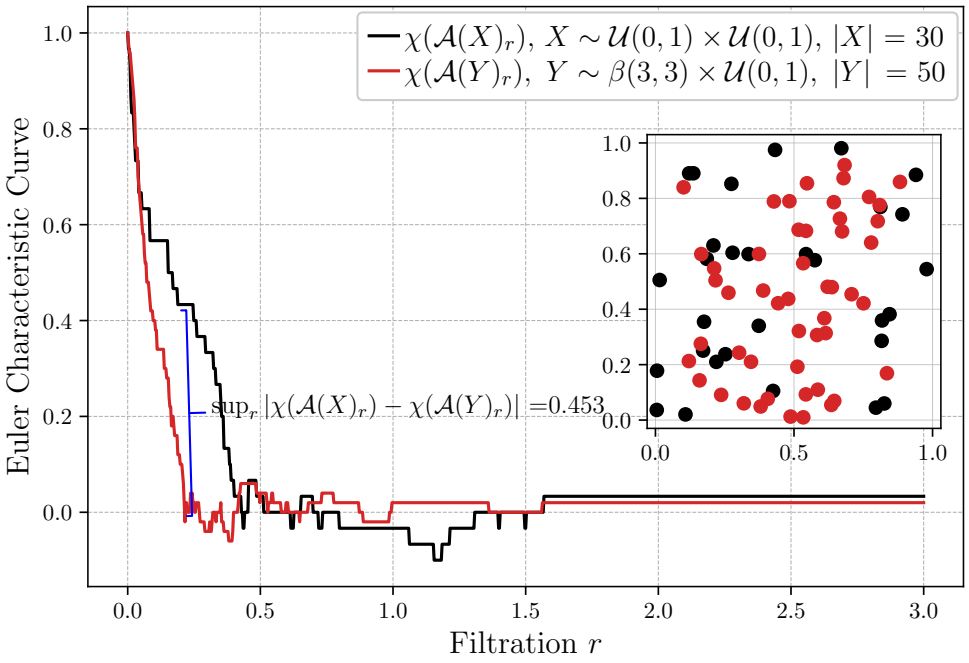


Figure 5.7: Normalized Euler Characteristic Curves of two samples of size 30 and 50 drawn from different distributions: $X \sim \mathcal{U}(0,1) \times \mathcal{U}(0,1)$ and $Y \sim \beta(3,3) \times \mathcal{U}(0,1)$.

as Cauchy, logistic distributions and mixture of Gaussians. This set of experiments allowed to assess how well TopoTests performs to recognize standard normal distributions.

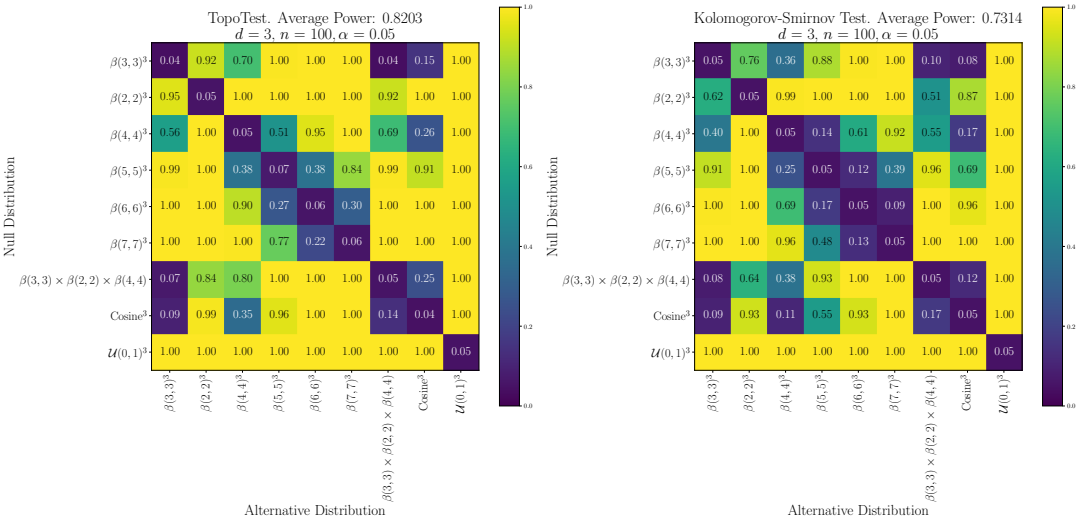
2. Fixing a family of distributions, and treat each of them as null distribution while all others are considered as alternative distribution. For each such a pair of distributions, the empirical power of the test, i.e. 1 minus probability of type II error, was computed using Monte Carlo methods. The result was visualized in a form of heat-maps.

3. In addition, for various dimensions, a relation between power of the test and number n of data points in the sample was examined (cf. Figure 5.13). As expected, the power of the test increases monotonically with the sample size.

In this section both simulations satisfying Assumption 1 and those that do not satisfy it (for instance multivariate normal) were considered. To theoretically underpin this approach, several ideas were suggested in Section 5.1.5. In practice, the fact that the Assumption 1 was not satisfied in some cases did not affect the test powers.

Remark 5.3.1. In this section we benchmark TopoTest by comparing its power with the power of Kolmogorov-Smirnov test, i.e. the probability that the test correctly rejects null hypothesis when the alternative distribution is different than null distribution. Since TopoTests is not able to distinguish different distributions with the same excess mass, which Kolmogorov-Smirnov can distinguish, the setting under which it operates (5.1) is different from the Kolmogorov-Smirnov setting (1.2), and

Figure 5.8: Average power of TopoTest (left panel) and Kolmogorov-Smirnov test for selected trivariate on compact support on $[0,1]^3$. Average power, at significance level $\alpha = 0.05$, is estimated based on $K = 1000$ Monte Carlo realizations for sample size $n = 100$.



hence the reported power of TopoTest might be overestimated. To mediate this effect a vast collection of distributions was considered.

5.3.1 Compactly Supported Distributions

As a first example a collection of distributions supported on three-dimensional unit cube $[0,1]^3$ was considered. The collection consisted of a number of three-fold Cartesian products of independent beta, cosine (rescaled to fit unit interval) and uniform univarite distributions. In such setup the Assumption 1 is fulfilled and developed theory can be applied straightforwardly. In Figure 5.8 the power of TopoTest was compared with power of Kolmogorov-Smirnov test for a collection of trivariate distributions on compact domain. Several sample sizes were considered but here only results obtained for $n = 100$ are reported as similar conclusions can be drawn for different values of n. The TopoTest provided higher power for vast majority of considered pairs of null and alternative distributions resulting in average power, at significance level $\alpha = 0.05$, for this collection of distributions to be 0.82 for TopoTest and 0.73 for Kolmogorov-Smirnov. In fact, for collection of distributions considered in Figure 5.8 in only one, out of 72, comparisons the power of Kolmogorov-Smirnov test was higher than the one for TopoTest, and the difference was slim (0.07 vs. 0.08).

5.3.2 Univariate Unbounded Distributions

In this section we consider a vast collection of univariate unbounded distribution represented on a computer (hence, restricted to a representable range of double precision numbers). The collection include normal distributions $\mathcal{N}(0, \sigma^2)$ with different values of σ, Cauchy, Laplace, Logistic distributions, Student's t-distributions with increasing number of degrees of freedom ν as well as Gaussian mixtures defined as $GM(p, \mu, \sigma) = p\mathcal{N}(0,1) + (1-p)\mathcal{N}(\mu, \sigma)$, for $p \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$,

$\mu \in \{0, 1\}$ and $\sigma \in \{\frac{1}{2}, 1, 2\}$. For completeness some distributions defined on compact support are considered as well.

Table 5.1 provides the empirical power of TopoTests, assessed based on $K = 5000$ Monte Carlo simulations, in distinguishing a standard normal $\mathcal{N}(0, 1)$ from a number of alternative distributions at significance level $\alpha = 0.05$.

As we can observe in Table 5.1, TopoTest outperformed the Kolmogorov-Smirnov test when distinguishing between the standard normal distribution from the normal distribution with variance different from 1, regardless of the sample size. The power of the TopoTest is also greater when the alternative distribution is Student's t-distribution: the difference compared to the Kolmogorov-Smirnov test was particularly pronounced when the number of degrees of freedom ν was small. When ν was 10 or more, the power of both tests is much lower, as expected, but still TopoTest outperformed the Kolmogorov-Smirnov test. Similar conclusion can be drawn for heavier tail alternative distributions such as Cauchy, Laplace or Logistic distribution: the empirical probability of type II error was always lower for TopoTest than for Kolmogorov-Smirnov counterpart. On the other hand, when Gaussian mixtures were considered, it was the Kolmogorov-Smirnov test that performs better, regardless of the value of mixing coefficient p.

5.3.3 Two and Three Dimensional Unbounded Distributions

In Table 5.2 result for collection of bivariate distributions are shown. The $MG(a)$ denotes a multivariate normal distribution with non-diagonal covariance matrix, i.e.

$$MG(a) = \mathcal{N}\left(0, \begin{bmatrix} 1 & a & a & \dots & a \\ a & 1 & a & \dots & a \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a & a & a & \dots & 1 \end{bmatrix}\right), \tag{5.17}$$

where the value of the parameter a varies from 0 to 1 to reflect increasing correlation of components.

Similarly to the univariate case, TopoTests provided lower type II errors in case of alternative distributions being products involving a Student's t-distribution. This conclusion holds also when one of the marginal distribution was a $\mathcal{N}(0, 1)$ and second being Student's t-distribution. A similar result is true for bivariate distributions being a Cartesian product involving Logistic or Laplace distribution. We notice that TopoTest usually provided higher efficiency in case of Gaussian mixtures. On the other hand, TopoTest is significantly weaker than Kolmogorov-Smirnov when considering correlated multivariate normal distributions MG. All of these conclusions can be generalized to three dimensional distributions as initiated by results in Table 5.3.

The last row of Tables 5.1, 5.2 and 5.3 show the average powers of TopoTest and Kolmogorov-Smirnov test for the considered set of alternative distributions. The average power of TopoTest is greater than that of Kolmogorov-Smirnov test for all studied sample sizes.

5.3.4 All-to-All Tests

Results presented in Tables 5.1,5.2,5.3 focused on the ability to discriminate the standard normal distribution from a set of different distributions. However in TopoTest one can choose arbitrary

Table 5.1: Empirical powers of the one-sample TopoTest for different alternative distributions and sample sizes n – the null distribution was standard normal $\mathcal{N}(0, 1)$. Corresponding powers of Kolmogorov-Smirnov tests are given in parenthesis for comparison – higher result is given in bold for easier comparison. Results for the significance level $\alpha = 0.05$. Empirical powers estimated based on $K = 5000$ Monte Carlo simulations.

	Sample size n				
Alternative Distribution	30	50	100	250	500
$\mathcal{N}(0, 0.50)$	**0.953** (0.417)	**0.997** (0.820)	1.000 (1.000)	1.000 (1.000)	1.000 (1.000)
$\mathcal{N}(0, 0.75)$	**0.278** (0.061)	**0.369** (0.097)	**0.705** (0.247)	**0.995** (0.734)	**1.000** (0.998)
$\mathcal{N}(0, 1.25)$	**0.222** (0.096)	**0.291** (0.123)	**0.477** (0.211)	**0.879** (0.459)	**0.998** (0.899)
$\mathcal{N}(0, 1.5)$	**0.519** (0.228)	**0.670** (0.327)	**0.956** (0.688)	**1.000** (0.990)	1.000 (1.000)
Laplace$(0, 1)$	**0.224** (0.055)	**0.309** (0.058)	**0.544** (0.084)	**0.918** (0.145)	**1.000** (0.534)
$\mathcal{U}(-\sqrt{3}, \sqrt{3})$	0.037 (**0.110**)	0.041 (**0.141**)	0.099 (**0.249**)	**0.840** (0.558)	**1.000** (0.930)
$\mathcal{U}(0, 1)$	1.000 (1.000)	1.000 (1.000)	1.000 (1.000)	1.000 (1.000)	1.000 (1.000)
$t(3)$	**0.280** (0.070)	**0.400** (0.066)	**0.674** (0.122)	**0.966** (0.267)	**1.000** (0.700)
$t(5)$	**0.151** (0.054)	**0.169** (0.054)	**0.306** (0.068)	**0.636** (0.080)	**0.918** (0.176)
$t(10)$	**0.084** (0.049)	**0.080** (0.043)	**0.111** (0.051)	**0.246** (0.053)	**0.346** (0.074)
$t(25)$	**0.059** (0.052)	**0.054** (0.041)	**0.066** (0.060)	**0.072** (0.045)	**0.081** (0.053)
Cauchy$(0, 1)$	**0.907** (0.281)	**0.971** (0.456)	**1.000** (0.850)	1.000 (1.000)	1.000 (1.000)
Logistic$(0, 1)$	**0.760** (0.322)	**0.903** (0.511)	**0.996** (0.885)	1.000 (1.000)	1.000 (1.000)
$0.9\mathcal{N}(0, 1) + 0.1\mathcal{N}(0, 0.5)$	**0.065** (0.042)	**0.048** (0.038)	**0.073** (0.072)	**0.090** (0.059)	**0.137** (0.093)
$0.7\mathcal{N}(0, 1) + 0.3\mathcal{N}(0, 0.5)$	**0.124** (0.052)	**0.136** (0.078)	**0.248** (0.136)	**0.542** (0.337)	**0.816** (0.784)
$0.5\mathcal{N}(0, 1) + 0.5\mathcal{N}(0, 0.5)$	**0.292** (0.088)	**0.375** (0.152)	**0.637** (0.404)	**0.978** (0.912)	0.999 (**1.000**)
$0.3\mathcal{N}(0, 1) + 0.7\mathcal{N}(0, 0.5)$	**0.544** (0.159)	**0.746** (0.329)	**0.961** (0.855)	1.000 (1.000)	1.000 (1.000)
$0.1\mathcal{N}(0, 1) + 0.9\mathcal{N}(0, 0.5)$	**0.852** (0.304)	**0.977** (0.672)	**1.000** (0.995)	1.000 (1.000)	1.000 (1.000)
$0.9\mathcal{N}(0, 1) + 0.1\mathcal{N}(0, 2)$	**0.092** (0.052)	**0.077** (0.050)	**0.143** (0.064)	**0.229** (0.056)	**0.413** (0.087)
$0.7\mathcal{N}(0, 1) + 0.3\mathcal{N}(0, 2)$	**0.256** (0.085)	**0.350** (0.098)	**0.627** (0.140)	**0.943** (0.315)	**1.000** (0.778)
$0.5\mathcal{N}(0, 1) + 0.5\mathcal{N}(0, 2)$	**0.514** (0.152)	**0.683** (0.212)	**0.952** (0.449)	**1.000** (0.933)	1.000 (1.000)
$0.3\mathcal{N}(0, 1) + 0.7\mathcal{N}(0, 2)$	**0.733** (0.291)	**0.898** (0.450)	**0.997** (0.858)	**1.000** (0.999)	1.000 (1.000)
$0.1\mathcal{N}(0, 1) + 0.9\mathcal{N}(0, 2)$	**0.875** (0.491)	**0.968** (0.750)	**1.000** (0.984)	1.000 (1.000)	1.000 (1.000)
$0.9\mathcal{N}(0, 1) + 0.1\mathcal{N}(1, 2)$	**0.096** (0.068)	**0.111** (0.063)	**0.171** (0.092)	**0.319** (0.135)	**0.548** (0.280)
$0.7\mathcal{N}(0, 1) + 0.3\mathcal{N}(1, 2)$	**0.318** (0.182)	**0.464** (0.249)	**0.747** (0.508)	**0.985** (0.932)	1.000 (1.000)
$0.5\mathcal{N}(0, 1) + 0.5\mathcal{N}(1, 2)$	**0.588** (0.453)	**0.760** (0.665)	**0.971** (0.948)	1.000 (1.000)	1.000 (1.000)
$0.3\mathcal{N}(0, 1) + 0.7\mathcal{N}(1, 2)$	**0.778** (0.747)	0.927 (**0.930**)	0.999 (0.999)	1.000 (1.000)	1.000 (1.000)
$0.1\mathcal{N}(0, 1) + 0.9\mathcal{N}(1, 2)$	0.889 (**0.921**)	0.987 (**0.990**)	1.000 (1.000)	1.000 (1.000)	1.000 (1.000)
Average Power	**0.446** (0.246)	**0.527** (0.338)	**0.659** (0.501)	**0.808** (0.643)	**0.866** (0.764)

Table 5.2: The same as Table 5.1 but for two dimensional distributions. Null distribution is $N_0 \sim \mathcal{N}(0, I_2)$, where I_2 is a 2×2 identity matrix. Empirical powers, based on $K = 1000$ Monte Carlo simulations. Alternative distributions include Gaussian mixtures of N_0, $N_1 \sim \mathcal{N}((1,1), 3I_2)$, $N_2 \sim \mathcal{N}((0,0), 3I_2)$ and $N_3 \sim \mathcal{N}((-1,-1), 3I_2)$.

	Sample size n				
Alternative Distribution	30	50	100	250	500
$MG(0.05)$	0.036 (**0.052**)	0.050 (0.050)	**0.049** (0.038)	0.061 (**0.070**)	**0.059** (0.048)
$MG(0.1)$	0.042 (**0.044**)	0.041 (**0.056**)	**0.048** (0.042)	0.052 (**0.074**)	0.065 (**0.096**)
$MG(0.2)$	0.040 (**0.073**)	0.064 (**0.114**)	0.060 (**0.106**)	0.062 (**0.170**)	0.062 (**0.298**)
$MG(0.3)$	0.046 (**0.072**)	0.064 (**0.130**)	0.071 (**0.134**)	0.090 (**0.368**)	0.121 (**0.702**)
$MG(0.5)$	0.093 (**0.124**)	0.115 (**0.258**)	0.200 (**0.478**)	0.369 (**0.952**)	0.652 (**1.000**)
$MG(0.7)$	**0.232** (0.229)	0.381 (**0.578**)	0.688 (**0.902**)	0.966 (**1.000**)	1.000 (1.000)
$\mathcal{U}(-\sqrt{3}, \sqrt{3}) \times \mathcal{U}(-\sqrt{3}, \sqrt{3})$	0.044 (**0.157**)	0.082 (**0.292**)	**0.487** (0.468)	**1.000** (0.942)	1.000 (1.000)
$\mathcal{U}(0,1) \times \mathcal{U}(0,1)$	1.000 (1.000)	1.000 (1.000)	1.000 (1.000)	1.000 (1.000)	1.000 (1.000)
$t(3) \times t(3)$	**0.399** (0.121)	**0.673** (0.176)	**0.956** (0.308)	**1.000** (0.838)	**1.000** (0.996)
$t(5) \times t(5)$	**0.152** (0.073)	**0.305** (0.104)	**0.609** (0.124)	**0.960** (0.304)	**0.999** (0.660)
$t(10) \times t(10)$	0.045 (**0.064**)	**0.094** (0.088)	**0.191** (0.078)	**0.470** (0.094)	**0.782** (0.100)
$t(25) \times t(25)$	0.039 (**0.047**)	0.066 (**0.068**)	**0.067** (0.044)	**0.096** (0.052)	**0.165** (0.058)
$\mathcal{N}(0,1) \times t(3)$	**0.096** (0.064)	**0.235** (0.086)	**0.466** (0.102)	**0.882** (0.244)	**0.993** (0.422)
$\mathcal{N}(0,1) \times t(5)$	0.059 (**0.062**)	**0.086** (0.068)	**0.196** (0.086)	**0.472** (0.122)	**0.787** (0.116)
$\mathcal{N}(0,1) \times t(10)$	0.041 (**0.043**)	0.052 (**0.060**)	**0.068** (0.060)	**0.141** (0.066)	**0.270** (0.072)
$0.9N_0 + 0.1N_1$	0.051 (**0.074**)	0.092 (**0.096**)	**0.184** (0.102)	**0.448** (0.238)	**0.701** (0.406)
$0.7N_0 + 0.3N_1$	**0.284** (0.257)	**0.519** (0.452)	**0.842** (0.782)	**0.998** (0.996)	1.000 (1.000)
$0.5N_0 + 0.5N_1$	0.600 (**0.637**)	**0.908** (0.902)	0.998 (0.998)	1.000 (1.000)	1.000 (1.000)
$0.3N_0 + 0.7N_1$	0.843 (**0.917**)	0.982 (**0.988**)	1.000 (1.000)	1.000 (1.000)	1.000 (1.000)
$0.1N_0 + 0.9N_1$	0.943 (**0.995**)	0.998 (**1.000**)	1.000 (1.000)	1.000 (1.000)	1.000 (1.000)
$0.9N_0 + 0.1N_2$	0.050 (**0.064**)	0.064 (**0.074**)	**0.128** (0.052)	**0.281** (0.080)	**0.511** (0.114)
$0.7N_0 + 0.3N_2$	**0.185** (0.110)	**0.369** (0.170)	**0.679** (0.236)	**0.982** (0.596)	**1.000** (0.900)
$0.5N_0 + 0.5N_2$	**0.487** (0.237)	**0.777** (0.422)	**0.982** (0.678)	**1.000** (0.984)	1.000 (1.000)
$0.3N_0 + 0.7N_2$	**0.746** (0.433)	**0.956** (0.702)	**0.999** (0.956)	1.000 (1.000)	1.000 (1.000)
$0.1N_0 + 0.9N_2$	**0.902** (0.665)	**0.996** (0.930)	1.000 (1.000)	1.000 (1.000)	1.000 (1.000)
$0.9N_0 + 0.05N_1 + 0.05N_3$	0.055 (**0.059**)	**0.080** (0.078)	**0.207** (0.080)	**0.453** (0.128)	**0.750** (0.178)
$0.7N_0 + 0.15N_1 + 0.15N_3$	**0.308** (0.137)	**0.566** (0.232)	**0.879** (0.384)	**0.998** (0.878)	**1.000** (0.996)
$0.5N_0 + 0.25N_1 + 0.25N_3$	**0.679** (0.371)	**0.918** (0.634)	**0.998** (0.858)	1.000 (1.000)	1.000 (1.000)
Average Power	**0.303** (0.256)	**0.412** (0.350)	**0.538** (0.432)	**0.671** (0.578)	**0.747** (0.649)

Table 5.3: The same as Table 5.1 but for three dimensional distributions. Null distribution is $N_0 \sim \mathcal{N}(0, I_3)$, where I_3 is a 3×3 identity matrix. Empirical powers, based on $K = 250$ Monte Carlo simulations. Alternative distributions include Gaussian mixtures of $N_0, N_1 \sim \mathcal{N}((1,1,1), 3I_3)$.

	Sample size n				
Alternative Distribution	30	50	100	250	500
$MG(0.05)$	**0.052** (0.028)	0.048 (**0.052**)	0.064 (**0.068**)	**0.062** (0.056)	**0.056** (0.044)
$MG(0.1)$	**0.056** (0.052)	0.062 (**0.112**)	**0.076** (0.068)	0.038 (**0.104**)	0.054 (**0.104**)
$MG(0.2)$	**0.084** (0.076)	0.062 (**0.120**)	0.086 (**0.128**)	0.074 (**0.328**)	0.084 (**0.592**)
$MG(0.3)$	0.084 (**0.104**)	0.080 (**0.216**)	0.134 (**0.252**)	0.168 (**0.776**)	0.276 (**0.992**)
$MG(0.5)$	0.204 (**0.212**)	0.252 (**0.576**)	0.524 (**0.852**)	0.854 (**1.000**)	0.994 (**1.000**)
$\mathcal{U}(-\sqrt{3}, \sqrt{3})^3$	0.048 (**0.176**)	0.156 (**0.408**)	**0.632** (0.568)	**1.000** (0.968)	1.000 (1.000)
$\mathcal{U}(0,1)^3$	1.000 (1.000)	1.000 (1.000)	1.000 (1.000)	1.000 (1.000)	1.000 (1.000)
$t(3)^3$	**0.624** (0.168)	**0.836** (0.388)	**0.998** (0.524)	**1.000** (0.996)	1.000 (1.000)
$t(5)^3$	**0.268** (0.056)	**0.402** (0.196)	**0.806** (0.240)	**0.992** (0.560)	**1.000** (0.936)
$t(10)^3$	0.048 (**0.064**)	0.108 (0.108)	**0.266** (0.080)	**0.624** (0.176)	**0.906** (0.276)
$Logistic(0,1)^3$	**0.988** (0.904)	**1.000** (0.996)	1.000 (1.000)	1.000 (1.000)	1.000 (1.000)
$Laplace(0,1)^3$	**0.496** (0.116)	**0.774** (0.220)	**0.990** (0.332)	**1.000** (0.924)	1.000 (1.000)
$N_0 \times t(5) \times t(5)$	**0.140** (0.052)	**0.238** (0.128)	**0.520** (0.120)	**0.824** (0.224)	**0.996** (0.480)
$N_0 \times N_0 \times t(5)$	**0.056** (0.028)	**0.082** (0.076)	**0.154** (0.064)	**0.304** (0.080)	**0.586** (0.116)
$0.9N_0 + 0.1N_1$	**0.100** (0.052)	0.110 (**0.132**)	**0.228** (0.116)	**0.502** (0.304)	**0.772** (0.500)
$0.5N_0 + 0.5N_1$	**0.792** (0.748)	**0.954** (0.944)	1.000 (1.000)	1.000 (1.000)	1.000 (1.000)
$0.1N_0 + 0.9N_1$	0.996 (**1.000**)	1.000 (1.000)	1.000 (1.000)	1.000 (1.000)	1.000 (1.000)
Average Power	**0.355** (0.284)	**0.421** (0.392)	**0.558** (0.436)	**0.673** (0.617)	**0.748** (0.708)

continuous distributions as null and alternative. Hence below we present power matrices where all possible pairs of null and alternative distributions formed from the previous set were considered – results are presented in Figures 5.9,5.10,5.11. For easier evaluation of the effectiveness of the TopoTest in comparison to Kolmogorov-Smirnov, the difference in power was shown in the figures. Hence, the blue region corresponds to combinations of null and alternative distribution for which the TopoTest yielded higher power while red regions reflect the combinations for which TopoTest was outperformed by Kolmogorov-Smirnov. White color stands for combinations for which both tests performed similar.

The analysis was conducted also dimension $d = 5$ as can be seen in Figure 5.12. For $d > 3$ the Kolmogorov-Smirnov test was not preformed due to too long computation time, hence results for TopoTest are presented only as this method provided feasible computational complexity.

As can be seen the TopoTest stayed sensitive enough to differentiate between multivariate normal distribution and Cartesian products of involving Student's t-distribution and standard normal as marginals, especially given that considered samples sizes are low for such high dimensional spaces.

The heatmap presented in Figure 5.10 reveals several prominent red-blocks, i.e. combinations of null and alternative distributions for which the power of the TopoTest is significantly lower than the power of KS test: e.g. the combination $G = p\mathcal{N}(0,1) + (1-p)\mathcal{N}(0,2)$ and $F = p\mathcal{N}(0,1) + (1-p)\mathcal{N}(\mu, 2)$, $\mu = 1$. This observation is related to the Lemma 5.1 by Vishwanath *et al.* [157] (c.f. Example 5.2.6) regarding equivalence in expected ECCs. Although the distributions F and G are not Euler equivalent and the condition (5.16) is not met but only approximately, the expected ECCs are quite similar for small values of μ making them hard to distinguish by the TopoTest test statistic (5.3). Similar situations holds for trivariate distributions as shown in Figure 5.11.

Figure 5.9: Comparison of the power of TopoTest and Kolmogorov-Smirnov one-sample tests in case of univariate probability distributions. In each matrix element a difference between power of TopoTest and Kolmogorov-Smirnov test was given. The difference in power was estimated based on $K = 1000$ Monte Carlo realizations. Left and right panels shows tests powers for sample sizes $n = 100$ and $n = 250$, respectively. The average power (excluding diagonal elements) of TopoTest is 0.722 (0.832) and 0.634 (0.794) for Kolmogorov-Smirnov for $n = 100$ ($n = 250$).

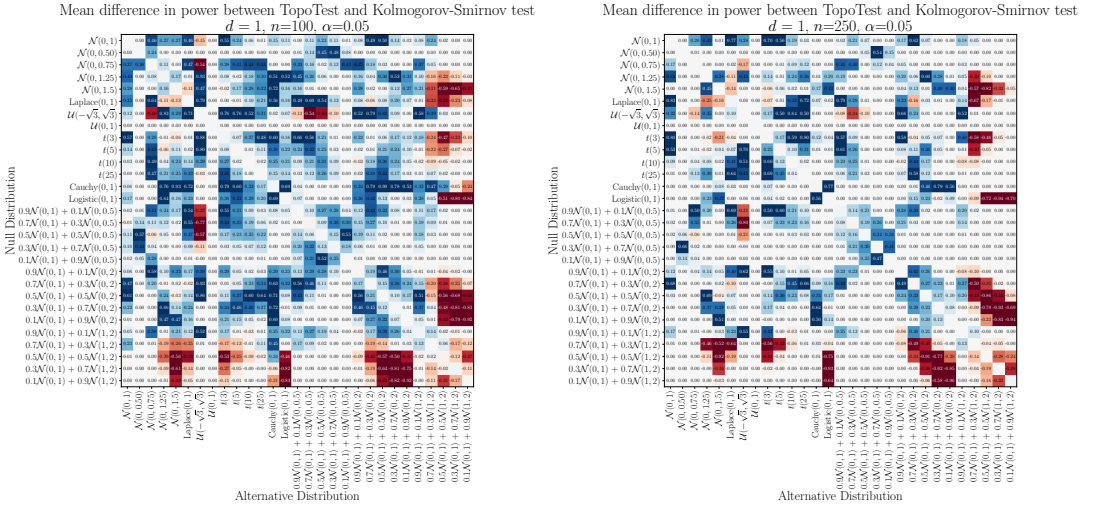


Figure 5.10: The same as Figure 5.9 but for bivariate distributions. Results based on $K = 1000$ Monte Carlo realizations. Average power is 0.642 (0.772) for TopoTest and 0.560 (0.720) for Kolmogorov-Smirnov for $n = 100$ ($n = 250$).

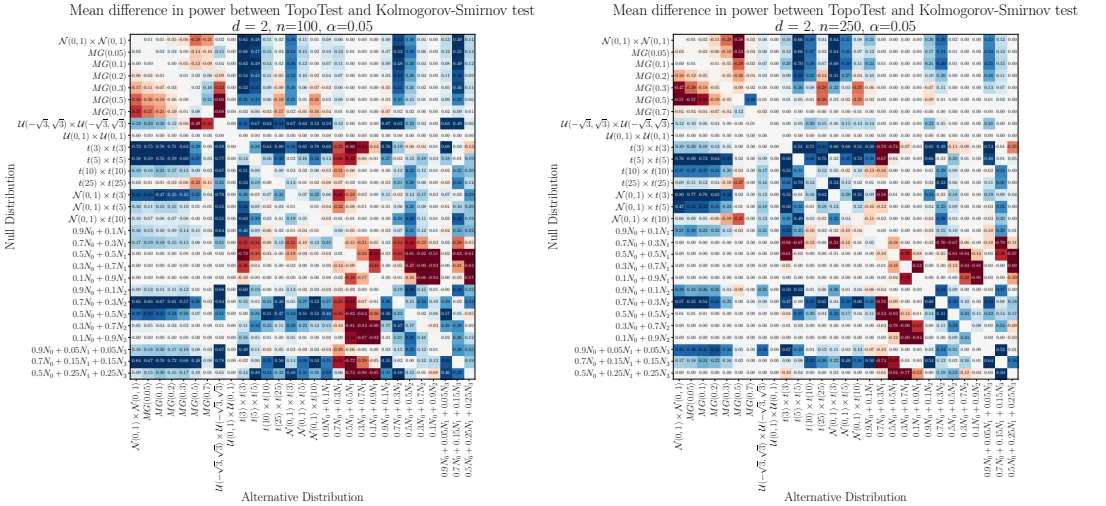


Figure 5.11: The same as Figure 5.9 but for three-dimensional distributions. Results based on $K = 250$ Monte Carlo realizations. Average power is 0.708 (0.824) for TopoTest and 0.602 (0.763) for Kolmogorov-Smirnov for $n = 100$ ($n = 250$).

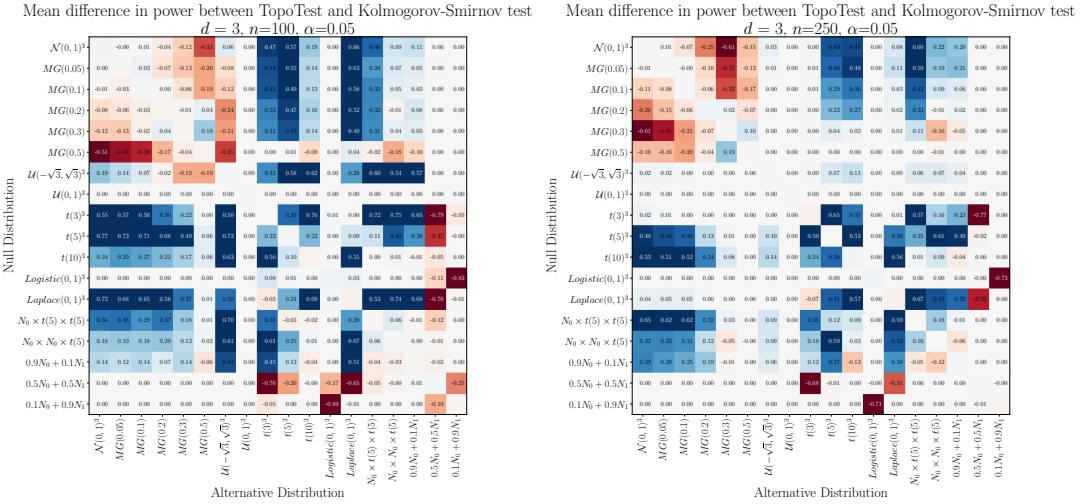


Figure 5.12: Average power of TopoTest for five dimension distributions, for sample sizes $n = 250$ and $n = 500$. Results based on $K = 1000$ Monte Carlo realizations.

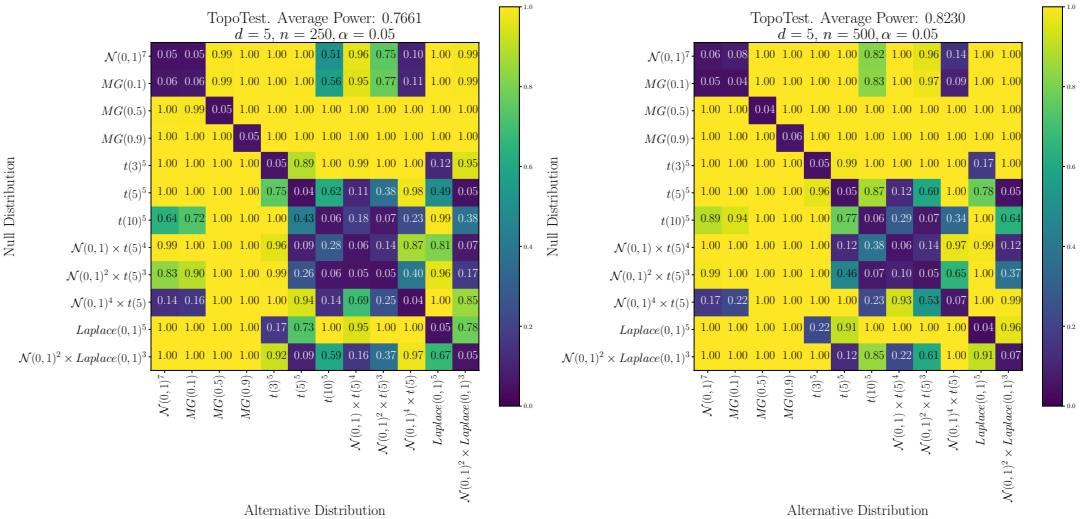
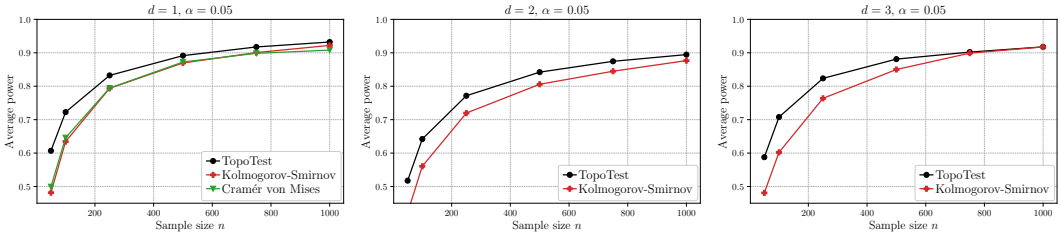


Figure 5.13: Average power of the TopoTest (black curve) and Kolmogorov-Smirnov (red curve) as a function of sample size n for dimensions $d = 1, 2, 3$. In case of $d = 1$ the average power of Cramér-von Mises (green curve) test was shown as well. To guide an eye the data points are connect by lines.



5.3.5 Dependence of the Test Power on Sample Size

The dependence of the power of TopoTest and Kolmogorov-Smirnov tests on the sample size n is shown in Figure 5.13 for random samples in dimensions $d = 1, 2, 3$. To compute average power, all combinations of null and alternative distributions, as considered in Figures 5.9, 5.10 and 5.11, were taken into account, except alternative being the same as null distribution. In all cases, the average power increased with sample size as expected. In case of univariate distribution (leftmost panel in Figure 5.13) the results obtained using Cramér-von Mises test were added for completeness. The overall performance of this test is similar to Kolmogorov-Smirnov, hence detailed analysis was omitted. The TopoTest however provides higher average power for all sample sizes regardless of the data dimension. It should be noted that powers presented in Figure 5.13 should not be directly compared across different dimensions as the actual value depends on the list of considered distributions which is different for each dimension.

5.4 Numerical Experiments, Two-Sample Problem

A numerical study was conducted also for two-sample problems, in which Algorithm 5.2 was applied. The two-sample problem was considered for completeness purpose as practical application is limited by high computational costs, therefore results presented here are restricted to comparison of empirical power of two-sample TopoTest and Kolmogorov-Smirnov tests in $d = 1$ (cf. Table 5.4) and $d = 2$ (cf. Table 5.5). Simulations showed that in both cases the TopoTest outperformed the Kolmogorov-Smirnov test: in the vast majority of examined cases the power of the former is greater. Moreover, the average power for TopoTest is greater than the corresponding average power of Kolmogorov-Smirnov test for all sample sizes n.

As in the Section 5.3, the above collection of distribution is examined also in all-to-all settings. The difference in average power between TopoTest and Kolmogorov-Smirnov tests are shown are shown Figure 5.14.

5.5 Real Data Analysis

In this section, we show two exemplary applications of the developed method to the analysis of real data.

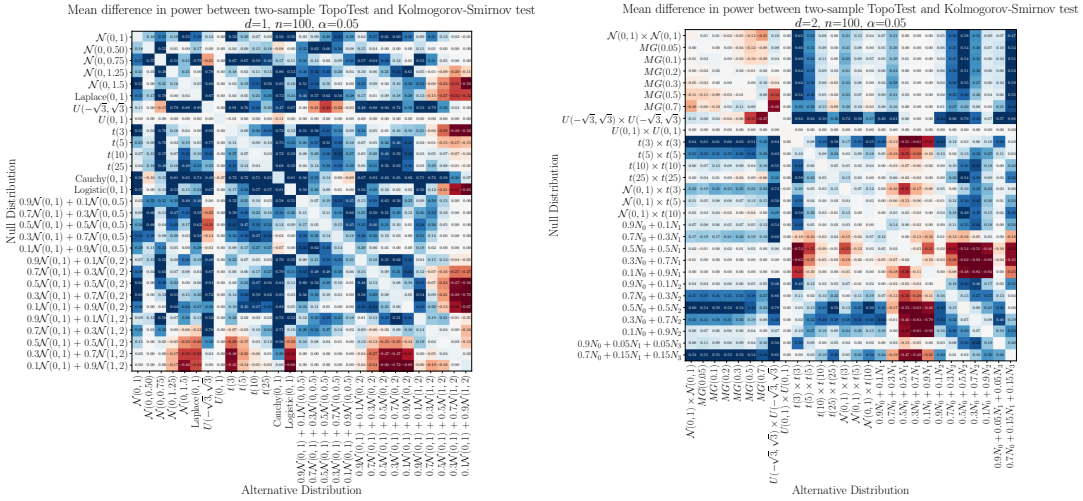
Table 5.4: Empirical powers of the two-sample TopoTest for different alternative distributions and sample sizes n – the null distribution is standard normal $\mathcal{N}(0,1)$. Corresponding powers of Kolmogorov-Smirnov tests are given in parenthesis for comparison – higher result is given in bold for easier comparison. Results for the significance level $\alpha = 0.05$ Empirical powers estimated based on $K = 500$ Monte Carlo realizations.

	Sample size n				
Second Sample Distribution	30	50	100	250	500
$\mathcal{N}(0, 0.50)$	**0.694** (0.218)	**0.890** (0.358)	**0.996** (0.816)	1.000 (1.000)	1.000 (1.000)
$\mathcal{N}(0, 0.75)$	**0.202** (0.054)	**0.290** (0.070)	**0.462** (0.114)	**0.858** (0.376)	**0.938** (0.790)
$\mathcal{N}(0, 1.25)$	**0.188** (0.056)	**0.166** (0.040)	**0.300** (0.110)	**0.682** (0.228)	**0.822** (0.474)
$\mathcal{N}(0, 1.5)$	**0.366** (0.084)	**0.468** (0.124)	**0.792** (0.240)	**0.984** (0.782)	0.984 (**0.994**)
Laplace$(0, 1)$	**0.154** (0.036)	**0.204** (0.046)	**0.458** (0.068)	**0.892** (0.076)	**0.992** (0.154)
$\mathcal{U}(-\sqrt{3}, \sqrt{3})$	**0.092** (0.042)	**0.094** (0.054)	**0.204** (0.082)	**0.756** (0.274)	**0.998** (0.592)
$\mathcal{U}(0, 1)$	**1.000** (0.970)	1.000 (1.000)	1.000 (1.000)	1.000 (1.000)	1.000 (1.000)
$t(3)$	**0.230** (0.024)	**0.276** (0.058)	**0.564** (0.046)	**0.930** (0.084)	**0.956** (0.220)
$t(5)$	**0.116** (0.038)	**0.124** (0.030)	**0.238** (0.036)	**0.568** (0.036)	**0.844** (0.072)
$t(10)$	**0.088** (0.048)	**0.082** (0.030)	**0.098** (0.028)	**0.204** (0.062)	**0.370** (0.052)
$t(25)$	**0.102** (0.036)	**0.062** (0.028)	**0.064** (0.040)	**0.094** (0.040)	**0.110** (0.046)
Cauchy$(0, 1)$	**0.784** (0.060)	**0.894** (0.118)	**0.914** (0.350)	0.906 (**0.956**)	0.916 (**1.000**)
Logistic$(0, 1)$	**0.494** (0.096)	**0.712** (0.164)	**0.948** (0.392)	**0.994** (0.942)	0.998 (**1.000**)
$0.9\mathcal{N}(0, 1) + 0.1\mathcal{N}(0, 0.5)$	**0.072** (0.036)	**0.092** (0.038)	**0.076** (0.048)	**0.104** (0.078)	0.082 (**0.086**)
$0.7\mathcal{N}(0, 1) + 0.3\mathcal{N}(0, 0.5)$	**0.124** (0.048)	**0.122** (0.068)	**0.188** (0.098)	**0.278** (0.206)	0.266 (**0.430**)
$0.5\mathcal{N}(0, 1) + 0.5\mathcal{N}(0, 0.5)$	**0.190** (0.072)	**0.242** (0.096)	**0.456** (0.178)	**0.638** (0.550)	0.610 (**0.938**)
$0.3\mathcal{N}(0, 1) + 0.7\mathcal{N}(0, 0.5)$	**0.334** (0.088)	**0.490** (0.176)	**0.810** (0.380)	**0.950** (0.922)	0.822 (**1.000**)
$0.1\mathcal{N}(0, 1) + 0.9\mathcal{N}(0, 0.5)$	**0.568** (0.172)	**0.782** (0.282)	**0.954** (0.674)	0.992 (**0.998**)	0.958 (**1.000**)
$0.9\mathcal{N}(0, 1) + 0.1\mathcal{N}(0, 2)$	**0.114** (0.040)	**0.102** (0.038)	**0.090** (0.048)	**0.220** (0.044)	**0.402** (0.076)
$0.7\mathcal{N}(0, 1) + 0.3\mathcal{N}(0, 2)$	**0.184** (0.030)	**0.272** (0.048)	**0.424** (0.058)	**0.814** (0.146)	**0.980** (0.338)
$0.5\mathcal{N}(0, 1) + 0.5\mathcal{N}(0, 2)$	**0.284** (0.038)	**0.502** (0.084)	**0.758** (0.152)	**0.992** (0.476)	**0.996** (0.934)
$0.3\mathcal{N}(0, 1) + 0.7\mathcal{N}(0, 2)$	**0.458** (0.100)	**0.722** (0.126)	**0.944** (0.344)	**1.000** (0.906)	1.000 (1.000)
$0.1\mathcal{N}(0, 1) + 0.9\mathcal{N}(0, 2)$	**0.604** (0.118)	**0.822** (0.276)	**0.988** (0.630)	0.998 (**1.000**)	0.996 (**1.000**)
$0.9\mathcal{N}(0, 1) + 0.1\mathcal{N}(1, 2)$	**0.086** (0.050)	**0.120** (0.042)	**0.128** (0.042)	**0.286** (0.074)	**0.548** (0.134)
$0.7\mathcal{N}(0, 1) + 0.3\mathcal{N}(1, 2)$	**0.210** (0.064)	**0.280** (0.108)	**0.540** (0.190)	**0.906** (0.630)	**0.974** (0.958)
$0.5\mathcal{N}(0, 1) + 0.5\mathcal{N}(1, 2)$	**0.354** (0.174)	**0.552** (0.330)	**0.814** (0.692)	**1.000** (0.990)	0.996 (**1.000**)
$0.3\mathcal{N}(0, 1) + 0.7\mathcal{N}(1, 2)$	**0.556** (0.380)	**0.744** (0.684)	**0.972** (0.952)	1.000 (1.000)	0.998 (**1.000**)
$0.1\mathcal{N}(0, 1) + 0.9\mathcal{N}(1, 2)$	**0.688** (0.616)	0.888 (**0.892**)	0.990 (**1.000**)	0.998 (**1.000**)	1.000 (1.000)
Average Power	**0.333** (0.135)	**0.428** (0.193)	**0.577** (0.315)	**0.752** (0.531)	**0.806** (0.653)

Table 5.5: The same as Table 5.4 but for $d = 2$. Standard bivariate normal is used as a null distribution. The MG distribution is defined in (5.17).

	Sample size n				
Second Sample Distribution	30	50	100	250	500
$MG(0.05)$	**0.084** (0.058)	0.052 (**0.066**)	**0.080** (0.066)	**0.066** (0.058)	0.058 (**0.086**)
$MG(0.1)$	0.060 (**0.072**)	**0.074** (0.064)	**0.078** (0.066)	0.036 (**0.076**)	0.060 (**0.092**)
$MG(0.2)$	**0.078** (0.062)	**0.080** (0.074)	0.052 (**0.074**)	0.060 (**0.124**)	0.074 (**0.196**)
$MG(0.3)$	**0.082** (0.062)	0.054 (**0.066**)	0.064 (**0.114**)	0.080 (**0.236**)	0.100 (**0.472**)
$MG(0.5)$	0.086 (**0.092**)	0.100 (**0.136**)	0.136 (**0.264**)	0.236 (**0.666**)	0.368 (**0.976**)
$MG(0.7)$	**0.142** (0.132)	0.226 (**0.254**)	0.374 (**0.582**)	0.764 (**0.986**)	0.958 (**1.000**)
$\mathcal{U}(-\sqrt{3},\sqrt{3}) \times \mathcal{U}(-\sqrt{3},\sqrt{3})$	**0.096** (0.090)	0.144 (**0.156**)	**0.346** (0.244)	**0.944** (0.584)	**1.000** (0.930)
$\mathcal{U}(0,1) \times \mathcal{U}(0,1)$	1.000 (1.000)	1.000 (1.000)	1.000 (1.000)	1.000 (1.000)	1.000 (1.000)
$t(3) \times t(3)$	**0.328** (0.042)	**0.494** (0.068)	**0.792** (0.140)	**0.990** (0.412)	**0.980** (0.868)
$t(5) \times t(5)$	**0.196** (0.050)	**0.224** (0.066)	**0.412** (0.086)	**0.806** (0.150)	**0.982** (0.298)
$t(10) \times t(10)$	**0.120** (0.054)	**0.110** (0.050)	**0.144** (0.064)	**0.274** (0.066)	**0.546** (0.108)
$t(25) \times t(25)$	**0.068** (0.040)	**0.076** (0.054)	**0.064** (0.058)	**0.080** (0.078)	**0.130** (0.052)
$\mathcal{N}(0,1) \times t(3)$	**0.156** (0.052)	**0.160** (0.064)	**0.288** (0.076)	**0.598** (0.128)	**0.866** (0.190)
$\mathcal{N}(0,1) \times t(5)$	**0.070** (0.052)	**0.112** (0.064)	**0.180** (0.052)	**0.304** (0.056)	**0.518** (0.118)
$\mathcal{N}(0,1) \times t(10)$	**0.082** (0.034)	0.054 (**0.062**)	**0.090** (0.054)	**0.088** (0.062)	**0.152** (0.086)
$0.9N_0 + 0.1N_1$	**0.098** (0.052)	**0.102** (0.080)	**0.136** (0.068)	**0.296** (0.154)	**0.454** (0.204)
$0.7N_0 + 0.3N_1$	**0.280** (0.120)	**0.376** (0.178)	**0.628** (0.414)	**0.956** (0.900)	0.998 (0.998)
$0.5N_0 + 0.5N_1$	**0.508** (0.292)	**0.694** (0.588)	0.914 (**0.922**)	1.000 (1.000)	1.000 (1.000)
$0.3N_0 + 0.7N_1$	**0.712** (0.636)	0.892 (**0.900**)	1.000 (1.000)	1.000 (1.000)	1.000 (1.000)
$0.1N_0 + 0.9N_1$	0.820 (**0.888**)	0.972 (**0.996**)	1.000 (1.000)	1.000 (1.000)	1.000 (1.000)
$0.9N_0 + 0.1N_2$	**0.108** (0.058)	**0.096** (0.076)	**0.118** (0.064)	**0.190** (0.066)	**0.332** (0.086)
$0.7N_0 + 0.3N_2$	**0.224** (0.064)	**0.250** (0.090)	**0.496** (0.106)	**0.840** (0.278)	**0.994** (0.658)
$0.5N_0 + 0.5N_2$	**0.352** (0.080)	**0.582** (0.140)	**0.858** (0.282)	**1.000** (0.806)	**1.000** (0.996)
$0.3N_0 + 0.7N_2$	**0.636** (0.140)	**0.810** (0.318)	**0.970** (0.664)	**1.000** (0.998)	1.000 (1.000)
$0.1N_0 + 0.9N_2$	**0.798** (0.246)	**0.918** (0.574)	**1.000** (0.914)	1.000 (1.000)	1.000 (1.000)
$0.9N_0 + 0.05N_1 + 0.05N_3$	**0.088** (0.050)	**0.094** (0.076)	**0.142** (0.076)	**0.304** (0.102)	**0.476** (0.130)
$0.7N_0 + 0.15N_1 + 0.15N_3$	**0.234** (0.084)	**0.400** (0.108)	**0.662** (0.188)	**0.968** (0.502)	**1.000** (0.904)
$0.5N_0 + 0.25N_1 + 0.25N_3$	**0.588** (0.128)	**0.786** (0.242)	**0.966** (0.554)	**1.000** (0.990)	1.000 (1.000)
Average Power	**0.289** (0.169)	**0.355** (0.236)	**0.464** (0.328)	**0.603** (0.481)	**0.680** (0.587)

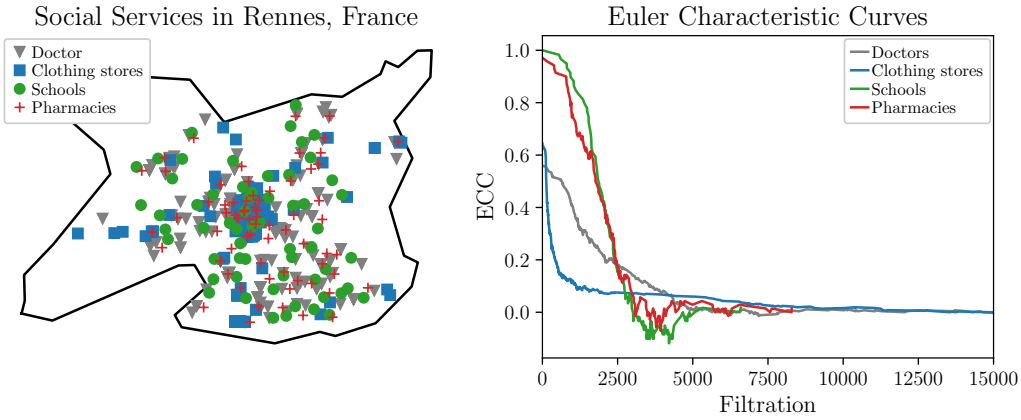
Figure 5.14: Difference in average power of two-sample TopoTest and two-sample Kolmogorov-Smirnov tests for univariate (left panel) and bivariate (right panel) distributions. In both cases sample sizes were $n = 100$ and $K = 500$ Monte Carlo realizations were performed to estimate the average power. Average power of TopoTest is 0.643 (0.537) while for Kolmogorov-Smirnov it is 0.453 (0.437) in $d = 1$ ($d = 2$).



First, we consider Fisher's *Iris data* [6, 76] in the one-sample setting. This data includes three multivariate samples corresponding to three different species of Iris, i.e. *Iris setosa*, *Iris virginica*, and *Iris versicolor*. There are 50 samples from each species, containing four measurements of the flower. We would like to determine if the distribution of each species follows a four-dimensional normal distribution. This can be formulated as a one-sample problem, where G is the distribution of a sample, and F is the specified four-dimensional normal distribution. F involves an unknown mean vector μ and unknown covariance matrix Σ. For each species, μ and Σ are estimated by sample mean and sample covariance matrix. Our one-sample test for testing $H_0 : G = F$ against $H_1 : G \neq F$ gave p-values of 0.057, 0.569 and 0.999 for *Iris setosa*, *Iris virginica* and *Iris versicolor*, respectively. These p-values indicate that, at significance level 0.05, H_0 should not be rejected for each of the Iris species. However, when the same procedure is applied to the entire Iris dataset (i.e. without splitting into species), the p-value is $< 10^{-4}$, hence the null hypothesis is to be rejected, which indicates that multivariate normal distribution does not fit whole Iris dataset. The conclusions are consistent with the literature [59].

In our second example, we consider a dataset introduced in [78] consisting of a collection of geographic locations of four distinct social services, i.e. doctor offices, clothing stores, schools, and pharmacies, in the municipality of Rennes, France. It is visualized in Figure 5.15 as a map. The two-sample TopoTest is used to detect if there are any significant differences in the distribution of those facilities. The test was conducted for all possible pairs. The p-values for all tests involving the distribution of clothing stores were below 10^{-4}, meaning that in the Algorithm 5.2 in all of $K = 10000$ iterations $d_{(p)} < D$, which indicates that their geographic distribution is significantly different from the distribution of doctor offices, schools, and pharmacies. Such conclusion is supported by the plots of corresponding ECCs (c.f. Figure 5.15, right panel): The curve computed for clothing stores (blue)

Figure 5.15: Spatial distribution of selected social services with the municipality of Rennes, France (left panel), corresponding Euler curves (right panel).



is visually distinct from other curves. Contrary, no statistical differences were observed between the distribution of pharmacies and the distribution of schools – the p-value of the TopoTest is 0.306. All the above conclusions are in agreement with the previous findings about that dataset made using the Fasano-Franceschi test [73, 127]. However, in addition to that, the TopoTest rejects the hypothesis of equal geographical distributions of doctor offices vs. pharmacies and doctor offices vs. schools (in both cases the p-value is below 10^{-4}), while the Fasano-Franceschi does not (p-value 0.881 and 0.435, respectively as computed using `fasano.franceschini.test` R package). This is an interesting observation in the context of previously discussed simulation study results, where we show that TopoTest is more powerful than the Kolmogorov-Smirnov test (closely related to the Fasano-Franceschi test) and hence more often correctly rejects the null hypothesis.

5.6 Discussion

Using Euler characteristic curves, we introduced a new framework for goodness-of-fit testing in arbitrary dimensions. In addition, we provide a theoretical justification of the method. Although the distribution of the test statistic is unknown for finite n, and contrary to the Kolmogorov-Smirnov test, depends on F, the asymptotic distribution is given by (5.4), while theorem 5.1.4 provides an upper bound on the type II error.

A simulation study was conducted to address the power of the TopoTest in comparison with Kolmogorov-Smirnov test. A one- and two-sample setting was considered. In both cases, the TopoTest in many cases yielded better performance than Kolmogorov-Smirnov. It should be however highlighted that Kolmogorov-Smirnov test and TopoTests operate in slightly different frameworks – the former in capable to distinguish between distributions that differ e.g. in location parameter while the TopoTests are insensitive to the changes of the distribution that leave the excess mass invariant, including shifts, rotations and reflections as described in Section 5.1.4.

6 Damage Identification in Rolling Element Bearings Using Topological Data Analysis

Abstract. The problem of bearing damage detection is of high practical importance. Instances characterized by non-Gaussian properties and time-varying operational conditions are of special interest, as many classical methods fail to detect damage in those cases. This work fills the gap by proposing a novel algorithm for detecting damages that is also successful in these challenging scenarios. Formulating the problem in the language of signal processing, the proposed algorithm detects the cyclic impulses p, being an evidence of fault of a machine, embedded in an unknown non-deterministic and non-Gaussian signal of background noise s. Using topological data analysis tools, the chapter presents a method to analyze the signal $s + p$ and determine the existence of the component p even if it has a small amplitude compared to s. The proposed technique is based on Takens' reconstruction theorem and uses persistent homology methods, which we motivate as a parameter-free generalization of recurrence plots. Specifically, we adapt the methodology of the previous chapter to Betti curves. This approach is agnostic to the model that generates s and p and outperforms alternative techniques. The method was successfully tested for Monte Carlo simulations, test rig data under different speed conditions, and industrial data with serious non-Gaussian disturbances.

Author's contributions. This chapter contains joint work with Justyna Hebda-Sobkowicz, Agnieszka Wyłomańska, Radosław Zimroz and Paweł Dłotko. As of the time of writing, it is being finalized for submission to an industrial engineering journal. N.H. is the lead author, conceived and implemented the TDA method and carried out the machine learning analysis. J.H.-S. provided the baseline results for CVB and (geometric) infograms. A.W. provided supervision for the signal processing aspects of the work, R.Z. provided data and engineering experience, P.D. supervised the TDA aspects of the work. The collaboration was initiated by P.D. and A.W..

This chapter is organized as follows. In Section 6.1, the works related to local damage detection, as well as topological data analysis are discussed. Section 6.2 presents the basic mathematical definitions used in the chapter. Section 6.3 introduces the methodology used to test hypothesis about healthy condition of given time series. In Section 6.4, simulation and laboratory test rig data analysis as well as real vibration signal data examination are presented. Finally, Section 6.5 is a summary with underlined practical information concluded from data analysis.

6.1 Related works

Vibration or acoustic damage detection is widely recognized in the literature as a powerful and popular tool to prevent catastrophic machine failure. The classical approach is based on identifying (detecting) certain properties of the signal of interest (SOI), typically impulsiveness or periodicity [8, 89], which are indicators of malfunctioning machines. The SOI is often hidden in the background noise (produced by the machine and environment) and is impossible to detect in the time domain. In the case of stationary background noise (usually assumed to be Gaussian), the analysis of the envelope spectrum or squared envelope spectrum (SES) of the given signal is sufficient and reveals information about cyclic events. However, industrial data are mostly complicated in terms of the spectral structure of the observed signal, thus various decompositions, such as EMD [111], SVD [110], STFT [116], are used to unravel the complex mixture of the signal. One of the interesting perspectives is to decompose the signal into the time-frequency domain for narrow-band components and use statistical analysis to select an informative frequency band (IFB) that carries information about local faults [47, 139]. Selected IFB is used for data filtration to significantly improve SNR and enable extraction of the local fault frequency (e.g., by SES analysis). In addition, the frequency obtained indicates the frequency of the component of the machine which is then treated as damaged.

A popular tool for IFB indication in the bi–frequency domain is the fast kurtogram [7]. It decomposes the signal into the frequency domain (1/3 - binary tree decomposition) for narrow band subsignals and uses kurtosis to detect the impulsiveness of the envelope of the given subsignal. The main limitation of the kurtogram is its sensitivity to non-Gaussian noise as a consequence of the used kurtosis statistic. In addition, it does not consider whether the identified cyclic component is impulsive. Many methods have been proposed to improve the kurtogram. One of the powerful examples being the infogram [8] which uses more robust statistic (negentropy) to test the impulsiveness of the envelope spectrum of the subsignals, and additionally to identify the existence of cyclicity in the same subsignals at a given step of the frequency decomposition. It considers both characteristics of the local fault and seems to be the ideal tool for damage detection. However, in the case of non-Gaussian noise, some limitations occur as pointed out in [89]. To address them, a Geometric Normalized Infogram [89], GNinfogram for short, was proposed. In the followingthe GNinfogram will be used for comparison purpose.

In addition, several more robust statistics were proposed, such as the Gini index [93, 159], or the conditional variance (CV) statistic [88]. The latter one shows superior performance for difficult noise nature (high energy, non-cyclic impulses). Therefore, the conditional variance-based selector (CVB-selector) was also chosen for comparison in this chapter.

Data acquisition for rotating machinery in real-world applications are exceedingly complex due to various factors such as fluctuations in rotating speeds, loads, and environmental noise. These

non-stationary conditions pose a significant challenge to detect local damage, and advanced methods are necessary [112]. Although machine learning techniques have become popular in this field [87], they require training data. It is essential to note that there is a discrepancy between the training data (the source domain, where the diagnostic model is learned) and the practical testing data (the target domain, where the learned model is deployed for real-time health monitoring), which is inevitable and can influence the results.

Despite the effectiveness of the above-mentioned techniques, they have some limitations, including the difficulty of implementation, computing time, the existence of the appropriate amount of the reference signals, and the ability to handle non-Gaussian noise effectively. Therefore, there is still a need to define simple, understandable algorithms that will effectively deal with various types of signal noise.

The new technique proposed in this chapter is based on topological data analysis (TDA), a new tool that utilizes robust invariant of shapes of considered sets. Recently, it has gained attention in the context of smart manufacturing and industry 4.0 – see the survey article [151] and the references therein. It has already been applied in [125] for periodicity detection in time-series processing. However, the approach presented therein is very sensitive to noise and fails to detect cyclicity under the amounts of noise encountered in the considered applications. Therefore, until now, only a handful of contributions have appeared in the literature to topological data analysis applied to condition monitoring. In a notable series of works [101, 102, 162, 163, 164], such techniques were found to be useful in the detection of chatter. In [137], authors use tools of time-delay embedding followed by computations of one-dimensional persistent homology to detect the existence and potential disruption of the rotary movement of DC motors. Doing so, the authors assume that a sufficiently high-persistence one-dimensional generator is an indication of a healthy engine, while lack of it indicates engine malfunction. However, sufficient details of the experiments are not provided, and hence it is unknown what severity of the malfunction is captured by the proposed method. Another application of TDA to electric motors was given in [158], in which Betti curves are shown to predict eccentricity fault. The time series in their work are phase current measurements, which are not impaired by added noise. A paradigm of time-delayed embedding followed by persistence homology computations is also applied in [98, 150] where the authors are testing the proposed methodology on a number of artificial time series as well as time series obtained from wearable devices. The use of persistent homology with a shortest path distance in time delay embedding was recently suggested by Fernandez et al. [74] and successfully applied to anomaly detection of an ECG time series. In none of these works, the significance of the findings is statistically investigated. Topological data analysis methods are not limited to time series data, but are also applied to detect errors in additive manufacturing [18] or wafer defect patterns in semiconductor manufacturing [103]. Moreover, the essential, as is shown in this work, effect of the type and level of noise in the signal has not yet been adequately addressed. Therefore, no clear demonstration of the practicality of the TDA approach in the field of mechanical damage detection can be found in the literature.

6.2 Preliminaries

6.2.1 Takens' Embedding for Dynamics Reconstruction

Let us start from Takens' embedding theorem [35]. Take a phase space manifold X and a diffeomorphism $f\colon X \to X$ generating a dynamics on X. One may think of X as a space of possible states of the machine considered and f as a working condition of machines that takes it from a state $X \ni x_i$ to $f(x_i) = x_{i+1} \in X$. Starting from an initial state x_0 the following states $x_0 = f^0(x_0), x_1 = f^1(x_0), x_2 = f^2(x_0), \ldots, x_n = f^n(x_0)$ describe the working condition of the machine. Moreover, we assume that X has dimension $\dim(X)$ (box counting) (see [35]). Informally, it means that one needs $\dim(X)$ degrees of freedom to characterize space X.

Take $m : X \to \mathbb{R}$, a continuously differentiable generic measurement function of the observable states of X. The values of m can be treated as the results of measurements using a sensor on the machine considered (Figure 6.1a). It associates the states of the system

$$x_0, x_1, x_2, \ldots, x_n$$

with the one-dimensional time series of measurements

$$Z = [m(x_0), m(x_1), m(x_2), \ldots, m(x_n)].$$

The celebrated Takens' Embedding Theorem [35] states that it is possible to reconstruct X and f given Z. The (state space) reconstruction $\mathbb{R}^d \supseteq R = [r_0, r_1, \ldots, r_k, \ldots]$ consists of time delay vectors in d-dimensional space defined for a *lag* $\tau > 0$ as follows:

$$\begin{aligned}
r_1 &= (m(x_0), m(x_\tau), m(x_{2\tau}), \ldots, m(x_{d\tau})) \\
r_2 &= (m(x_1), m(x_{\tau+1}), m(x_{2\tau+1}), \ldots, m(x_{d\tau})) \\
&\vdots \\
r_k &= (m(x_k), m(x_{\tau+k}), m(x_{2\tau+k}), \ldots, m(x_{d\tau+k})) \\
&\ldots
\end{aligned}$$

Takens' theorem states that if $d \geqslant 2\dim(X) + 1$, the space R with map $\hat{f} : R \to R$, $\hat{f}(r_i) = r_{i+1}$ is dynamically equivalent to the map f on X. Although the theorem holds for an arbitrary lag, a proper choice of τ is important (see [97, Chapter 3]; it can be estimated using mutual information). Typically, the dimension d of the attractor is not known but It can be estimated from data using standard techniques [90, 99], like false nearest neighbors (FNN). A different heuristic for choosing τ and dim was suggested by Perea and Harer [125, p. 803] in the context of periodicity scoring using one-dimensional persistence. They recommend to select them in such a way that the product $\dim \cdot \tau$ approximates the length of the period of the signal. This is practical in our setting, as we know the potential period length from the rotational speed of the machine.

6.2.2 From Recurrence Plots to Persistent Homology

In what follows, we will not use the information on dynamics, but will purely restrict ourselves to analyze the shape of the reconstructions R (Figure 6.1b). We rely on the assumption that two sets R_1, R_2 generated by the same dynamics f on X will have a similar shape. Given a third set R_3 generated by different dynamics, we expect its shape to be different from those of R_1 and R_2. It is important to note that while the shapes of R_1 and R_2 generated by the same dynamics will be similar, it is possible that the shape of R_3 generated by different dynamics can also be similar to those of R_1 and R_2. However, if the shape of R_3 is different from that of R_1 and R_2, then we can safely conclude that R_3 was generated from a different dynamics than R_1 and R_2.

To make such a comparison, stable characteristics of the shape of R are needed. A classical tool for this task are *recurrence plots*, introduced in the seminal paper of Eckmann et al. [67]. They construct an $R \times R$ matrix with entry at (i, j) equal to 1 if r_i and r_j are close to each other and 0 otherwise (Figure 6.1c). To be precise, we will consider r_j to be close to r_i if $\|r_i - r_j\| \leqslant \varepsilon$ for some $\varepsilon > 0$. (Note that Eckmann et al. in their original article suggest choosing ε_i such that $B(r_i, \varepsilon_i)$ contains the ten nearest neighbors, this relation is not symmetric, and we instead take ε independently of i, j.)

A different way of visualizing the same recurrence information is via *recurrence networks*[64]: This is a network which has the points of R as nodes and a connection between r_i and r_j if $\|r_i - r_j\| \leqslant \varepsilon$ (where $i \neq j$). In graph-theoretic language, we say that the recurrence plot is the adjacency matrix of the recurrence network. We will see an example in Figures 6.1b and 6.1c. Recurrence in the dynamics gives rise to cycles in the recurrence network, which means we can detect features of the dynamics via topological features of the recurrence network. However, three subsequent points in R might be close enough together that they form a cycle in the network, but we do not wish to consider this cycle as evidence for recurrence in the dynamics; i.e. presence of a periodic feature in the signal. Moreover, the question of how to choose the proximity parameter $\varepsilon > 0$ remains. To address these two problems, we employ *Vietoris-Rips persistence*: We consider the Vietoris-Rips complex as a generalization of recurrence networks as follows: While the recurrence network is built from 0- and 1-dimensional blocks, i.e. nodes and edges, the Vietoris-Rips complex also has 2-dimensional building blocks, namely triangles. The small triangular cycles which we regarded as problematic in the recurrence network perspective thus become filled in. This motivates the idea to *consider only those cycles as features of the Vietoris-Rips complex which are not filled in by triangles*. In other word, we are led to consider its one-dimensional *homology*.

The notion of *persistence* means that we consider *all* possible values of ε and keep track of the topological features of the Vietoris-Rips complex as ε increases from 0 towards ∞. Note that as ε increases, edges and triangles will appear but never vanish; connected nodes stay connected.

It is the presence of 1-dimensional features in persistent homology which capture the circular shape of the state space reconstruction we consider, and thus we would it regard it as evidence for recurrence.

Indeed, consider this construction in combination with the previous ideas of time delay embedding. Figure 6.1a shows a sampled time series. It gives rise to the state space reconstruction shown in Figure 6.1b, where we also draw connections between nearby points to form the recurrence network at $\varepsilon = 0.5$. The 1-dimensional features of the corresponding Vietoris-Rips persistence, $\mathrm{Dgm}(H_1(\mathcal{R}(R)))$, are shown in the diagram 6.1d. Points in the rectangle with lower right corner at $(0.5, 0.5)$ (shown in gray) represent loops which were formed earlier (i.e. $\varepsilon < 0.5$) and have not yet been filled in. These are the 1-dimensional topological features present in the Vietoris-Rips complex

at $\varepsilon = 0.5$. Recall that taking the number of features present at ε as a function of ε gives rise to the *Betti curve* (Definition 2.3.24). Formulated in the broader context of the mechanical problem under investigation, bearing faults manifest themselves in periodic features in the signal, which we regard as recurrent feature of the dynamics. This is reflected by a circular shape of the state space reconstruction, which in turn yields a 1-dimensional topological feature in the Vietoris-Rips complex present (*persisting*) for a long range of ε-values, for which the Betti curve is therefore non-zero.

In the experiments presented in this chapter, we use a variation known as weighted Vietoris-Rips complexes (cf. Definition 2.2.31 and [5]) designed for noisy data. (Note that taking the distance to nearest neighbors into account bears some similarity to how Eckmann et al. suggested to choose a radius ε_i around r_i as the distance to the tenth-nearest neighbor [67].)

In the presented methodology section, we will operate on $\beta_1(H_1(\mathcal{R}(R)))$ in order to draw conclusions about R. We restrict ourselves to 1-dimensional features of the state space because they have already been used in the literature [125]. In addition, Betti curves are known to be robust in the L^1 distance to small perturbations in the data [51, 61].

6.3 Methodology

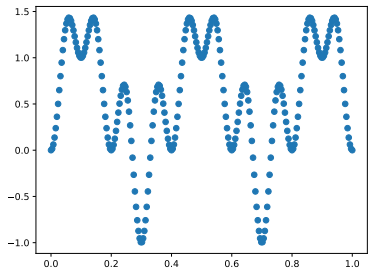
This section presents the methodology used to test a hypothesis about time series allowing to discriminate those that do not contain cyclic impulses from those that contain them. In a nutshell, the proposed pipeline consists of three main parts:

Algorithm 6.1: Computing TDA Betti curve summary from signal Z

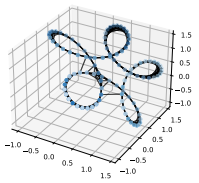
Function `BettiCurveFromSignal`(Z)**:**

$\tau \leftarrow$ `MutualInformationDelayEstimator`(Z)
$d \leftarrow$ `FNNDimensionEstimator`(Z, τ)
$R \leftarrow$ `TimeDelayEmbedding`(Z, τ, d)
$\overline{R_Z} \leftarrow \left\{ \frac{r - \text{mean}(r)}{\|r - \text{mean}(r)\|} : r \in R_Z \right\}$
$DTM \leftarrow$ `DistanceToMeasure`$(\overline{R}_Z, n_{neighbors}$=5)
$\beta_1^Z(\varepsilon) \leftarrow \dim($`PersistentHomology`$(DTM, \varepsilon))$

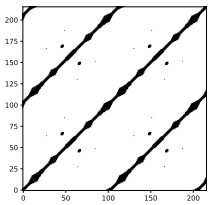
1. Given a time series Z, motivated by Takens' reconstruction theorem, we construct its *time delay embedding* R_Z to reconstruct the phase space of dynamical system generating Z via a generic observable. We employ the normalization procedure of [125], in which the one-dimensional persistent homology of time delay embeddings was used to construct a periodicity score. One first projects onto the hyperplane orthogonal to $(1, \ldots, 1)$ and then to the unit sphere on that subspace:
$$\overline{R_Z} = \left\{ \frac{r - \text{mean}(r)}{\|r - \text{mean}(r)\|} : r \in R_Z \right\}.$$
2. The tools of *persistent homology* in particular one-dimensional *Betti curves*, β_1^Z for short, are used to measure the shape of the normalized phase space $\overline{R_Z}$. Specifically, we use a distance-to-measure Rips filtration [5] with weights equal the distance to five nearest neighbors.



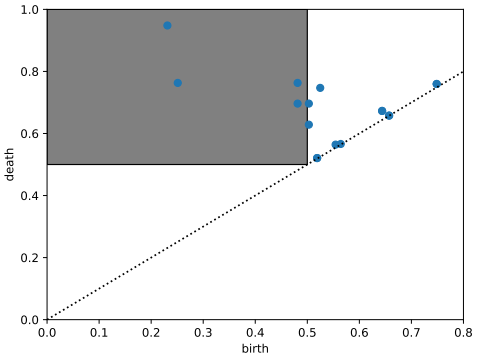
(a) Example of a sampled periodic time series.



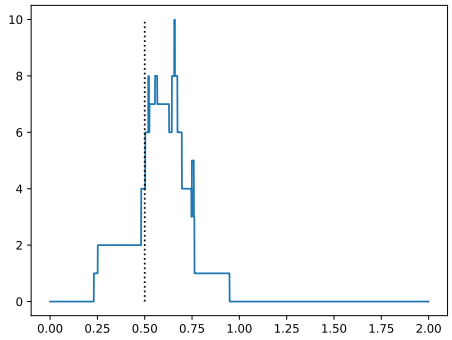
(b) Time delay embedding with drawn-in recurrence network at $\varepsilon = 0.5$.



(c) Recurrence plot using $\varepsilon = 0.5$.



(d) Persistence diagram of the time delay embedding with drawn-in rectangle containing cycles present at scale $\varepsilon = 0.5$.



(e) Betti curve.

Figure 6.1: The topological data analysis signal processing pipeline. The original signal 6.1a is transformed to a high dimensional point cloud via Takens' embedding 6.1b. On that, a recurrence network is built, which has adjacency matrix equal the recurrence plot 6.1c at a given radius (here 0.5). As ε varies from 0 to ∞, loops appear and subsequently get filled in with triangles, this information is stored in the persistence diagram 6.1d. The number of features present at a given ε can be read off by counting the points to the left and above $(\varepsilon, \varepsilon)$. In our example, these are the four points in the grey region left and above of $(0.5, 0.5)$. Plotting this number as a function of the varying ε yields the Betti curve 6.1e, which thus serves as a functional summary of the geometry and topology of the (reconstructed) state space.

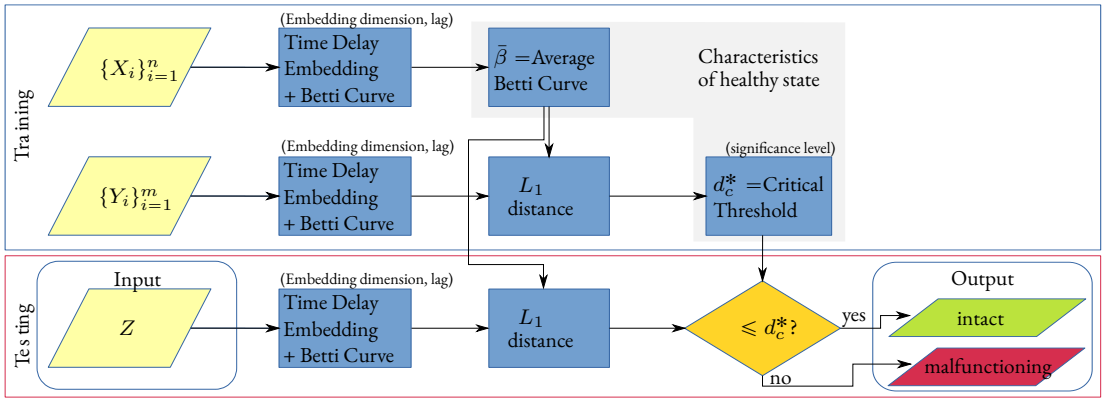


Figure 6.2: Illustration of the proposed methodology. In order to obtain a signature of a healthy state, generated by the fixed model s, a collection of time series $\{X_i\}_{i=1}^n$ is sampled from s, their time delay embedding and the corresponding Betti curves are computed. By averaging them the curve $\bar{\beta}$ is obtained. In order to have an information about the deviation from the mean of Betti curves generated by s, another collection of time series $\{Y_i\}_{i=1}^m$ is computed along with their Betti curves. On a given significance level c, the critical value d_c^* of L^1 distances is computed. Any time series Z giving rise, via the proposed pipeline, to a Betti curve that is farther away from $\bar{\beta}$ than d_c^*, is considered not to be sampled from s and therefore containing some component additional to s. The same pipeline can also be used for any vector summary of the signal in place of Betti curves.

3. For a number of time series and their Betti curve descriptors, statistical tests and standard machine learning techniques are applied in order to discriminate different classes of time series generated by different dynamics.

A pseudo-code of the proposed methodology can be found in Algorithm 6.1 and a visual summary in Figure 6.2.

6.3.1 Statistical Testing

We propose a statistical hypothesis testing framework for functional/vector summaries of time series. Such summaries include classical techniques like CVB and (geometric normalized) infograms, which were originally introduced to identify IFB. Moreover, we analyze a summary based on topological data analysis, namely Betti curves and compare it to the present spectral methods. Below, the set-up is illustrated using Betti curves, the same ideas are then also applied to CVB and (geometric normalized) infograms. Moreover, one can combine different vectorizations by simply concatenating the vectors. Firstly, in order to test the methodology on large-scale computations, Betti curves will be incorporated into a framework of statistical hypothesis testing as developed in [63] (which is Chapter 5 in this thesis). While this adaption does not enjoy the theroretical guarantees of the ECC of a Čech/Alpha complex, there are several reasons motivating this apprroach. We found the embedding dimension to be too high to be computationally tractable using Alpha complexes or ECCs. However, the two-skeleton of the Rips is quick to compute independently of the ambient dimension and one-dimensional persistence has previously been used for periodicity scoring [125] and Betti curves were used for classification [150] of time series. The null hypothesis to test is that the machine is intact/healthy, in other words,

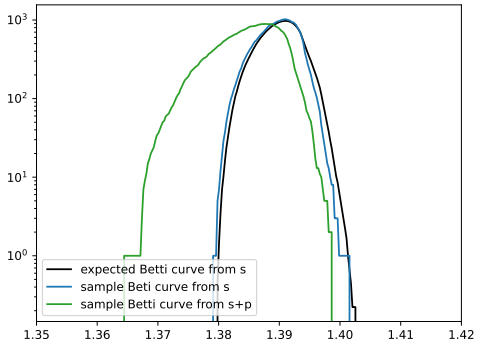


Figure 6.3: The expected Betti curve under the null hypothesis is shown in black. A sample Betti curve from the null model (α-stable noise with $\alpha = 1.9$) shown in blue follows the expected curve closely. A sample Betti curve from cyclic impulses added to the noise of the null model shown in green is significantly different, evidenced by a large L^1 distance.

the signal only contains noise but no impulses. In this instance, we assume to be able to sample n time series $\{X_i\}_{i=1}^n$ from the null distribution, i.e. healthy machine. Subsequently their time delay embeddings and 1-dimensional Betti curves $\beta_1^{X_1}, \dots \beta_1^{X_n}$ are computed. The averaged Betti curve $\bar{\beta_1} = \frac{1}{n}\sum_{j=1}^n \beta_1^{X_j}$ is obtained as a signature of a healthy state.

In order to determine the level of variation of the signal under the null hypothesis, additional m time series $\{Y_i\}_{i=1}^m$ are sampled and their Betti curves $\beta_1^{Y_1}, \dots \beta_1^{Y_m}$ are obtained. Note that L^1 puts the most weight of all L^p metrics in differences in the support, which are what we are particularly interested in – cf. Figure 6.3. Subsequently their L^1 distances $d_1, \dots, d_m$ from $\bar{\beta_1}$ are computed, $d_j = \int_0^\infty |\beta_1^{Y_j}(r) - \bar{\beta_1}(r)|\mathrm{d}r$. Given a significance level $c > 0$, a threshold value of a distance d_c^* is chosen such that $d_j > d_c^*$ only for $c \cdot m$ of the sample distances d_j.

In the described scenario take a new time series Z. Our null hypothesis is that Z is sampled from a healthy state. In order to test the hypothesis, time delay embedding, of the same parameters, followed by 1-dimensional Betti curve computations are performed to obtain β_1^Z. If the L^1 distance between β_1^Z and $\bar{\beta_1}$ is greater than d_c^*, the null hypothesis is rejected. In the other case, there is no evidence to reject it. For an illustration of the deviation from the expected Betti curve, see Figure 6.3.

This methodology allows for large-scale Monte Carlo testing of the efficiency of the presented procedure. The results of those tests are discussed in Section 6.4.1. Furthermore, note that no data about malfunctioning machines was needed as input.

6.3.2 Further Analysis of Topological Signatures

While the pipeline taking at the input time series and putting out its Betti curves can be used for large scale Monte-Carlo comparisons, it has much broader applicability. Let us consider the case of multiple time series $\{X_i^1\}_{i=1}^{l_1}, \{X_i^2\}_{i=1}^{l_2}, \dots, \{X_i^k\}_{i=1}^{l_k}$. The Betti curves corresponding to them may be computed, vectorized and used for classifications. This approach is illustrated in particular, in Section 6.4.3, where both support vector classifier as well as two dimensional principal component

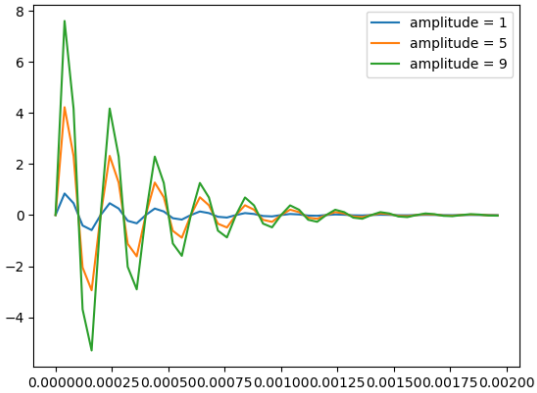


Figure 6.4: Different amplitudes of a single exponential-modulated pulse.

projection of the vectorized Betti curves are used to understand different conditions of the machine. For background on these standard machine learning techniques, we refer the reader to the textbook [94]. We employ them for real world data, of which not enough samples are available for our hypothesis testing approach.

6.4 Results

All the experiments were carried out on a workstation computer with an AMD Ryzen Threadripper PRO 5955WX 16-Cores CPU and 256 GB of RAM running Ubuntu 22.04.1 LTS. The code can be found on the author's github. The CVB curve is usually discretized on a equi-spaced grid of points as a vector in 257-dimensional space. For the Betti curves, we use a grid of size 257 to improve comparability with CVB and GNinfogram. The TDA computations were carried out using the library [117].

6.4.1 Simulated Data Analysis

The considered simulated signals are composed of two main elements:

- The s component being in this case the α-stable noise that represents the background noise (compare top left panel of Figure 6.5). The α-stable distribution is a statistical model that can capture non-Gaussian or impulsive behavior in data. It can be treated as the generalization of the Gaussian distribution, as for the parameter $\alpha = 2$ ($\alpha \in]0, 2]$ - stability index) it reduces to Gaussian noise, whereas for $\alpha < 2$ the non-periodic impulses appear (distribution becomes more impulsive) which are treated here as the specific characteristic of machine work. In particular, low values of α provides heavy-tailed distributions. In this chapter, we apply the symmetric version of this distribution. The symmetric α – stable distributed random variable A is defined by the characteristic function [138] of the following form:

$$\phi_A(\theta) = \mathbb{E}\left(e^{iA\theta}\right) = \begin{cases} e^{-\sigma^\alpha |\theta|^\alpha}, & \alpha \neq 1, \\ e^{-\sigma|\theta|}, & \alpha = 1, \end{cases}$$

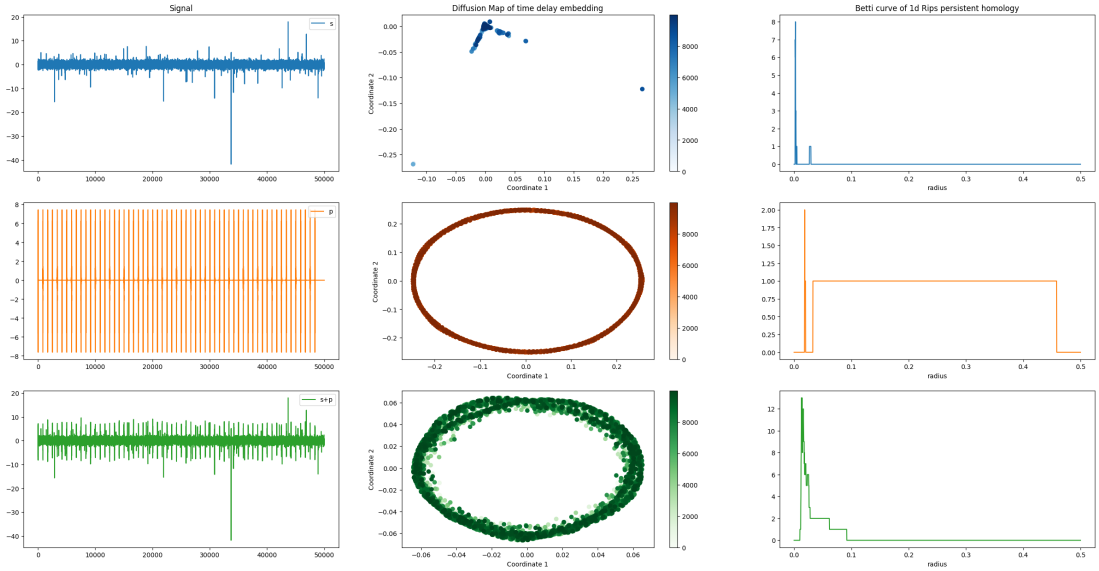


Figure 6.5: Illustration of the set-up of our synthetic dataset. We show signals (left row), (the first two diffusion map components [20] of) their time delay embedding using dimension 1666, delay 3 (middle row) coloured by time index, and the Betti curve of 1-d Rips persistent homology of a subsample of size 128. Diffusion maps [53] are a non-linear dimensionality reduction technique which we use purely for visualization here; our TDA algorithms operate in the high-dimensional state-space reconstruction. The top row shows pure α-stable noise with $\alpha = 1.95$ (s in the text) – the time delay embedding does not admit interesting topological structure and the Betti curve is very narrow. The middle row shows cyclic impulses with amplitude 7 on their own (p in the text) – we see how the time delay embedding follows a cyclic trajectory and the Betti curve is very wide. The bottom row shows the setting of our interest, in which the signal is the sum of the previous two ($s + p$ in the text). Although the time delay embedding is rather noisy, we see a clear topological feature, namely a cycle, which is reflected by a somewhat wider Betti curve than for the pure noise case. This leads to a significantly increased L^1-distance compared to what would be expected under the pure noise model, compare Figure 6.3.

where $\sigma > 0$ is the scale parameter. The details of the α-stable distribution can be found in [138]. The α–stable distribution model is often used to describe the statistical features of non-Gaussian signals [80].

- The p component represents the exponential-modulated sinusoidal periodic impulses related to a local defect (compare Figure 6.4 for different amplitudes of a single impulse and left row, middle panel in Figure 6.5 for the periodic impulses). One can find more about the model of the simulated signal in [161].

In this experiment we measure the test power, that is, how often a time series is classified as containing cyclic impulses when it actually does. Hence power as close to 1 (bright color in Figure 6.6) as possible is desirable. All time series are of length 50000. We considered $\alpha \in \{1.6, 1.65 \ldots, 2\}$ and impulse amplitude $\in \{5, 5.5, \ldots, 9\}$ (see Figure 6.4 for an illustration of some amplitude values for a single impulse without any noise). For each α, we simulate the average curve under H_0 using 1000

samples consisting purely of α-stable noise. Then we compute the acceptance threshold using 1000 independent samples. Subsequently, for each level of impulse amplitude, we simulate 1000 time series with this kind of signal and count how often we reject the null hypothesis. The summary of our extensive Monte Carlo simulation study on synthetic data can be seen in Figure 6.6.

The figures present the test power in relation to the amplitude of the impulses (horizontal axis) and the α-parameter of the ambient noise (vertical axis). The infogram-based selector [8] shows poor performance except for the Gaussian case $\alpha = 2$. The geometric normalized infogram increases the test power considerably. While the CVB selector has similar, only slightly better power compared to the previous, the TDA approach yields the best results as illustrated by the big bright region in the figure. For a specific example, consider the amplitude 7 and $\alpha = 1.8$ – CVB has a power of 0.546 whereas TDA admits 0.96, significantly better. The difference in power is shown in Figure 6.7, which is the difference of the number for TDA minus those for CVB from Figure 6.6. Values less than zero indicate better performence of CVB, values greater than zero show that TDA outperforms CVB. One can observe the latter case holds for a range of parameters; in particular, for high amplitudes and low values of α. Notably, the simulations for the TDA approach take 67 seconds, which is significantly faster than the CVB approach, which takes 837 seconds. Infograms are the slowest to compute with several hours.

For the following experiments we focus on CVB and TDA because they perform the best in the simulation study.

6.4.2 Laboratory Test Rig Data Analysis

To test the presented methodology on non–synthetic data a test rig presented in Figure 6.8 has been used to acquire experimental data in laboratory conditions. It contains an electric motor, gearbox, couplings, and two bearings, of which one was deliberately damaged. We[1] recorded a 40-second-long vibration signal with a sampling rate of 50 kHz using an accelerometer (KISTLER Model 8702B500), which was stacked horizontally to the shaft bearing.

Data has been acquired using two channels and four different rotating speeds for each baseline and faulty state, totaling in 16 time series. This situation presents the challenge of data scarcity, therefore we cannot proceed with the hypothesis testing pipeline outlined above. However, the results for the synthetic data establish that Betti curves as a featurization/functional summary of a time series is on the level with or even better than state of the art spectral methods. In order to gauge the significance of the presented methodology we first split each time series into ten parts. Then we compute the Betti curves as described in Section 6.2 as well as 6.3.2. The Figure 6.9 shows a plot of the first two principal components of the obtained Betti curves. We observe a clear separation between healthy and malfunctioning machines across all speeds and channels. Even the different speeds can be distinguished with the bare eye.

Moreover, Figure 6.10 shows the L^1 norms of the Betti curves showing the significance of the observed differences. Focusing on a single channel and speed we observe that the norms of healthy machines are consistently higher and do not overlap with those of malfunctioning machines. In addition, we trained a linear Support Vector Machine (SVM) [56] classifier 100 times, leaving out

[1] collaborators JHS and RZ at the Faculty of Geoengineering, Mining and Geology; Wrocław Universisty of Science and Technology

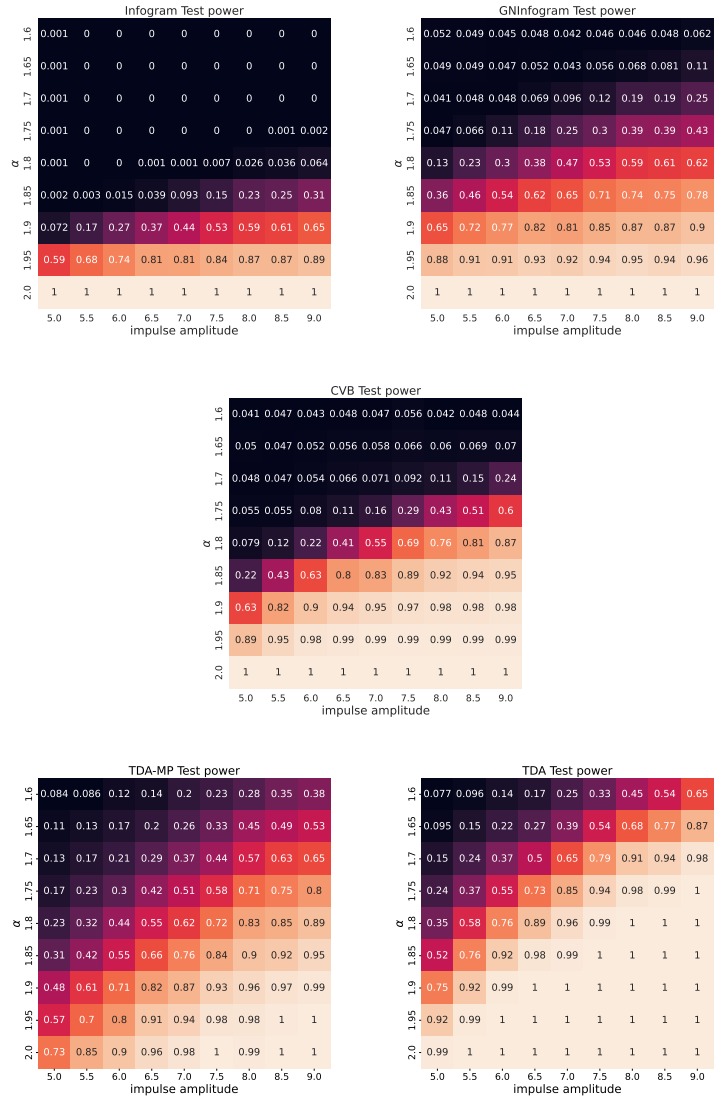


Figure 6.6: Test powers when the noise is from an α-stable distribution, obtained using different descriptors: infogram [8], geometric-normalized infogram [89], conditional variance based (CVB) [88], TDA maximum persistence [101]; TDA Betticurves. Brighter colors and higher numbers correspond to better performance.

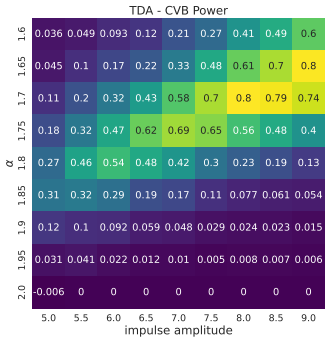


Figure 6.7: Difference of power between TDA and CVB approach: The yellow region indicates parameter choices in which TDA significantly outperforms CVB.

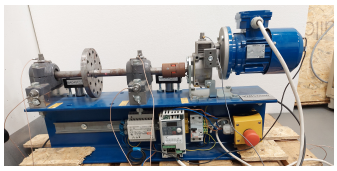


Figure 6.8: Test rig used in the experiment.

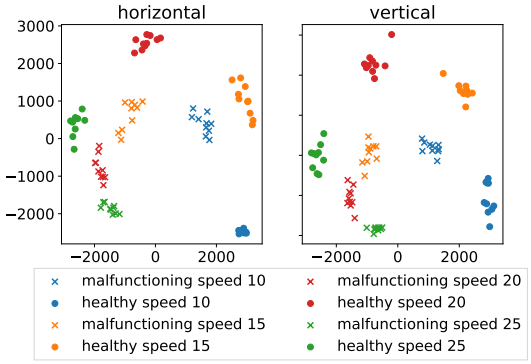


Figure 6.9: PCA of Betti curves from test rig measurements.

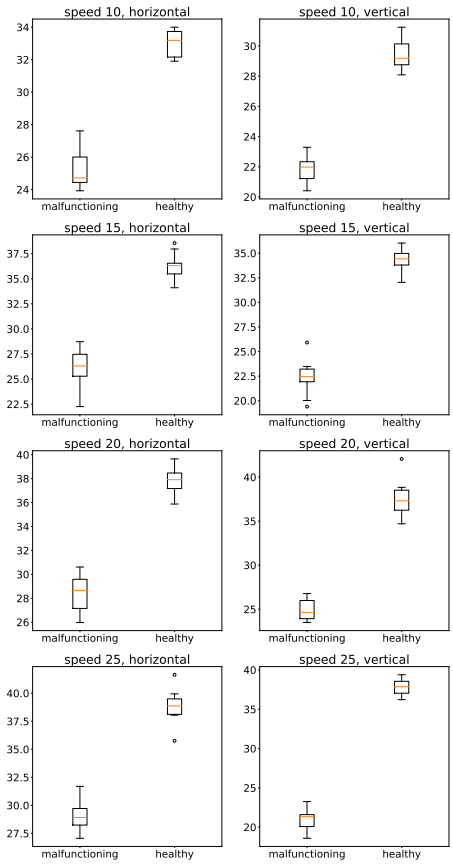


Figure 6.10: Boxplots of L^1 norms of Betti curves from test rig measurements.

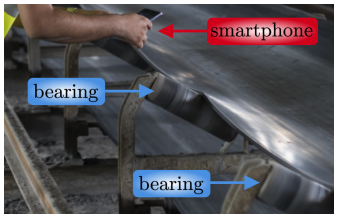


Figure 6.11: An idler inspection and using a smartphone for acoustic data acquisition.

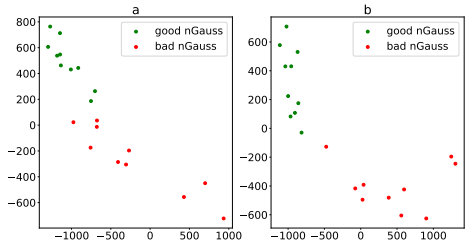


Figure 6.12: PCA of Betti curves from real-world measurements.

half of the data at random for each iteration. Testing on the previously left-out part of the data, we consistently obtain perfect 100% separation of good and malfunctioning machines.

6.4.3 Industrial Data Analysis

To test the presented methodology on an industrial site, an acoustic signal from a bearing installed in an idler in a belt conveyor has been recorded. Rotating components known as idlers are utilized in belt conveyors to provide support to the moving belt, as shown in Figure 6.11. These idlers typically comprise a shaft, two bearings, and a coating. The duration of the signal is c.a. 10 seconds with a sampling frequency of 48 kHz. Manual analysis (signal by signal) has shown that some idlers produce cyclic impulsive noise. Thus, by visual inspection we consider the signals as non-Gaussian distributed, which renders damage detection challenging.

We investigate such time series with the non-Gaussian type of baseline noise. As in the previous cases we have two channels (labelled 'a' and 'b' in the figures) and we split each time-series into ten parts enabling some basic statistical analysis. In Figure 6.12 we show the first two principal components of the Betti curves of the time delay embeddings. We report the best results, which were obtained for $\dim = 98, \tau = 96$, which leads to $\tau \cdot \dim$ being around one period (which we know from the rotational speed of the machine); this follows the suggested parameter choice of [125]. The parameters obtained from FNN and mutual information estimation led to slightly worse results. The data is much more noisy than for the test rig from the laboratory, however, we still see some separation in the PCA, although it is not visible in the boxplots of the L^1 norms in Figure 6.13. As it is more sensible to separate in the ambient space than in a projection, we trained a support vector machine classifier with linear kernel using cross validation with a 50-50 train-test split and 100 iterations, a simple standard machine learning tool [94, Chapter 9]. This means we perform training 100 times, leaving out half of the data of either class at random in each iteration. Then we test the classifier each time on the

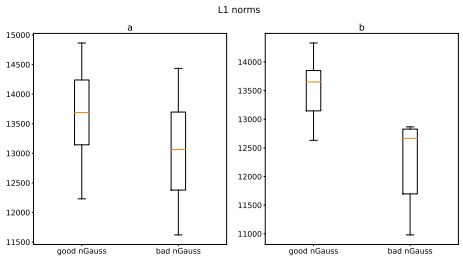


Figure 6.13: Boxplots of L^1 norms of Betti curves from real-world measurements.

	Laboratory test rig data				Industrial data
	Speed 10	Speed 15	Speed 20	Speed 25	
CVB	100 ± 0	100 ± 0	100 ± 0	100 ± 0	88 ± 7
TDA	100 ± 0	100 ± 0	100 ± 0	100 ± 0	92 ± 6
CVB +TDA	100 ± 0	100 ± 0	100 ± 0	100 ± 0	96 ± 5

Table 6.1: Mean accuracy of the SVM classifier $[\%]$ and standard deviation.

other half of the data. We achieve a classification accuracy of 0.92 ± 0.06, an improvement over CVB (0.88 ± 0.07). Hence, the topological descriptors allow to detect damage in rolling element bearings even in challenging real-world industrial settings. We also investigated a combination of Betti curves with CVB by simply concatenating the normalized vectors. This increased the classification accuracy to (0.96 ± 0.05). For even further validation of the method, more measurements would be required.

Discussion

Our findings are summarized in Table 6.1. We note that the data from the test rig admits perfect separation accuracy in each case. This reflects the controlled conditions under which the data was obtained; such ideal behavior cannot be expected in an actual industrial context. Indeed, we see that the industrial data we studied poses a bigger challenge. In the presence of a non-Gaussian type of noise, TDA outperforms the CVB method, and combining both yields even higher scores. However, as we show in the simulation study in section 6.4.1, we expect that these results strongly depend on the degree of non-Gaussianity (amplitude and amount of non-cyclic impulses, not related to local fault). Moreover, the time series we considered in sections 6.4.2 and 6.4.3 were split into pieces, which then constitute highly dependent samples, influencing the classification problem.

6.5 Conclusions

In this chapter, we target to detect the existence of cyclic impulses p embedded in an unknown (non-deterministic, non-Gaussian) signal s. The proposed technique is based on Takens' reconstruction theorem and uses methods of persistent homology to detect subtle changes in dynamics generating the signal $p + s$ compared to dynamics generating the signal containing only the s component.

The performance of the proposed method was tested for various scenarios showing a considerable improvement compared to the state of the art. In particular, as observed in the simulation study, our method excels in regimes with high amplitude of cyclic impulses embedded in very impulsive/non-Gaussian noise. Notably, combining CVB, which is a state of the art technique using spectral methods, with our new topological approach yields the best results. We interpret this finding as evidence that the information gained using TDA was not detected by the spectral method.

We hope that the proposed procedure can be useful for damage detection in a wide range of industrial scenarios as a general failure detection technique that is agnostic to many conditions under which the considered signal was generated. We plan to apply the proposed method for other types of noise as well as for more complicated signals (for example from planetary gearbox).

7 Density Sensitive Bifiltered Dowker Complexes via Total Weight

Abstract. In this chapter, we introduce a density-sensitive bifiltration on Dowker complexes. Previously, Dowker complexes were studied to address directional or bivariate data whereas density-sensitive bifiltrations on Čech and Vietoris–Rips complexes were suggested to make them more robust. We combine these two lines of research, noting that the superlevels of the total weight function of a Dowker complex can be identified as an instance of Sheehy's multicover filtration. An application of the multicover nerve theorem then provides a form of Dowker duality that is compatible with this filtration. As an application, we find that the subdivision intrinsic Čech complex admits a smaller model; we also establish its robustness by linking it to the subdivision-Rips bifiltration. Moreover, regarding the total weight function as a counting measure, we generalize it to arbitrary measures and prove a density-sensitive stability theorem for the case of probability measures. Additionally, we provide an algorithm to calculate the appearances of simplices in our bifiltration and present computational examples.

Author's contributions. This chapter contains joint work with Jan Spaliński [91]. J.S. proposed to study the two-parameter persistence of the multineighbor complex in a TDA context. N.H. is the lead author, conceptualized the general theory in terms of Dowker complexes, proved the stability results and implemented the algorithm.

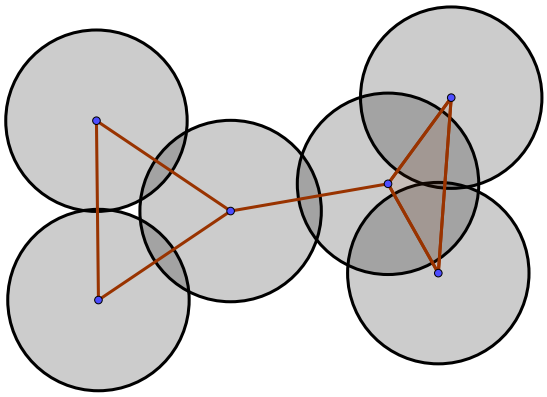


Figure 7.1: Repeating Figure 2.3a: From the perspective of Dowker complexes, the points in the intersections are witnesses (compare the discussion in Example 2.2.29). In this chapter, we are going to take the size of these intersections into account.

In this chapter, we are going to study bifiltrations of Dowker complexes (Definition 2.2.24) in order to make them sensitive to the density of the data and robust against outliers. Recall that the Dowker complex $\mathcal{D}(X, Y, R)$ of a relation $R \subseteq X \times Y$ between two sets X, Y has vertex set X and $\varnothing \neq \sigma \subseteq X$ forms a simplex if there is some *witness* $y \in Y$ with $\sigma \times \{y\} \subseteq R$. We will refine the condition of existence of a witness by asking for the number of witnesses to exceed a threshold m. In Example 2.2.29, we saw that we can view the Čech complex of some $X \subseteq (Z, d)$ as Dowker complex via

$$\mathcal{C}(X)_r = \mathcal{D}(X, Z, R_r), \qquad R_r = \{(x, z) \colon d(x, z) \leqslant r\}.$$

Inspecting Figure 7.1, we observe that the set of witnesses of a k-simplex is the intersection of the r-balls around the corresponding $k + 1$ points. While the Čech construction only asks about the existence of a witness, we are interested in how many there are. For a finite metric space (X, d), the intrinsic Čech complex is the Dowker complex $\mathcal{I}(X)_r = \mathcal{D}(X, X, R_r)$, see for instance [36, 45]. In this case, we can simply count the size of the intersections of r-balls. More generally, we will require a measure to capture the mass of these intersections and impose it as a second filtration parameter in Section 7.1. Section 7.2 contains results pertaining to robustness and stability of this bifiltration. Finally, in Section 7.3, we provide an algorithm to calculate such bifiltered complexes from data and provide computational experiments.

As an additional prerequisite, recall the notion of contiguity of simplicial maps:

Definition 7.0.1. Let $f, g \colon K \to K'$ be simplicial maps. They are *contiguous* if for each simplex $\sigma = [v_0, \dots v_k] \in K$, we have the simplex $f(\sigma) \cup g(\sigma) \in K'$, i.e. a simplex formed by the vertices

$$\{f(v_0), \dots, f(v_k), g(v_0), \dots g(v_k)\}.$$

Contiguous simplicial maps are a combinatorial model for homotopic maps in the following sense:

Lemma 7.0.2 ([145, Lemma 3.5.2 and Theorem 4.3.9])**.** *Two contiguous simplicial maps $f, g\colon K \to K'$ induce homotopic geometric realizations $|f|, |g|\colon |K| \to |K'|$. Moreover, they induce chain-homotopic chain maps f_*, g_*, and in particular the same map in homology.*

Contiguity arguments were also originally used for Dowker's duality theorem [65], as well as for the stability result of [45] and the functorial Dowker duality of [50].

7.1 The Total Weight Filtration

As we have seen above, there are in general multiple witnesses for the presence of a simplex. Let us count them:

Definition 7.1.1. The *total weight* function is

$$t\colon \mathcal{D}(X,Y,R) \to \mathbb{N} \cup \{\infty\},\ t(\sigma) = |\{y \in Y : \sigma \times \{y\} \subseteq R\}|.$$

For $m \in \mathbb{N}$, we set $\mathcal{D}(X,Y,R)_m = \{\sigma \in \mathcal{D}(X,Y,R)\colon t(\sigma) \geqslant m\}$.

Observe that we recover the whole Dowker complex for $m = 1$, i.e. $\mathcal{D}(X,Y,R)_1 = \mathcal{D}(X,Y,R)$.

Lemma 7.1.2 ([134, Proposition 2])**.** *The superlevel sets of the total weight function $\mathcal{D}(X,Y,R)_m$ form a filtration by subcomplexes. That is, for $m \leqslant m'$, we have $\mathcal{D}(X,Y,R)_{m'} \subseteq \mathcal{D}(X,Y,R)_m$.*

Proof. First observe that if $\sigma \subseteq \tau$, the total weight satisfies $t(\sigma) \geqslant t(\tau)$, because any witness of τ in particular witnesses σ. Hence, the superlevels of t are indeed subcomplexes. Moreover, they form a filtration because $t(\sigma) \geqslant m' \Rightarrow t(\sigma) \geqslant m$ as $m \leqslant m'$. □

By this lemma, we view $\mathcal{D}(X,Y,R)$ as a functor $]0,\infty[^{op} \to \mathbf{Simp}$. A special case of this filtration has appeared in the following context, which motivated the present study:

Example 7.1.3. Let $G = (V,E)$ be a simple graph, i.e. without loops or multiple edges between the same two vertices. The Dowker complex of the adjacency relation is the neighborhood complex of the graph, introduced by Lovasz [115] in his proof of the Kneser conjecture. It has simplices formed by sets of vertices which have a common neighbor in G. The subcomplex of total weight at least m, for some $m \in \mathbb{N}$, is the m-(multi)neighbor complex of the graph of [12]. The case of Erdös-Renyi graphs has been of particular interest. The adjacency relation of such a graph is given by a symmetric $n \times n$ matrix with zeros on the diagonal, entries above the diagonal independently equal to 1 with probability p and equal to 0 with probability $1-p$, and entries below the diagonal given by symmetry. It turns out that as the dimension of the matrix n increases to infinity, for a large number of choices of p_n (probability) and m_n (number of neighbors), the resulting Dowker complexes are Linial-Meshulam complexes – random simplicial complexes which are d-dimensional with all faces of a lower dimension present and the d-faces present with a fixed probability. Moreover, for each finite simplicial complex X and each m the there exists a threshold $b_m(X)$ such that if $p_n = n^{-1/b}$, where $b > \beta_m(X)$, then asymptotically almost surely the random Dowker complex described above will contain a copy of X (and if $b < \beta_m(X)$ it will not). The details are worked out in [12]. We will return to Dowker complexes of random relations with numerical experiments in Example 7.3.3.

Crucially, one can regard this total weigh filtration of a Dowker complex as an instance of the multicover filtration introduced by Sheehy [143] (see also Definition 2.2.32 as well as [27, 41]), where the cover is given by rows of the matrix that represents the relation, which we make precise below. By the multicover nerve theorem [27, 41, 143], the multicover filtration of $\mathcal{D}(X, Y, R)$ corresponds to the subdivision filtration of $\mathcal{D}(Y, X, R^\top)$.

Theorem 7.1.4. *Let $R \subseteq X \times Y$ be a relation such that*

1. *for all $y \in Y$, there are only finitely many $x \in X$ such that $(x, y) \in R$ (R is* column-finite*),*
2. *for all $x \in X$, there are only finitely many $y \in Y$ such that $(x, y) \in R$ (R is* row-finite*).*

Then we have a weak equivalence of filtrations $|\mathcal{D}(X, Y, R)_\bullet| \simeq |\mathcal{S}(\mathcal{D}(Y, X, R^\top))_\bullet|$, where $\mathcal{S}$ is the subdivision filtration (Definition 2.2.35). Moreover, the weak equivalence is natural in the following sense: If X, Y are fixed and $R_\bullet \colon [0, \infty[\to \mathbf{Rel}$ is a filtration of column- and row-finite relations between X and Y, then we have a weak equivalence of the two bifiltrations

$$\begin{aligned}]0, \infty[^{op} \times [0, \infty[&\to \mathbf{Top}, \\ (m, r) &\mapsto |\mathcal{D}(X, Y, R_r)_m|, \\ (m, r) &\mapsto |\mathcal{S}(\mathcal{D}(Y, X, R_r^\top))_m|. \end{aligned}$$

Proof. Consider the simplices determined by the elements $y \in Y$, i.e.

$$\Delta_y = \{x \in X \colon (x, y) \in R\} \subseteq \mathcal{D}(X, Y, R);$$

they are indeed simplices because the sets are finite by the column-finiteness assumption. Then $\mathcal{A} = \{\Delta_y\}_{y \in Y}$ is a covering of $\mathcal{D}(X, Y, R)$ by (abstract) simplices. As the intersection of simplices is again a simplex, this gives rise to a good cover of a compactly generated space after geometric realization. By construction, the total weight of a simplex counts how many times it is covered. In order to invoke the multicover nerve theorem (Theorem 2.2.36,) we need to check that our cover satisfies the conditions of Theorem 2.2.5. It is locally finite because of the row-finiteness of R. To show that it is locally finite-dimensional, consider any $y \in Y$ and set

$$k_y = |\{y' \in Y \colon \exists x \in X \text{ such that both } (x, y), (x, y') \in R\}|.$$

Since y is in relation with only finitely many x, which in turn each are in relation with only finitely many y', we infer that $k_y < \infty$. Then for any $J \subseteq Y$ with $y \in J$ and $\bigcap_{y' \in J} \Delta_{y'} \neq \varnothing$, we have $|J| \leqslant k_y$ by construction. Moreover, for any $T \subseteq Y$, the set $A_T = \bigcap_{y \in T} \Delta_y$ is a finite simplex if it is non-empty. The latching set

$$L(T) = \bigcup_{T \subsetneq J \subseteq Y} A_J \subseteq A_T$$

is a union of simplices. Thus, after geometrically realizing, the latching space $|L(T)|$ is a closed subcomplex (by the closure-finite property of CW complexes) of the geometric simplex $|A_T|$ and hence satisfies the homotopy extension property.

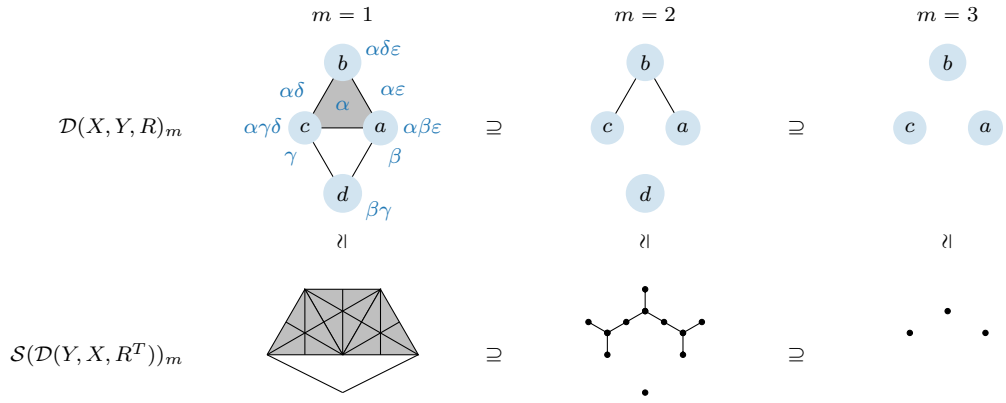


Figure 7.2: Applying Theorem 7.1.4 to the the complex from Example 2.2.27. The top row is the filtration of $\mathcal{D}(X, Y, R)$ by total weight, the bottom row is Sheehy's subdivision filtration [143] applied to the dual Dowker complex $\mathcal{D}(Y, X, R^\top)$.

Finally, to establish the claimed naturality, we consider row- and column-finite relations $R_r \subseteq X \times Y$ for any $r \geqslant 0$ such that $R_r \subseteq R_s$ whenever $r \leqslant s$. In the notation of Theorem 2.2.36, we set F to be the functor

$$\begin{aligned} F\colon [0, \infty[&\to \mathbf{Simp}, \\ r &\mapsto \mathcal{D}(X, Y, R_r), \\ r \leqslant s &\mapsto \mathcal{D}(X, Y, R_r) \hookrightarrow \mathcal{D}(X, Y, R_s). \end{aligned}$$

Moreover, we let $\mathcal{U} = \{U^y\}_{y \in Y}$ be the set of functors

$$\begin{aligned} U^y\colon [0, \infty[&\to \mathbf{Simp}, \\ r &\mapsto \{x \in X\colon (x, y) \in R_r\} \\ r \leqslant s &\mapsto \{x \in X\colon (x, y) \in R_r\} \hookrightarrow \{x \in X\colon (x, y) \in R_s\}. \end{aligned}$$

For each fixed r, the space U^y_r is a simplex we called Δ_y above. Therefore, they form a good cover and satisfy the conditions of [14] after postcomposing with geometric realization. Moreover, for $r \leqslant s$, both $F_{r \leqslant s}$ and all $U^y_{r \leqslant s}$ are inclusions of subcomplexes. Thus, $U^y_{r \leqslant s}$ is the restriction of $F_{r \leqslant s}$. Geometric realization preserves restrictions, hence Theorem 2.2.36 indeed applies to complete the proof. □

See Figure 7.2 for an illustration of the theorem. One perspective on the previous theorem is that the subdivision filtration of a Dowker complex admits a smaller, thus computationally tractable, equivalent. We will come back to this perspective later on, as a variation of Example 7.1.3. Let us point out two desirable potential extensions of the above theorem, which we leave for future research:

Remark 7.1.5. First, it has been variously observed that Dowker duality is equivalent to the nerve lemma [23, p. 1851], [50, Theorem 27]; this raises the question of what the equivalent formulation of the multicover nerve theorem is in terms of Dowker duality. Second, Björner's nerve lemma for covering

simplical complexes [23, Theorem 10.6] does not require any finiteness conditions. One might be able to avoid these in the multicover case as well.

This last issue is relevant because there can be uncountably many witnesses, as we saw in Example 2.2.29 on the Čech complex. To measure their size in such a case, we need the set Y in the relation to be endowed with a measure. If Y is finite, we can take the counting measure to recover the total weight, as we will see in Example 7.1.9.

Definition 7.1.6. Let X be a set, (Y, Σ, μ) a measure space; let $\Lambda\colon X \times Y \to \mathbb{R}$ a function such that for all $x \in X$, the restricted map $y \mapsto \Lambda(x, y)$ is measurable (with respect to $\mathfrak{B}(\mathbb{R})$ on the codomain). Define the *measure Dowker bifiltration*

$$\begin{aligned}\mathcal{MD}(X, (Y, \Sigma, \mu), \Lambda)\colon]0, \infty[^{op} \times [0, \infty[&\to \mathbf{Simp},\\ X \supseteq \sigma \in \mathcal{MD}(X, (Y, \Sigma, \mu), \Lambda)_{m,r} &\Leftrightarrow 0 < |\sigma| < \infty \text{ and}\\ &\mu(\{y \in Y\colon \Lambda(x, y) \leqslant 2r \text{ for all } x \in \sigma\}) \geqslant m.\end{aligned}$$

In the special case in which X and Y are subsets of a common ambient metric space (Z, d), we will use the shorthand notation

$$\mathcal{MD}(X, \mu, \Lambda) = \mathcal{MD}(X, (Y, \mathfrak{B}(Y), \mu), \Lambda);$$

if $\Lambda = d|_{X \times Y}$, we omit it from the notation.

In words, one includes σ in $\mathcal{MD}(X, \mu)_{m,r}$ if the intersection of the $2r$-balls centered at the points in σ has at least measure m with respect to μ.

Example 7.1.7. Fixing r, the complexes $\mathcal{MD}(X, Z)_{\bullet,r}$ form a filtration of the Čech complex at scale $2r$. Recalling Example 2.2.29 and Figure 2.3a, the set of witnesses of a k-simplex is the intersection of the corresponding $k + 1$ balls. With the new filtration parameter, we control precisely the measure of these intersections. Observe the relation to the measure bifiltration, in which we also include balls of radius r if their mass exceeds a threshold; however, one does not impose further restrictions on the mass of the intersection there.

Lemma 7.1.8. *Let X be a subset of a metric measure space (Z, d, μ). If $(Z' d')$ is another metric space and $\varphi\colon (Z, d) \to (Z', d')$ is an isometry, then it induces an isomorphism of filtered simplicial complexes $\mathcal{MD}(X, \mu) \xrightarrow{\cong} \mathcal{MD}(\varphi(X), \varphi_{\#}(\mu))$.*

Proof. First fixing m, r, we want to show

$$[x_0, \dots, x_k] \in \mathcal{MD}(X, \mu)_{m,r} \Leftrightarrow [\varphi(x_0), \dots, \varphi(x_k)] \in \mathcal{MD}(\varphi(X), \varphi_{\#}(\mu))_{m,r}, \text{ i.e.}$$
$$\mu(\{z \in Z\colon d(z, x_i) \leqslant 2r \text{ for all } i\}) \geqslant m \Leftrightarrow \varphi_{\#}\mu(\{z' \in Z'\colon d'(z', \varphi(x_i)) \leqslant 2r \text{ for all } i\}) \geqslant m$$

To this end, we compute

$$\begin{aligned}
&\varphi_{\#}\mu(\{z' \in Z' \colon d'(z', \varphi(x_i)) \leqslant 2r \text{ for all } i\}) \\
&= \mu(\varphi^{-1}(\{z' \in Z' \colon d'(z', \varphi(x_i)) \leqslant 2r \text{ for all } i\}) \\
&\overset{(*)}{=} \mu(\varphi^{-1}(\varphi(\{z \in Z \colon d(z, x_i) \leqslant 2r \text{ for all } i\}))) \\
&= \mu(\{z \in Z \colon d(z, x_i) \leqslant 2r \text{ for all } i\}).
\end{aligned}$$

The last equality is due to φ being bijective; the second to last equality $(*)$ requires some elaboration: In fact, φ restricts to a bijection

$$\{z \in Z \colon d(z, x_i) \leqslant 2r \text{ for all } i\} \to \{z' \in Z' \colon d'(z', \varphi(x_i)) \leqslant 2r \text{ for all } i\}$$

because it is an isometry. Indeed,

$$\begin{aligned}
d(z, x_i) \leqslant 2r \text{ for all } i &\Rightarrow d'(\varphi(z), \varphi(x_i)) \leqslant 2r \text{ for all } i, \\
\text{and } d'(z', \varphi(x_i)) \leqslant 2r \text{ for all } i &\Rightarrow d(\varphi^{-1}(z'), x_i) \leqslant 2r \text{ for all } i.
\end{aligned}$$

Finally, since the structure maps of the filtration are (by definition) inclusions of subcomplexes, they commute with the simplicial map induced by φ. □

Example 7.1.9. Let X, Y be subsets of a common ambient metric space (Z, d) and $\mu = \mu_Y$ be the counting measure of Y. Then $\mathcal{MD}(X, \mu_Y)_{m,r} = \mathcal{D}(X, Y, R_{2r})_m$, where $R_r = \{(x, y) \colon d(x, y) \leqslant r\}$. In this way, the measure Dowker bifiltration generalises ordinary Dowker complexes endowed with the total weight filtration.

Example 7.1.10. Consider the empirical measure $\mu_X = \sum_{x \in X} \delta_x$ of a point cloud $X \subseteq \mathbb{R}^d$. The measure Dowker bifiltration $\mathcal{MD}(X, \mu_X)$, is a multineighbor complex [12] of a geometric graph with loops. That is, given $r \geqslant 0$ and $m \in \mathbb{N}$, consider the graph $G = (V, E)$ which has $V = X$ and edges $\{x, x'\}$ whenever $d(x, x') \leqslant 2r$ (note the relation to the 1-skeleton of the Čech complex!). Here, we explicitly allow and even enforce the existence of a loop at each vertex. The m-neighbor complex of this graph has σ as a simplex if its vertices have m common neighbors in G. This is equivalent to saying $\mu_X(\{x' \in X \colon d(x, x') \leqslant 2r \text{ for all } x \in \sigma\}) \geqslant m$.

For an explicit example, consider X to be the four vertices of the unit square in $\mathbb{R}^2$, the slice of its associated measure Dowker bilfiltration for $r = 0.6$ is shown in Figure 7.3. Observe that non-trivial second homology appears for $m = 1$, even though the point cloud is embedded in $\mathbb{R}^2$. In contrast, Alpha or Čech complexes of points in the plane can have non-trivial homology only up to dimension 1.

In the setting of this example, we can apply the total weight Dowker duality (Theorem 7.1.4) to obtain:

Corollary 7.1.11. *Let X be a finite metric space. We have a weak equivalence of filtrations $\mathcal{MD}(X, \mu_X)_{m,r/2} \simeq \mathcal{S}(\mathcal{D}(X, X, \{d \leqslant r\}))_m$.*

Definition 7.1.12. The latter complex is the *subdivision intrinsic Čech complex*,

$$\mathcal{SI}(X)_{m,r} := \mathcal{S}(\mathcal{D}(X, X, \{d \leqslant r\}))_m.$$

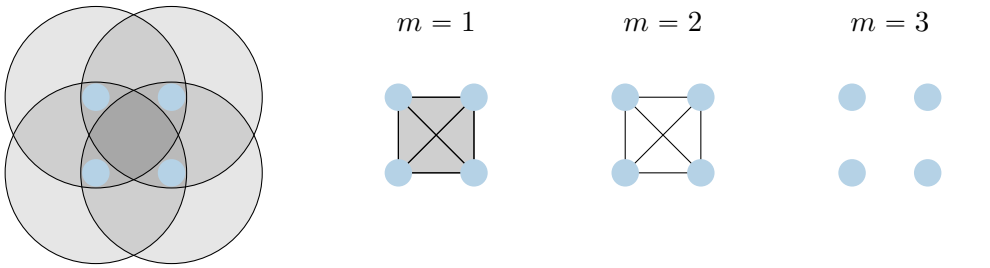


Figure 7.3: Consider X to be four points on the edges of a square in $\mathbb{R}^2$ and $\mu = \mu_X$ the associated counting measure.

We say that in the setting of Corollary 7.1.11, the measure Dowker bifiltration is a model for the subdivision intrinsic Čech bifiltration. While this complex is not easy to draw in the example from Figure 7.3 on paper, it is an instructive exercise to imagine what it will look like.

7.2 Robustness and Stability

This section is split into two parts. First, we focus on the special case of Example 7.1.10 and use the previous corollary to establish a robustness result. In the second part, we consider general metric probability spaces and show a density-sensitive stability theorem (Theorem 7.2.4) which entails a kind of law of large numbers (Theorem 7.2.9) as a corollary.

7.2.1 Counting Measure of a Finite Metric Space

In this section, we consider (X, d) to be a finite metric space and endow it with its counting measure μ_X. Recall the characterization of the intrinsic Čech complex as Dowker complex $\mathcal{I}(X)_r = \mathcal{D}(X, X, \{d \leqslant r\})$. We obtain a result similar to [27, Theorem 1.7], but with a smaller multiplicative factor. After employing Corollary 7.1.11, it suffices to see that the subdivision intrinsic Čech complex approximates the subdivision Rips complex (which is robust) in the homotopy interleaving distance. Indeed, this approximation result was obtained by Lesnick and McCabe [108], but it was not yet published at the time when this article first appeared, which is why we include it for the sake of completeness (without claiming originality). Our model for the subdivision intrinsic Čech complex has the advantage of being a bifiltration, whereas the model considered in [108] is not.

Lemma 7.2.1. *We have the following interleaving:*

$$\mathcal{I}(X)_r \subseteq \mathcal{R}(X)_{2r}, \qquad \mathcal{R}(X)_r \subseteq \mathcal{I}(X)_r.$$

Proof. On the one hand, let $\sigma = [x_0, \ldots, x_k] \in \mathcal{I}(X)_r$. Then there is some $x \in X$ such that $d(x, x_i) \leqslant r$ for all i. Thus, by the triangle inequality, $d(x_i, x_j) \leqslant 2r$ for all i, j. Hence, $\sigma \in \mathcal{R}(X)_{2r}$.

On the other hand, let $\sigma = [x_0, \dots, x_k] \in \mathcal{R}(X)_r$. Then, by definition, $d(x_i, x_j) \leqslant r$ for all i, j. Thus $\bigcap_{i=0}^{k} \overline{B}_r(x_i) \supseteq \sigma$, which is non-empty. (The closed balls are understood with respect to the finite metric space.) Thus, $\sigma \in \mathcal{I}(X)_r$. □

For $\delta > 0$, consider the forward shift

$$\begin{aligned} \alpha^\delta \colon [0,\infty[^{op} \times [0,\infty[&\to [0,\infty[^{op} \times [0,\infty[\\ (m, r) &\to (m - \delta, 2r + \delta). \end{aligned}$$

Theorem 7.2.2. *Consider two non-empty finite metric spaces endowed with their empirical probability measures $(X_1, d_1, \nu_{X_1}), (X_2, d_2, \nu_{X_2})$. Then the homology of their measure Dowker complexes*

$$H_*(\mathcal{MD}(X_1, \nu_{X_1})),\ H_*(\mathcal{MD}(X_2, \nu_{X_2})))$$

are α^δ-interleaved functors for any $\delta > d_{GPr}(\nu_{X_1}\nu_{X_2})$.

Proof. First, recall from Corollary 7.1.11 that

$$\mathcal{MD}(X_i, \nu_{X_i})_{m,r} = \mathcal{MD}(X_i, \mu_{X_i})_{|X_i|m,r} \overset{\text{Cor. 7.1.11}}{\simeq} \mathcal{S}(\mathcal{D}(X_i, X_i, \{d \leqslant 2r\})_{m|X_i|} = \mathcal{S}^n(\mathcal{I}(X_i)_{2r})_m,$$

where μ_{X_i} denotes the counting measure and the superscript $\mathcal{S}^n$ indicates the normalized subdivision filtration as in Definition 2.2.39. Taking (normalized) subdivision filtrations in Lemma 7.2.1, we get

$$\mathcal{S}^n(\mathcal{I}(X_i)_r)_m \subseteq \mathcal{S}^n(\mathcal{R}(X_i)_{2r})_m, \qquad \mathcal{S}^n(\mathcal{R}(X_i)_r)_m \subseteq \mathcal{S}^n(\mathcal{I}(X_i)_r)_m.$$

Now robustness of subdivision-Rips implies the desired result. Namely, we get the following comparisons of normalized bifiltrations:

$$\begin{aligned} &\mathcal{MD}(X_1, \nu_{X_1})_{m,r} && \\ \simeq &\mathcal{S}^n(\mathcal{I}(X_1)_{2r})_m && \text{by Corollary 7.1.11} \\ \subseteq &\mathcal{SR}^n(X_1)_{m,4r} && \text{by Lemma 7.2.1} \\ \overset{\sim}{\underset{\delta}{\longleftrightarrow}} &\mathcal{SR}^n(X_2)_{m-\delta,2(2r+\delta)} && \text{by Theorem 2.2.41} \\ \subseteq &\mathcal{S}^n(\mathcal{I}(X_2)_{2(2r+\delta)})_{m-\delta} && \text{by Lemma 7.2.1} \\ \simeq &\mathcal{MD}(X_2, \nu_{X_2})_{m-\delta,2r+\delta} && \text{by Corollary 7.1.11,} \end{aligned}$$

where $\overset{\sim}{\underset{\delta}{\longleftrightarrow}}$ denotes a δ-homotopy interleaving. For X_1 and X_2 interchanged, we obtain the analogous statement. After applying homology, these form interleaving morphisms as they are compositions of interleaving morphisms. □

For completeness, we also include the following relation with the degree Rips bifiltration:

Proposition 7.2.3. *Let X be a non-empty finite metric space with associated counting measure μ_X We have the following interleaving:*

$$\mathcal{MD}(X, \mu_X)_{m,r} \subseteq \mathcal{DR}(X, \mu_X)_{m,4r}, \qquad \mathcal{DR}(X, \mu_X)_{m,r} \subseteq \mathcal{MD}(X, \mu_X)_{m-1,r}.$$

Proof. Let $\sigma = [x_0, \dots, x_k] \in \mathcal{MD}(X, \mu_X)_{m,r}$ be a k-simplex, i.e. $|X \cap \bigcap_{x\in\sigma} \overline{B}_{2r}(x)| \geqslant m$. As $m = 0$ is excluded in Definition 7.1.6, this means the pairwise distances are bounded as $d(x_i, x_j) \leqslant 4r$. Thus in particular, every vertex has at least degree $m - 1$ in $sk^1(\mathcal{R}(X)_{4r})$; consequently, σ is a clique in there.

Vice versa, let $\sigma = [x_0, \dots, x_k] \in \mathcal{DR}(X, \mu_X)_{m,r}$ be a k-simplex. That is, each x_i has at least $m - 1$ other data points in an r-neighbourhood, k of which are in σ. Then by the triangle inequality, $\overline{B}_r(x_i) \subseteq \bigcap_{x\in\sigma} \overline{B}_{2r}(x)$. Therefore, all the data points within distance r of any x_i lie within distance $2r$ of all x_i. In particular, $|X \cap \bigcap_{x\in\sigma} \overline{B}_{2r}(x)| \geqslant m - 1$. □

While one could also apply stability results of the degree Rips bifiltration now to obtain stability of this measure Dowker bifiltration, we will provide stronger bounds in a more general setting in the next section.

7.2.2 General Metric Probability Spaces

We present a stability result about the measure Dowker bifiltration, similar in spirit to the results of [136, 142]. However, our theorem explicitly only concerns homology:

Theorem 7.2.4. *Suppose (Z, d) is a Polish space, endowed with Borel Σ-algebra $\mathfrak{B}(Z)$. Let $X_1, X_2 \in \mathfrak{B}(Z)$ and let μ_1, μ_2 be measures on $(Z, \mathfrak{B}(Z))$. Then for any $k \in \mathbb{N}$, we have*

$$d_I(H_k(\mathcal{MD}(X_1, \mu_1)), H_k(\mathcal{MD}(X_2, \mu_2))) \leqslant \max(\{d_H(X_1, X_2), d_{Pr}(\mu_1, \mu_2)\}),$$

where d_H is the Hausdorff distance (Definition 2.1.1) and d_{Pr} is the Prokhorov metric (Definition 2.1.6).

Our strategy is to take any $\delta \geqslant \max(\{d_H(X_1, X_2), d_{Pr}(\mu_1, \mu_2)\})$ and show that the persistence modules are δ-interleaved. For the proximity relation $C = \{(x, y) \colon d(x, y) \leqslant \delta\} \subseteq X_1 \times X_2$ consider the canonical projections $X_1 \xleftarrow{\pi_{X_1}} C \xrightarrow{\pi_{X_2}} X_2$. As $\delta \geqslant d_H(X_1, X_2)$, these projection maps are surjective. A relation with this feature is sometimes called a correspondence, hence the notation C here. We follow the the general line of thought of the proofs of [45], although carefully adapted to the two-parameter setting in the following lemma.

Lemma 7.2.5. *In the setting of Theorem 7.2.4, let $C = \{(x, y) \in X_1 \times X_2 \colon d(x, y) \leqslant \delta\}$. Denote the canonical projections as $X_1 \xleftarrow{\pi_{X_1}} C \xrightarrow{\pi_{X_2}} X_2$. For any subset $\sigma \subseteq X_1$, set $C(\sigma) = \pi_{X_2}(\pi_{X_1}^{-1}(\sigma)) \subseteq X_2$. Then for any simplex $\sigma \in \mathcal{MD}(X_1, \mu_1)_{m,r}$, every finite subset of $C(\sigma)$ is a simplex in $\mathcal{MD}(X_2, \mu_2)_{m-\delta,r+\delta}$.*

Proof. Let $\tau \subseteq C(\sigma)$ be finite; it being a simplex in $\mathcal{MD}(X_2, \mu_2)_{m-\delta,r+\delta}$ amounts to

$$\mu_2\left(\bigcap_{y\in\tau} \overline{B}_{2(r+\delta)}(y)\right) \geqslant m - \delta.$$

This holds true by the estimate

$$\begin{aligned} m - \delta &\leqslant \mu_1\left(\bigcap_{x\in\sigma} \overline{B}_{2r}(x)\right) - \delta \\ &\leqslant \mu_2\left(\left(\bigcap_{x\in\sigma} \overline{B}_{2r}(x)\right)^{\delta}\right) \\ &\leqslant \mu_2\left(\bigcap_{y\in\tau} \overline{B}_{2(r+\delta)}(y)\right). \end{aligned}$$

Here, the first inequality is by definition of $\sigma \in \mathcal{MD}(X_1, \mu_1)$; the second inequality is due to the definition of the Prokhorov metric and because $\delta \geqslant d_{Pr}(\mu_1, \mu_2)$; the third inequality is because

$$\left(\bigcap_{x\in\sigma} \overline{B}_{2r}(x)\right)^{\delta} \subseteq \bigcap_{y\in\tau} \overline{B}_{2(r+\delta)}(y).$$

Indeed, take $z' \in Z$ with $d(z, z') < \delta$ for some $z \in \bigcap_{x\in\sigma} \overline{B}_{2r}(x)$. Let $y \in \tau$ be arbitrary and $x \in \sigma$ such that $(x, y) \in C$. This exists because $\delta \geqslant d_H(X_1, X_2)$ and $y \in C(\sigma)$. Finally, the triangle inequality yields

$$d(y, z') \leqslant d(y, x) + d(x, z) + d(z, z') \leqslant \delta + 2r + \delta = 2(r + \delta). \qquad \square$$

Proof of Theorem 7.2.4. Let $f\colon X_1 \to X_2$ be such that $\forall x \in X_1\colon (x, f(x)) \in C$ (this exists because the canonical projections are surjective). This induces a simplicial map $\mathcal{MD}(X_1, \mu_1)_{m,r} \to \mathcal{MD}(X_2, \mu_2)_{m-\delta,r+\delta}$ for all m, r. Namely, if $\sigma = [x_0, \dots, x_k] \in \mathcal{MD}(X_1, \mu_1)_{m,r}$, then $\{f(x_0), \dots, f(x_k)\}$ is a subset of $C(\sigma)$ and hence forms a simplex in $\mathcal{MD}(X_2, \mu_2)_{m-\delta,r+\delta}$ by the preceding lemma.

These simplicial maps commute with the inclusion maps of the filtration; i.e. f induces a map of bifiltered complexes and thus in persistent homology. Now, suppose $g\colon X_1 \to X_2$ is another map with $(x, g(x)) \in C$ for all $x \in X_1$. For any simplex $\sigma = [x_0, \dots, x_l] \in \mathcal{MD}(X_1, \mu_1)_{m,r}$ the set

$$\{f(x_0), \dots, f(x_l), g(x_0), \dots, g(x_l)\}$$

is a subset of $C(\sigma)$ and thus, by Lemma 7.2.5 a simplex in $\mathcal{MD}(X_2, \mu_2)_{m-\delta,r+\delta}$. That is, the induced simplicial maps f, g are contiguous and thus induce the same map in homology (by Lemma 7.0.2). By symmetry, we obtain the diagonal maps in the following diagram, where the horizontal maps are induced by the filtration inclusions:

$$\begin{array}{ccccc} H_*(\mathcal{MD}(X_1, \mu_1)_{\bullet,\bullet}) & \longrightarrow & H_*(\mathcal{MD}(X_1, \mu_1)_{\bullet-\delta,\bullet+\delta}) & \longrightarrow & H_*(\mathcal{MD}(X_1, \mu_1)_{\bullet-2\delta,\bullet+2\delta}) \\ & \searrow\!\!\nearrow & & \searrow\!\!\nearrow & \\ H_*(\mathcal{MD}(X_2, \mu_2)_{\bullet,\bullet}) & \longrightarrow & H_*(\mathcal{MD}(X_2, \mu_2)_{\bullet-\delta,\bullet+\delta}) & \longrightarrow & H_*(\mathcal{MD}(X_2, \mu_2)_{\bullet-2\delta,\bullet+2\delta}) \end{array}$$

The commutativity is again established by a contiguity argument, following [45, Proposition 4.2]: Indeed, let $f\colon X_1 \to X_2$ as before, and correspondingly $f^\top\colon X_2 \to X_1$ such that $(f^\top(y), y) \in C$

for all $y \in X_2$. That is, $f^\top$ induces the upward-right diagonal maps in the diagram. Let again $\sigma = [x_0, \dots, x_l] \in \mathcal{MD}(X_1, \mu_1)_{m,r}$. It remains to see that the composition $H_*(f^\top) \circ H_*(f)$ and the structure map of the filtration are contiguous, i.e. that we have a simplex

$$\tau = \{f^\top(f(x_0)), \dots, f^\top(f(x_l)), x_0, \dots, x_l\} \in \mathcal{MD}(X_1, \mu_1)_{m-2\delta, r+2\delta}.$$

Now, recall $f(\sigma)$ is a simplex in $\mathcal{MD}(X_2, \mu_2)_{m-\delta, r+\delta}$. Thus, by applying Lemma 7.2.5 shifted by δ and with the roles of X_1, X_2 interchanged to $f(\sigma)$, it suffices to see that τ is a finite subset of

$$C^\top(f(\sigma)) = \{x \in X_1 : \exists x_i \in \sigma \text{ such that } (x, f(x_i)) \in C\}.$$

But this is immediate from the constructions because for all i, we have $(x_i, f(x_i)) \in C$ by definition of f and $(f^\top(f(x_i)), f(x_i)) \in C$ by definition of $f^\top$. Again, a symmetric argument establishes commutativity of the other half of the diagram. □

Remark 7.2.6. Note that the interleaving in homology does not arise from an interleaving of spaces in a straight-forward way. However, if one dropped homology from the preceding disscussion, one would get a bound for the so-called homotopy-commutative interleaving distance [25] between the two filtrations. We conjecture that this can actually be strengthened to the homotopy interleaving distance.

It is furthermore worth noting that the set $\{d(x, y) \leqslant \delta\} \subseteq X_1 \times X_2$ takes on three different roles in our discussion:

1. It is the relation defining the Dowker complex.
2. It is the correspondence inducing the interleaving maps in homology,
3. It appears in the optimal transport characterization of the Prokhorov metric: this distance is the infimal δ such no more than δ of the mass need to be transported over a distance greater than δ, i.e. outside of $\{d(x, y) \leqslant \delta\}$ (recall Definitions 2.1.5, 2.1.6).

As a direct consequence of our stability theorem, we obtain a robustness result for Dowker complexes built on a fixed set of landmarks.

Corollary 7.2.7. *Consider a fixed finite set of landmarks in some metric space, $X \subseteq (Z, d)$. Given two point clouds $Y_1, Y_2 \subseteq Z$, we have*

$$d_I(H_*(\mathcal{MD}(X, \nu_{Y_1})), H_*(\mathcal{MD}(X, \nu_{Y_2}))) \leqslant d_{Pr}(\nu_{Y_1}, \nu_{Y_2}),$$

where $\nu_Y = \frac{1}{|Y|} \sum_{y \in Y} \delta_y$ is the empirical probability measure.

Another consequence is Gromov-Hausdorff-Prokhorov stability of the measure Dowker bifiltration of metric probability spaces:

Corollary 7.2.8. *For two metric probability spaces $(X_1, \mathfrak{B}(X_1), \nu_1)$, $(X_2, \mathfrak{B}(X_2), \nu_2)$, we have*

$$d_I(H_*(\mathcal{MD}(X_1, \nu_1)), H_*(\mathcal{MD}(X_2, \nu_2))) \leqslant d_{GHPr}((X_1, \nu_1), (X_2, \nu_2)).$$

Proof. Assume we have isometric embeddings into a common Polish space $X_1 \xrightarrow{\varphi} Z \xleftarrow{\psi} X_2$. Then

$$\begin{aligned}
&d_I(H_*(\mathcal{MD}(X_1,\nu_1)), H_*(\mathcal{MD}(X_2,\nu_2)))\\
&= d_I(H_*(\mathcal{MD}(\varphi(X_1),\varphi_\#(\nu_1))), H_*(\mathcal{MD}(\psi(X_2),\psi_\#(\nu_2))))\\
&\leqslant \max\{d_H(\varphi(X_1),\psi(X_2)), d_{Pr}(\varphi_\#\nu_1,\psi_\#\nu_2)\}.
\end{aligned}$$

Here, we used Lemma 7.1.8 for the first equality and Theorem 7.2.4 for the inequality. As φ, ψ are arbitrary, we can take the infimum over all such embeddings to get the desired assertion. □

As a third consequence, we obtain a consistency result: The interlaving distance between the homology of the measure Dowker bicomplex of a finite sample and the one of the true underlying metric probability space converges to zero in probability as the sample size goes to infinity. This can be thought of as a kind of 'law of large numbers' – the complex built on the empirical point sample converges to the true underlying bifiltration (at least in homology). An analogous result for degree-Rips was previously known, see in particular [136, Lemma 107], whose proof we follow closely. Yet another result of similar type is Theorem 3.11 in [27].

Theorem 7.2.9. *Let (X,μ) be a metric probability space with compact support* $\operatorname{supp}(\mu) = A$. *Let $(x_i)_{i\in\mathbb{N}}$ be an infinite sequence of i.i.d. samples from μ. Set $X_n = \{x_1,\dots,x_n\}$ and $\nu_{X_n} = n^{-1}\sum_{i=1}^n \delta_{x_i}$ the corresponding empirical probability measure. Then for all $\varepsilon > 0$,*

$$\lim_{n\to\infty} \mathbb{P}[d_I(H_*(\mathcal{MD}(X_n,\nu_{X_n})), H_*(\mathcal{MD}(A,\mu)) > \varepsilon] = 0.$$

Proof. Let $\varepsilon > 0$. From the stability theorem (Theorem 7.2.4) we have

$$\mathbb{P}[d_I(H_*(\mathcal{MD}(X_n,\nu_{X_n})), H_*(\mathcal{MD}(A,\mu))) > \varepsilon] \leqslant \mathbb{P}[\max(d_H(X_n,A), d_{Pr}(\nu_{X_n},\mu)) > \varepsilon].$$

As $n \to \infty$, the empirical measures converge almost surely [66, Theorem 11.4.1], $\nu_{X_n} \to \mu$, and thus, also in the Prokhorov distance. Moreover, as A is compact, there are $a_1,\dots,a_N$ such that $A \subseteq B_{\varepsilon/2}(a_1) \cup \ldots \cup B_{\varepsilon/2}(a_N)$. Now each of these balls has positive mass under μ, which is almost surely approximated by the empirical measures:

$$\nu_{X_n}(B_{\varepsilon/2}(a_i)) \xrightarrow[n\to\infty]{a.s.} \mu(B_{\varepsilon/2}(a_i)) > 0.$$

Therefore, there are almost surely sample points falling into those balls, which means

$$A \subseteq B_\varepsilon(x_1) \cup \ldots \cup B_\varepsilon(x_n)$$

and consequently $\lim_{n\to\infty} \mathbb{P}[d_H(X_n,A) > \varepsilon] = 0$. Thus

$$\mathbb{P}[\max(d_H(X_n,A), d_{Pr}(\nu_{X_n},\mu)) > \varepsilon] \to 0 \text{ as } n \to \infty,$$

as desired. □

7.3 Computational Results

For the algorithmic aspects, we focus on the discrete-combinatorial version of the measure Dowker bifiltration, i.e. we assume X and Y are finite and endow Y with its counting measure μ_Y. Note that the measure Dowker bifiltration is *multi-critical*, which means that a simplex need not appear at a unique minimal bidegree in $]0,\infty[^{op}\times[0,\infty[$ but rather at a collection of mutually incomparable bidegrees. In order to compute them, we make use of the following characterization:

Lemma 7.3.1. *Let X, Y be finite sets, μ_Y the counting measure of Y and $\Lambda\colon X\times Y\to\mathbb{R}$ a function. The simplex $\sigma=[x_0,\dots,x_k]\in\mathcal{MD}(X,\mu_Y,\Lambda)$ appears in bidegrees $(m,r_m(\sigma))$ where $r_m(\sigma)$ is the m^{th} smallest value of $\left\{\max_i \frac{1}{2}\Lambda(x_i,y)\colon y\in Y\right\}$.*

Proof. The simplex appears as soon as there are m witnesses. In other words,

$$r_m(\sigma)=\min\{r>0\colon \mu_Y(\{y\colon \Lambda(x,y)\leqslant 2r \text{ for all } x\in\sigma\})\geqslant m\}.$$

Now the map $r\mapsto\mu(\{y\colon \Lambda(x,y)\leqslant r \text{ for all } x\in\sigma\})$ is monotonically increasing and starts from 0. Hence, it reaches the value m after increasing m times. □

Observe that the only way for a simplex to be critical, i.e. to have a single bidegree of appearance is if $y\mapsto\max\{\frac{1}{2}\Lambda(x,y)\colon x\in\sigma\}$ is a constant function.

We use the preceding lemma to construct a list of simplices with their appearances recursively, adapting the classical algorithm of [167] to Algorithm 7.1. Knowing the witnesses of a simplex σ, we go to a coface $\tau=\sigma\cup\{j\}$ of codimension 1. Its witnesses are given by the intersection of the witnesses of σ and those of $\{j\}$. The bidegree of appearance is computed by sorting the first entries, up to some specified m_{max}, of $\left\{\max_i \frac{1}{2}\Lambda(x_i,y)\colon y\in Y\right\}$.

Let us briefly discuss runtime and size aspects. In the worst case there is $y\in Y$ such that $X\times\{y\}\subseteq R_r$ for some r, which means that the Dowker complex will be a filtration of the complete simplex on X, which has $2^{|X|}$ simplices. Consequently, its $\dim_{max}$-skeleton has $O(|X|^{\dim_{max}})$ simplices.. For each simplex, we have to store up to m_{max} bidegrees of appearance. They are computed by sorting the first m_{max} entries of an array of size $|Y|$, which is known as the partial sorting problem and can be implemented via a combination of heap-select and heap-sort giving complexity $O(|Y|\log(m_{max}))$. This leads to a total run-time of $O(|X|^{\dim_{max}}\cdot|Y|\cdot\log(m_{max}))$ for the skeleton of the bifiltered Dowker complex. For small values of $\dim_{max}$, as one needs for low-dimensional persistent homology, we found this to be computionally tractable. The computational bottleneck in our experiments is consistently the homology computation, although we admit that the RIVET software [146] we employed for its ease of use is not state of the art in terms of speed, which is [15]. In the case of Euclidean proximity as the relation, it might be interesting to speed up the construction using a geometric data structure for storing nearest neighbors. Even more interesting would be to decrease the size of the complex in a way similar to how Alpha complexes are much smaller but equivalent to Čech. Note that the naive approach of just intersecting with the Alpha complex at scale $2r$ does indeed change the homotopy type, as can be observed in the example of Figure 7.3: When $m=1$, $\mathcal{MD}$ has non-trivial second homology and this stays true if we wiggle the points minimally to move into general position. But the Alpha complex of points in $\mathbb{R}^2$ cannot have any second homology.

Algorithm 7.1: Computing the bifiltered measure Dowker complex.

Input: A finite set X of size n with elements labelled 0 through $n-1$, a finite set Y, a matrix $\Lambda \in \mathbb{R}^{X\times Y}$, $m_{max} \in \mathbb{N}$, $\dim_{max} \in \mathbb{N}$.

Output: A list of simplices of $\mathcal{D}(X, Y, \Lambda)$ with bidegrees of appearance.

```
SimplexList ← [] /* global variable                                   */
for k = n − 1 to 0 do
    AppendUpperCofaces({k}, Λ[k]))/* Λ[k] denotes k^th row           */
return SimplexList;
Function AppendUpperCofaces(σ, WitnessValues):
    sorted ← SmallestElements(WitnessValues, m_max);
    Appearances ← {(sorted[i]/2, i) : 0 < i < m_max, sorted[i] ⩽ r_max};
    SimplexList ← SimplexList ∪ (σ, Appearances);
    if dim(σ) ⩽ dim_max then
        for j = max(σ) + 1 to n − 1 do
            τ ← σ ∪ {j};
            CommonWitnessValues ← (max{WitnessValues[i], Λ[j][i]})_{i∈{0,...,|Y|}};
            AppendUpperCofaces(τ, CommonWitnessValues);
```

Before we conclude the chapter, we present some computational results, which we hope do not just illustrate the ideas presented in this work, but also will stimulate further applications of the measure Dowker bifiltration.

Example 7.3.2. Inspired by the experiment of [27, Appendix A], we consider three point clouds in the plane, illustrated in Figure 7.4a.

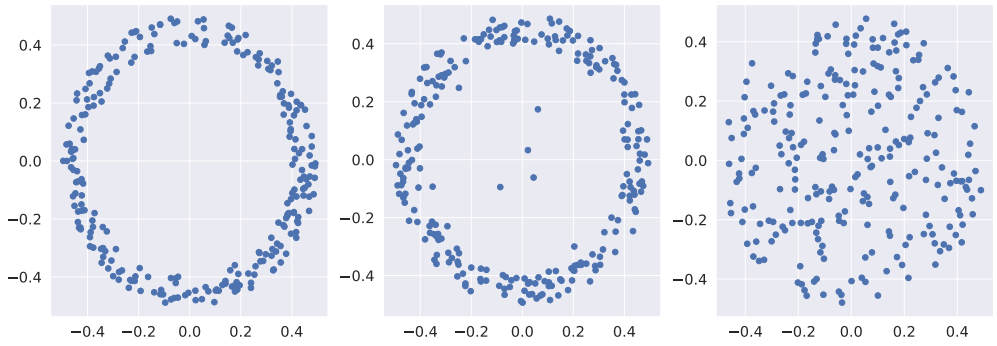
- X contains 100 points uniformly sampled from an annulus with inner radius 0.4 and outer radius 0.5,
- Y contains 94 points from the same annulus, and 6 points sampled uniformly from the disk of radius 0.4.
- Z consists of 100 points sampled uniformly from the disk of radius 0.5.

We consider the measure Dowker complex of an equispaced 10×10-grid S with respect to the counting measures from Corollary 7.2.7 (Figure 7.4b) as well the m-neighbor bifiltration from example 7.1.10 (Figure 7.4c) with m up to 50. Then we compute the Hilbert function, that is the dimensions of H_1 (recall Definition 2.3.24)

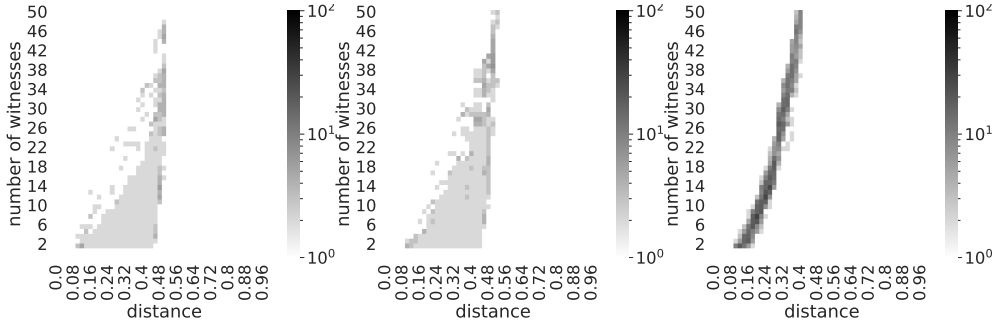
$$\mathrm{hf}^{H_1(\mathcal{MD}(X,\mu_X))} \colon]0,\infty[^{op}\times[0,\infty[\to \mathbb{N};\ (m,r)\mapsto \dim(H_1(\mathcal{MD}(X,\mu_X)_{m,r},\mathbb{Z}/2)).$$

on a 50×50 grid $G \subset]0,\infty[^{op}\times[0,\infty[$, using the RIVET software [146].

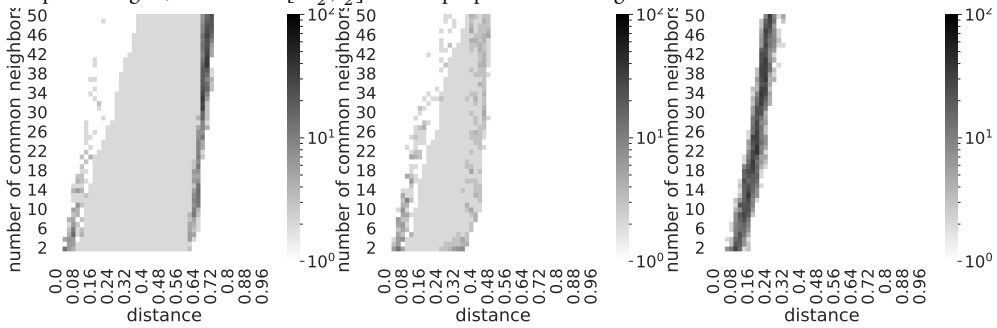
For comparison, we repeat the same computations using the degree Rips bifiltration with results displayed in Figure 7.4d, as in the original experiments of [27] and further studied in [135].



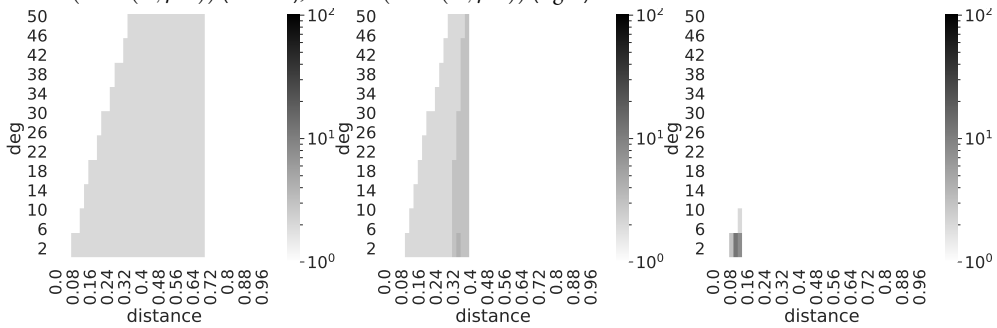
(a) Two noisy annuli X (left) and Y (middle) as well as a uniform sample Z (right) from a disk.



(b) The Hilbert functions of $H_1(\mathcal{MD}(S, \mu_X))$ (left), $H_1(\mathcal{MD}(S, \mu_Y))$ (middle), and $H_1(\mathcal{MD}(S, \mu_Z))$ (right) sampled on a grid, where $S \subseteq [-\frac{1}{2}, \frac{1}{2}]^2$ is an equispaced 10×10 grid of landmarks.



(c) The Hilbert functions of our Dowker model of subdivision intrinsic Čech, $H_1(\mathcal{MD}(X, \mu_X))$ (left), $H_1(\mathcal{MD}(Y, \mu_Y))$ (middle), and $H_1(\mathcal{MD}(Z, \mu_Z))$ (right).



(d) The Hilbert functions of the degree Rips bifiltrations, $H_1(\mathcal{DR}(X, \mu_X))$ (left), $H_1(\mathcal{DR}(Y, \mu_Y))$ (middle), and $H_1(\mathcal{DR}(Z, \mu_Z))$ (right).

Figure 7.4: The results of the computations from Example 7.3.2. Each Hilbert function is evaluated on an equispaced 50×50 grid. Note that the color-scale is logarithmic.

Example 7.3.3. Consider Λ to be a matrix with i.i.d. uniform entries from $[0, 1]$. If we fix a sublevel set of $\Lambda \leqslant p$ as the relation, we obtain an Erdös-Renyi hypergraph in the sense of [13]. We can keep track of the dimension of homology as p varies, cf. Figure 7.5. Studying vanishing thresholds for this two-parameter persistent homology is an intriguing direction for future research. A first step in this direction can be seen in [12] in the setting of m-neighbor complexes of Erdös-Renyi graphs.

Example 7.3.4. Consider the dataset [75] of gene expressions from 20531 genes of 801 patients with five different types of cancer. We regard this as 801 points in $\mathbb{R}^{20531}$; however, the Euclidean distance is not very meaningful due to the curse of dimensionality [19]. Instead, we consider the k-nearest neighbor matrix with respect to the cosine distance. Explicitly, the cosine distance between $x_1, x_2 \in \mathbb{R}^d$ is

$$d_C(x_1, x_2) = 1 - \frac{\langle x_1, x_2 \rangle}{\|x_1\| \|x_2\|}.$$

The filtration of relations

$$R_k = \{(x_1, x_2) \colon |\{x \in X \colon d_C(x_1, x) < d_C(x_1, x_2)\}| \leqslant k\}$$

is then encoded by the sublevel sets of the matrix

$$\Lambda_{ij} = k \Leftrightarrow j \text{ is the } k^{\text{th}} \text{ nearest neighbor of } i.$$

Note that both filtration parameters are on the same scale, as opposed to degree Rips for innstance, which is conceptually nice and might help to interpret the results. We compute H_0 of the bifiltered Dowker complex for the number of nearest neighbors k up to 64 and the total weight up to 64. In other words, two patients end up in the same connected component, which we interpret as a cluster, if and only if they have m common points among their respective $2k$ nearest neighbors. Of course, for $m > 2k$, there are no points at all. We can inspect the appearance and merging of clusters using the Hilbert function, shown in the left panel of Figure 7.6. Moreover, we can visualize the data based on a force-directed graph layout of the 1-skeleton of the Dowker complex for a given choice of k and m. In the right panel of Figure 7.6, this is done for $k = 30, m = 12$ at the top, which yields the true number of clusters 5. The colors represent the true label, i.e. which type of cancer the patient has. When we set $k = 60, m = 20$ in the bottom right panel, we only get three connected components, yet the cluster structure remains visible. This hints at a connection between Dowker complexes and dimensionality reduction techniques like UMAP, which builds a graph on a high dimensional point cloud by looking at the distance to the k-nearest neighbor and embeds it using a force-directed layout [118].

Example 7.3.5. Dowker and neighborhood complexes have previously been used with great success by Liu et al. for predicting protein-ligand binding affinity [113, 114], a task in computer aided drug design. We follow the setup of Liu et al. to create the complex, which is common in both referenced works, just that we have an additional filtration parameter. Given a protein-ligand pair, build a bipartite Dowker complex which has the ligand atoms of a fixed kind as vertices and use protein atoms of fixed type as witnesses. We use all possible combinations of ligand atoms from $\{C, N, O, S, P, F, Cl, Br, I\}$ and protein atoms from $\{C, N, O, S\}$. Then we proceed to compute persistent homology in dimensions 0 and 1 of each such Dowker complex, bifiltered by both distance (up to 100 ångströms) and total weight

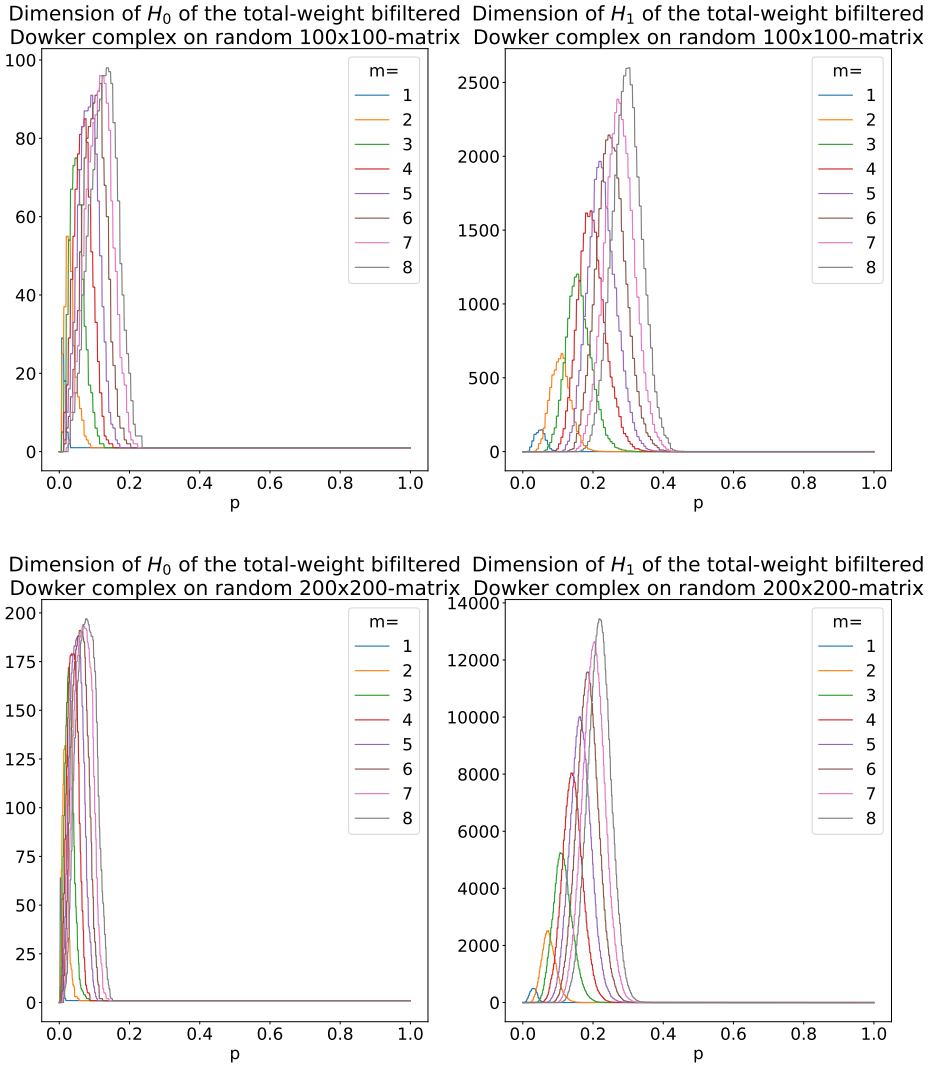


Figure 7.5: Hilbert functions of H_0 (left) and H_1 (right) for $n \times n$ matrices with random uniform entries ($n = 100$ (top), $n = 200$ (bottom)) for some small values of m.

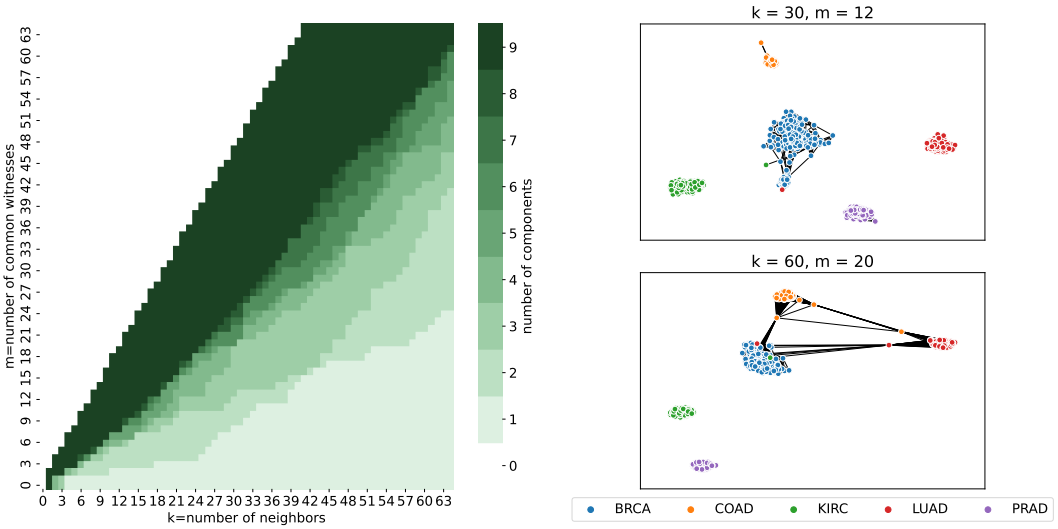


Figure 7.6: The gene expression clustering and dimensionality reduction of Example 7.3.4. We show hf^M, where $M = H_0((\mathcal{MD}(X, X, \Lambda)))$, in the left panel, and two particular choices of the 1-skeleton of $\mathcal{MD}(X, X, \Lambda)_{m,k}$ embedded using a force-directed layout on the right panel. In the left panel, colors correspond to the number of connected components, where cyan includes everything $\geqslant 9$. In the right panel, the colors encode the true label, i.e. the type of cancer as shown in the legend.

m_{max}	1	2	3	4	5	6	7	8
train R^2	0.96	0.95	0.96	0.96	0.95	0.96	0.96	0.96
test R^2	0.41	0.35	0.44	0.46	0.54	0.51	0.48	0.57

m_{max}	9	10	11	12	13	14	15	16
train R^2	0.96	0.96	0.95	0.95	0.95	0.95	0.95	0.95
test R^2	0.46	0.47	0.47	0.42	0.50	0.51	0.50	0.52

Table 7.1: Pearson R^2 of the binding affinity prediction of Example 7.3.5

(up to 16). The persistence modules are then vectorized via the Hilbert functions; we concatenate all the vectors to obtain a "topological fingerprint" of the protein-ligand pair. Note that [113, 114] use more sophisticated vectorization methods based on persistent Laplacians. However, we are mainly interested in the question how much additional information the introduction of the second (i.e. total weight) filtration carries. We train a random forest regression with the "PDBbind-refined" dataset using the library "DeepChem" accompanying the book [129]. The test data is "PDBbind-core", for which we report the prediction accuracy in table 7.1. Notably, we observe higher accuracy for two-parameter than for the one-parameter setup (which corresponds to the column m=1). However, we are not able to reproduce the even much higher scores of [113, 114], who use the persistent Laplacians and more sophisticated vectorizations, capturing more information that just the Hilbert function.

Bibliography

1. J. Abate and P. P. Valkó. "Multi-Precision Laplace Transform Inversion". *International Journal for Numerical Methods in Engineering* 60:5, 2004, pp. 979–993. DOI: 10.1002/nme.995. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/nme.995.
2. R. Abraham, J.-F. Delmas, and P. Hoscheit. "A Note on the Gromov-Hausdorff-Prokhorov Distance between (Locally) Compact Metric Measure Spaces". *Electronic Journal of Probability* 18, none 2013. ISSN: 1083-6489. DOI: 10.1214/EJP.v18-2116.
3. H. Adams, T. Emerson, M. Kirby, R. Neville, C. Peterson, P. Shipman, S. Chepushtanova, E. Hanson, F. Motta, and L. Ziegelmeier. "Persistence Images: A Stable Vector Representation of Persistent Homology". *J. Mach. Learn. Res.* 18, 2017, p. 35. ISSN: 1532-4435; 1533-7928/e.
4. R. J. Adler and J. E. Taylor. *Gaussian Inequalities*. Springer Monogr. Math. Springer, New York, NY, 2007. DOI: 10.1007/978-0-387-48116-6.
5. H. Anai, F. Chazal, M. Glisse, Y. Ike, H. Inakoshi, R. Tinarrage, and Y. Umeda. "DTM-Based Filtrations". In: *Topological Data Analysis*. Ed. by N. A. Baas, G. E. Carlsson, G. Quick, M. Szymik, and M. Thaule. Springer International Publishing, Cham, 2020, pp. 33–66. ISBN: 978-3-030-43408-3.
6. E. Anderson. "The Species Problem in Iris". *Annals of the Missouri Botanical Garden* 23:3, 1936, pp. 457–509. ISSN: 0026-6493. DOI: 10.2307/2394164.
7. J. Antoni. "Fast Computation of the Kurtogram for the Detection of Transient Faults". *Mech. Syst. Signal Process.* 21:1, 2007, pp. 108–124. DOI: 10.1016/j.ymssp.2005.12.002.
8. J. Antoni. "The Infogram: Entropic Evidence of the Signature of Repetitive Transients". *Mech. Syst. Signal Process.* 74, 2016, pp. 73–94. DOI: 10.1016/j.ymssp.2015.04.034.
9. J. Antoni and R. B. Randall. "The Spectral Kurtosis: Application to the Vibratory Surveillance and Diagnostics of Rotating Machines". *Mech. Syst. Signal Process.* 20:2, 2006, pp. 308–331. DOI: 10.1016/j.ymssp.2004.09.002.
10. E. Arias-Castro. *Principles of Statistical Analysis: Learning from Randomized Experiments*. Institute of Mathematical Statistics Textbooks. Cambridge University Press, Cambridge, 2022. DOI: 10.1017/9781108779197.
11. N. Atienza, R. Gonzalez-Díaz, and M. Soriano-Trigueros. "On the Stability of Persistent Entropy and New Summary Functions for Topological Data Analysis". *Pattern Recognition* 107, 2020, p. 107509. ISSN: 0031-3203. DOI: 10.1016/j.patcog.2020.107509.
12. E. Babson and J. Spaliński. *From Erdos-Renyi Graphs to Linial-Meshulam Complexes via the Multineighbor Construction*. 10, 2023. DOI: 10.48550/arXiv.2309.05149. arXiv: 2309.05149 [math]. preprint.

13. M. Barthelemy. "Class of Models for Random Hypergraphs". *Physical Review E* 106:6, 2022, p. 064310. DOI: 10.1103/PhysRevE.106.064310.

14. U. Bauer, M. Kerber, F. Roll, and A. Rolle. "A Unified View on the Functorial Nerve Theorem and Its Variations". *Expositiones Mathematicae*, 2023. DOI: 10.1016/j.exmath.2023.04.005. URL: https://doi.org/10.1016%2Fj.exmath.2023.04.005.

15. U. Bauer, F. Lenzen, and M. Lesnick. "Efficient Two-Parameter Persistence Computation via Cohomology". In: *DROPS-IDN/v2/Document/10.4230/LIPIcs.SoCG.2023.15*. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2023. DOI: 10.4230/LIPIcs.SoCG.2023.15.

16. U. Bauer and M. Lesnick. "Induced Matchings and the Algebraic Stability of Persistence Barcodes." *J. Comput. Geom.* 6:2, 2015, pp. 162–191. ISSN: 1920-180X/e.

17. R. Beals. *Advanced Mathematical Analysis*. Ed. by P. R. Halmos and C. C. Moore. Vol. 12. Graduate Texts in Mathematics. Springer, New York, NY, 1973. ISBN: 978-0-387-90065-0. DOI: 10.1007/978-1-4684-9886-8.

18. M. Behandish, A. M. Mirzendehdel, and S. Nelaturi. "A Classification of Topological Discrepancies in Additive Manufacturing". *Computer-Aided Design* 115, 2019, pp. 206–217. ISSN: 0010-4485. DOI: 10.1016/j.cad.2019.05.032.

19. R. Bellman. *Dynamic Programming*. Princeton University Press, 1957. ISBN: 069107951X.

20. T. Berry, J. R. Cressman, Z. Gregurić-Ferenček, and T. Sauer. "Time-Scale Separation from Diffusion-Mapped Delay Coordinates". *SIAM Journal on Applied Dynamical Systems* 12:2, 2013, pp. 618–649. DOI: 10.1137/12088183X. eprint: https://doi.org/10.1137/12088183X. URL: https://doi.org/10.1137/12088183X.

21. P. Billingsley. *Convergence of Probability Measures*. 2nd ed. Wiley Series in Probability and Statistics. Wiley, 1999. ISBN: 978-0-471-19745-4. DOI: 10.1002/9780470316962.

22. H. B. Bjerkevik, M. B. Botnan, and M. Kerber. "Computing the Interleaving Distance Is NP-Hard". *Found Comput Math* 20:5, 1, 2020, pp. 1237–1271. ISSN: 1615-3383. DOI: 10.1007/s10208-019-09442-y. URL: https://doi.org/10.1007/s10208-019-09442-y.

23. A. Björner. "Topological Methods". In: *Handbook of Combinatorics (Vol. 2)*. MIT Press, Cambridge, MA, USA, 1996, pp. 1819–1872. ISBN: 978-0-262-07171-0.

24. N. Blaser, M. Brun, O. H. Gardaa, and L. M. Salbu. *Core Bifiltration*. 2, 2024. arXiv: 2405.01214 [cs, math]. preprint.

25. A. Blumberg and M. Lesnick. "Universality of the Homotopy Interleaving Distance". *Trans. Amer. Math. Soc.* 376:12, 2023, pp. 8269–8307. ISSN: 0002-9947, 1088-6850. DOI: 10.1090/tran/8738. URL: https://www.ams.org/tran/2023-376-12/S0002-9947-2023-08738-4/.

26. A. J. Blumberg, I. Gal, M. A. Mandell, and M. Pancia. "Robust Statistics, Hypothesis Testing, and Confidence Intervals for Persistent Homology on Metric Measure Spaces". *Foundations of Computational Mathematics* 14:4, 2014, pp. 745–789. DOI: 10.1007/s10208-014-9201-4.

27. A. J. Blumberg and M. Lesnick. "Stability of 2-Parameter Persistent Homology". *Found Comput Math*, 2022. ISSN: 1615-3383. DOI: 10.1007/s10208-022-09576-6. URL: https://doi.org/10.1007/s10208-022-09576-6.

28. O. Bobrowski. "Algebraic Topology of Random Fields and Complexes". PhD thesis. Haifa: Technion, 2012.

29. O. Bobrowski and R. J. Adler. "Distance Functions, Critical Points, and the Topology of Random Čech Complexes". *Homology, Homotopy and Applications* 16:2, 2014, pp. 311–344. ISSN: 15320073, 15320081. DOI: 20141124094211.

30. O. Bobrowski and M. Kahle. "Topology of Random Geometric Complexes: A Survey". *J Appl. and Comput. Topology* 1:3-4, 2018, pp. 331–364. ISSN: 2367-1726, 2367-1734. DOI: 10.1007/s41468-017-0010-0.

31. O. Bobrowski and S. Mukherjee. "The Topology of Probability Distributions on Manifolds". *Probability Theory and Related Fields* 161, 2013. DOI: 10.1007/s00440-014-0556-x.

32. O. Bobrowski and S. Weinberger. "On the Vanishing of Homology in Random Čech Complexes". *Random Structures & Algorithms* 51:1, 2017, pp. 14–51. DOI: 10.1002/rsa.20697. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/rsa.20697.

33. M. Botnan and M. Lesnick. "An introduction to multiparameter persistence". English. In: *Representations of Algebras and Related Structures*. 2023, pp. 77–150. DOI: https://doi.org/10.4171/ecr/19/4.

34. G. E. Bredon. *Topology and Geometry*. Vol. 139. Graduate Texts in Mathematics. Springer, New York, NY, 1993. ISBN: 978-1-4419-3103-0. DOI: 10.1007/978-1-4757-6848-0. URL: http://link.springer.com/10.1007/978-1-4757-6848-0.

35. H. Broer and F. Takens. "Reconstruction and Time Series Analysis". In: *Dynamical Systems and Chaos*. Springer New York, New York, NY, 2011, pp. 205–242. ISBN: 978-1-4419-6870-8. DOI: 10.1007/978-1-4419-6870-8_6. URL: https://doi.org/10.1007/978-1-4419-6870-8_6.

36. M. Brun, B. García Pascual, and L. M. Salbu. "Determining Homology of an Unknown Space from a Sample". *European Journal of Mathematics* 9:4, 2023, p. 90. ISSN: 2199-675X, 2199-6768. DOI: 10.1007/s40879-023-00683-4.

37. M. Brun and L. M. Salbu. "The Rectangle Complex of a Relation". *Mediterranean Journal of Mathematics* 20:1, 2023, p. 7. ISSN: 1660-5446, 1660-5454. DOI: 10.1007/s00009-022-02213-0.

38. P. Bubenik and T. Vergili. "Topological Spaces of Persistence Modules and Their Properties". *J Appl. and Comput. Topology* 2:3-4, 2018, pp. 233–269. ISSN: 2367-1726, 2367-1734. DOI: 10.1007/s41468-018-0022-4. arXiv: 1802.08117.

39. M. Carrière. *3D Shape Segmentation Using TDA*. IPython notebook. URL: https://github.com/MathieuCarriere/sklearn-tda/tree/master/example/3DSeg.

40. G. Casella and R. L. Berger. *Statistical Inference*. Duxbury Press, Pacific Grove, CA, 2002.

41. N. J. Cavanna, K. P. Gardner, and D. R. Sheehy. "When and Why the Topological Coverage Criterion Works". In: *Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms*. Society for Industrial and Applied Mathematics, 2017, pp. 2679–2690. ISBN: 978-1-61197-478-2. DOI: 10.1137/1.9781611974782.177.

42. W. Chachólski and H. Riihimaki. "Metrics and Stabilization in One Parameter Persistence". *SIAM Journal on Applied Algebra and Geometry* 4:1, 2020, pp. 69–98. DOI: 10.1137/19M1243932.

43. F. Chazal, D. Cohen-Steiner, M. Glisse, L. J. Guibas, and S. Y. Oudot. "Proximity of Persistence Modules and Their Diagrams". In: *Proceedings of the 25th Annual Symposium on Computational Geometry - SCG '09*. ACM Press, Aarhus, Denmark, 2009, p. 237. ISBN: 978-1-60558-501-7. DOI: 10.1145/1542362.1542407.

44. F. Chazal, D. Cohen-Steiner, L. J. Guibas, F. Mémoli, and S. Y. Oudot. "Gromov-Hausdorff Stable Signatures for Shapes Using Persistence". In: *Computer Graphics Forum*. Vol. 28. 5. Wiley Online Library, 2009, pp. 1393–1403.

45. F. Chazal, V. de Silva, and S. Oudot. "Persistence Stability for Geometric Complexes". *Geom Dedicata* 173:1, 2014, pp. 193–214. ISSN: 1572-9168. DOI: 10.1007/s10711-013-9937-z.

46. X. Chen, A. Golovinskiy, and T. Funkhouser. "A Benchmark for 3D Mesh Segmentation". *ACM Transactions on Graphics (Proc. SIGGRAPH)* 28:3, 2009. DOI: 10.1145/1531326.1531379.

47. Y. Cheng, Z. Wang, and W. Zhang. "Combined Square Envelope Spectrum by Integrating Multiband Bearing Fault Information". *IEEE Sens. J.* 23:3, 2023, pp. 2495–2506. DOI: 10.1109/JSEN.2022.3227244.

48. S. N. Chiu and K. I. Liu. "Generalized Cramér–von Mises Goodness-of-Fit Tests for Multivariate Distributions". *Computational Statistics & Data Analysis* 53:11, 2009, pp. 3817–3834. ISSN: 0167-9473. DOI: 10.1016/j.csda.2009.04.004. URL: https://www.sciencedirect.com/science/article/pii/S016794730900139X.

49. S. N. Chiu, D. Stoyan, W. S. Kendall, and J. Mecke. *Stochastic Geometry and Its Applications*. John Wiley & Sons, 2013. ISBN: 978-1-118-65825-3.

50. S. Chowdhury and F. Mémoli. "A Functorial Dowker Theorem and Persistent Homology of Asymmetric Networks". *J Appl. and Comput. Topology* 2:1, 2018, pp. 115–175. ISSN: 2367-1734. DOI: 10.1007/s41468-018-0020-6. URL: https://doi.org/10.1007/s41468-018-0020-6.

51. Y.-M. Chung and A. Lawson. "Persistence Curves: A Canonical Framework for Summarizing Persistence Diagrams". *Adv Comput Math* 48:1, 2022, p. 6. ISSN: 1572-9044. DOI: 10.1007/s10444-021-09893-4.

52. D. Cohen-Steiner, H. Edelsbrunner, and J. Harer. "Stability of Persistence Diagrams". *Discrete and Computational Geometry* 37:1, 2007, pp. 103–120. ISSN: 1432-0444. DOI: 10.1007/s00454-006-1276-5.

53. R. R. Coifman, S. Lafon, A. B. Lee, M. Maggioni, B. Nadler, F. Warner, and S. W. Zucker. "Geometric Diffusions as a Tool for Harmonic Analysis and Structure Definition of Data: Diffusion Maps". *Proc Natl Acad Sci U S A* 102:21, 2005, pp. 7426–7431. ISSN: 0027-8424. DOI: 10.1073/pnas.0500334102. pmid: 15899970.

54. R. Corbet, M. Kerber, M. Lesnick, and G. Osang. "Computing the Multicover Bifiltration". *Discrete & Computational Geometry*, 2023. ISSN: 1432-0444. DOI: 10.1007/s00454-022-00476-8.

55. W. Crawley-Boevey. "Decomposition of Pointwise Finite-Dimensional Persistence Modules." *J. Algebra Appl.* 14:5, 2015, p. 8. ISSN: 0219-4988; 1793-6829/e. DOI: 10.1142/S0219498815500668.

56. N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods.* Cambridge University Press, 2000. ISBN: 978-0-511-80138-9.

57. R. B. D'Agostino and M. A. Stephens, eds. *Goodness-of-Fit Techniques*. Chapman & Hall/CRC, Boca Raton, 1986.

58. V. De Silva and R. Ghrist. "Coverage in Sensor Networks via Persistent Homology". *Algebr. Geom. Topol.* 7:1, 25, 2007, pp. 339–358. ISSN: 1472-2739, 1472-2747. DOI: 10.2140/agt.2007.7.339. URL: http://www.msp.org/agt/2007/7-1/p16.xhtml.

59. S. S. Dhar, B. Chakraborty, and P. Chaudhuri. "Comparison of Multivariate Distributions Using Quantile-Quantile Plots and Related Tests". *Bernoulli* 20:3, 2014, pp. 1484–1506. ISSN: 1436-3240. DOI: 10.3150/13-BEJ530. JSTOR: i40112044. URL: https://www.jstor.org/stable/i40112044.

60. P. Dlotko, M. Carrière, and M. Royer. "Persistence Representations". In: *GUDHI User and Reference Manual*. 3.4.1. GUDHI Editorial Board, 2021.

61. P. Dłotko and D. Gurnari. "Euler Characteristic Curves and Profiles: A Stable Shape Invariant for Big Data Problems". *GigaScience* 12, 2023, giad094. ISSN: 2047-217X. DOI: 10.1093/gigascience/giad094. URL: https://doi.org/10.1093/gigascience/giad094.

62. P. Dłotko and N. Hellmer. "Bottleneck Profiles and Discrete Prokhorov Metrics for Persistence Diagrams". *Discrete Comput Geom*, 2023. ISSN: 1432-0444. DOI: 10.1007/s00454-023-00498-w. URL: https://doi.org/10.1007/s00454-023-00498-w.

63. P. Dłotko, N. Hellmer, Ł. Stettner, and R. Topolnicki. "Topology-Driven Goodness-of-Fit Tests in Arbitrary Dimensions". *Stat Comput* 34:1, 2023, p. 34. ISSN: 1573-1375. DOI: 10.1007/s11222-023-10333-0.

64. R. V. Donner, Y. Zou, J. F. Donges, N. Marwan, and J. Kurths. "Recurrence Networks—a Novel Paradigm for Nonlinear Time Series Analysis". *New Journal of Physics* 12:3, 2010, p. 033025. DOI: 10.1088/1367-2630/12/3/033025. URL: https://dx.doi.org/10.1088/1367-2630/12/3/033025.

65. C. H. Dowker. "Homology Groups of Relations". *Annals of Mathematics* 56:1, 1952, pp. 84–95. ISSN: 0003-486X. DOI: 10.2307/1969768. JSTOR: 1969768. URL: https://www.jstor.org/stable/1969768.

66. R. M. Dudley. *Real Analysis and Probability*. 2nd ed. Cambridge Studies in Advanced Mathematics. Cambridge University Press, Cambridge, 2002. ISBN: 978-0-521-80972-6. DOI: 10.1017/CBO9780511755347.

67. J.-P. Eckmann, S. O. Kamphorst, and D. Ruelle. "Recurrence Plots of Dynamical Systems". *Europhysics Letters* 4:9, 1987, p. 973. DOI: 10.1209/0295-5075/4/9/004. URL: https://dx.doi.org/10.1209/0295-5075/4/9/004.

68. H. Edelsbrunner. "The Union of Balls and Its Dual Shape". *Discrete Comput Geom* 13:3, 1, 1995, pp. 415–440. ISSN: 1432-0444. DOI: 10.1007/BF02574053. URL: https://doi.org/10.1007/BF02574053.

69. H. Edelsbrunner and J. Harer. *Computational Topology - an Introduction.* American Mathematical Society Providence, RI, Providence, RI, 2010, pp. xii + 241. ISBN: 978-0-8218-4925-5.

70. H. Edelsbrunner, D. Letscher, and A. Zomorodian. "Topological Persistence and Simplification". *Discrete Comput Geom* 28:4, 2002, pp. 511–533. ISSN: 0179-5376, 1432-0444. DOI: 10.1007/s00454-002-2885-2. URL: http://link.springer.com/10.1007/s00454-002-2885-2.

71. H. Edelsbrunner and G. Osang. "The Multi-Cover Persistence of Euclidean Balls". *Discrete & Computational Geometry* 65:4, 2021, pp. 1296–1313. ISSN: 0179-5376, 1432-0444. DOI: 10.1007/s00454-021-00281-9.

72. A. Efrat, A. Itai, and M. J. Katz. "Geometry Helps in Bottleneck Matching and Related Problems." *Algorithmica* 31:1, 2001, pp. 1–28. ISSN: 0178-4617; 1432-0541/e. DOI: 10.1007/s00453-001-0016-8.

73. G. Fasano and A. Franceschini. "A Multidimensional Version of the Kolmogorov–Smirnov Test". *Monthly Notices of the Royal Astronomical Society* 225:1, 1987, pp. 155–170. ISSN: 0035-8711. DOI: 10.1093/mnras/225.1.155.

74. X. Fernández, E. Borghini, G. Mindlin, and P. Groisman. "Intrinsic Persistent Homology via Density-Based Metric Learning". *Journal of Machine Learning Research* 24:75, 2023, pp. 1–42. URL: http://jmlr.org/papers/v24/21-1044.html.

75. S. Fiorini. *Gene Expression Cancer RNA-Seq*. UCI Machine Learning Repository. 2016. DOI: https://doi.org/10.24432/C5R88H. URL: https://archive.ics.uci.edu/dataset/401/gene+expression+cancer+rna+seq.

76. R. A. Fisher. "The Use of Multiple Measurements in Taxonomic Problems". *Annals of Eugenics* 7:2, 1936, pp. 179–188. ISSN: 2050-1439. DOI: 10.1111/j.1469-1809.1936.tb02137.x.

77. T. Fleckenstein and N. Hellmer. *When Do Two Distributions Yield the Same Expected Euler Characteristic Curve in the Thermodynamic Limit?* 2024. arXiv: 2401.04580 [math.PR].

78. J.-M. Floch, E. Marcon, and F. Puech. "Spatial Distribution of Points". In: ed. by V. Loonis and M.-P. de Bellefon. Insee-Eurostat, 2018, pp. 71–111. URL: https://www.insee.fr/en/information/3635545.

79. A. Freund, M. Andreatta, and J.-L. Giavitto. "Lattice-Based and Topological Representations of Binary Relations with an Application to Music". *Annals of Mathematics and Artificial Intelligence* 73:3, 2015, pp. 311–334. ISSN: 1573-7470. DOI: 10.1007/s10472-014-9445-3.

80. M. Gao, G. Yu, and T. Wang. "Impulsive Gear Fault Diagnosis Using Adaptive Morlet Wavelet Filter Based on Alpha-Stable Distribution and Kurtogram". *IEEE Access* 7, 2019, pp. 72283–72296.

81. J. Garland, E. Bradley, and J. D. Meiss. "Exploring the Topology of Dynamical Reconstructions". *Physica D: Nonlinear Phenomena*. Topology in Dynamics, Differential Equations, and Data 334, 2016, pp. 49–59. ISSN: 0167-2789. DOI: 10.1016/j.physd.2016.03.006.

82. A. L. Gibbs and F. E. Su. "On Choosing and Bounding Probability Metrics". *International statistical review* 70:3, 2002, pp. 419–435.

83. F. Godi. "Bottleneck Distance". In: *GUDHI User and Reference Manual*. 3.4.1. GUDHI Editorial Board, 2021.

84. A. Goel, K. D. Trinh, and K. Tsunoda. "Strong Law of Large Numbers for Betti Numbers in the Thermodynamic Regime". *J Stat Phys* 174:4, 2019, pp. 865–892. ISSN: 1572-9613. DOI: 10.1007/s10955-018-2201-z.

85. R. Gonzalez and P. Wintz. "Digital Image Processing". Applied Mathematics and Computation 13, 1977.

86. A. Greven, P. Pfaffelhuber, and A. Winter. "Convergence in Distribution of Random Metric Measure Spaces (Λ-coalescent Measure Trees)". *Probability Theory and Related Fields* 145:1-2, 2009, pp. 285–322. ISSN: 0178-8051, 1432-2064. DOI: 10.1007/s00440-008-0169-3.

87. J. Guo, Q. He, D. Zhen, and F. Gu. "Intelligent Fault Detection for Rotating Machinery Using Cyclic Morphological Modulation Spectrum and Hierarchical Teager Permutation Entropy". *IEEE Trans. Ind. Informat.*, 2022, pp. 1–10. DOI: 10.1109/TII.2022.3185293.

88. J. Hebda-Sobkowicz, R. Zimroz, M. Pitera, and A. Wyłomańska. "Informative Frequency Band Selection in the Presence of Non-Gaussian Noise–a Novel Approach Based on the Conditional Variance Statistic with Application to Bearing Fault Diagnosis". *Mech. Syst. Signal Process.* 145, 2020, p. 106971. DOI: 10.1016/j.ymssp.2020.106971.

89. J. Hebda-Sobkowicz, R. Zimroz, A. Wyłomańska, and J. Antoni. "Infogram Performance Analysis and Its Enhancement for Bearings Diagnostics in Presence of Non-Gaussian Noise". *Mech. Syst. Signal Process.* 170, 2022, p. 108764. DOI: 10.1016/j.ymssp.2021.108764.

90. R. Hegger and H. Kantz. "Improved False Nearest Neighbor Method to Detect Determinism in Time Series Data". *Phys. Rev. E* 60:4, 1999, pp. 4970–4973. DOI: 10.1103/PhysRevE.60.4970. URL: https://link.aps.org/doi/10.1103/PhysRevE.60.4970.

91. N. Hellmer and J. Spaliński. *Density Sensitive Bifiltered Dowker Complexes via Total Weight*. 2024. arXiv: 2405.15592 [math].

92. J. E. Hopcroft and R. M. Karp. "An $n^{5/2}$ Algorithm for Maximum Matchings in Bipartite Graphs". *SIAM J. Comput.* 2:4, 1, 1973, pp. 225–231. ISSN: 0097-5397. DOI: 10.1137/0202019.

93. B. Hou, D. Wang, T. Yan, Y. Wang, Z. Peng, and K.-L. Tsui. "Gini Indices II and III: Two New Sparsity Measures and Their Applications to Machine Condition Monitoring". *IEEE ASME Trans Mechatron* 27:3, 2021, pp. 1211–1222.

94. G. James, D. Witten, T. Hastie, and R. Tibshirani. *An Introduction to Statistical Learning*. Vol. 103. Springer Texts in Statistics. Springer New York, New York, NY, 2013. ISBN: 978-1-4614-7137-0. DOI: 10.1007/978-1-4614-7138-7. URL: http://link.springer.com/10.1007/978-1-4614-7138-7.

95. A. Justel, D. Peña, and R. Zamar. "A Multivariate Kolmogorov-Smirnov Test of Goodness of Fit". *Statistics & Probability Letters* 35:3, 1997, pp. 251–259. ISSN: 0167-7152. DOI: 10.1016/s0167-7152(97)00020-5. URL: https://www.sciencedirect.com/science/article/pii/S0167715297000205.

96. M. Kahle. "Random Geometric Complexes". *Discrete & Computational Geometry* 45:3, 2011, pp. 553–573. ISSN: 0179-5376, 1432-0444. DOI: 10.1007/s00454-010-9319-3. arXiv: 0910.1649.

97. H. Kantz and T. Schreiber. *Nonlinear Time Series Analysis*. 2nd ed. Cambridge University Press, 2003. DOI: 10.1017/CBO9780511755798.

98. A. Karan and A. Kaygun. "Time Series Classification via Topological Data Analysis". *Expert Syst. Appl.* 183:C, 2021. ISSN: 0957-4174. DOI: 10.1016/j.eswa.2021.115326. URL: https://doi.org/10.1016/j.eswa.2021.115326.

99. M. B. Kennel, R. Brown, and H. D. I. Abarbanel. "Determining Embedding Dimension for Phase-Space Reconstruction Using a Geometrical Construction". *Phys. Rev. A* 45:6, 1992, pp. 3403–3411. DOI: 10.1103/PhysRevA.45.3403. URL: https://link.aps.org/doi/10.1103/PhysRevA.45.3403.

100. M. Kerber, D. Morozov, and A. Nigmetov. "Geometry Helps to Compare Persistence Diagrams." *ACM J. Exp. Algorithm.* 22, 2017, p. 20. ISSN: 1084-6654/e. DOI: 10.1145/3064175.

101. F. A. Khasawneh and E. Munch. "Chatter Detection in Turning Using Persistent Homology". *Mechanical Systems and Signal Processing* 70, 2016, pp. 527–541. DOI: 10.1016/j.ymssp.2015.09.046.

102. F. A. Khasawneh, E. Munch, and J. A. Perea. "Chatter Classification in Turning Using Machine Learning and Topological Data Analysis". *IFAC-PapersOnLine* 51:14, 2018, pp. 195–200. DOI: 10.1016/j.ifacol.2018.07.222.

103. S. Ko and D. Koo. "A Novel Approach for Wafer Defect Pattern Classification Based on Topological Data Analysis". *Expert Systems with Applications* 231, 2023, p. 120765. ISSN: 0957-4174. DOI: 10.1016/j.eswa.2023.120765.

104. J. T. N. Krebs, B. Roycraft, and W. Polonik. "On Approximation Theorems for the Euler Characteristic with Applications to the Bootstrap". *Electronic Journal of Statistics* 15:2, 2021, pp. 4462–4509. ISSN: 1935-7524, 1935-7524. DOI: 10.1214/21-EJS1898.

105. T. Lacombe, M. Cuturi, and S. Oudot. *Large Scale Computation of Means and Clusters for Persistence Diagrams Using Optimal Transport*. 2018. arXiv: 1805.08331 [cs, stat].

106. M. Ledoux. *The Concentration of Measure Phenomenon*. Vol. 89. Mathematical Surveys and Monographs. American Mathematical Society, Providence, Rhode Island, 2005. DOI: 10.1090/surv/089.

107. M. Lesnick. *Notes on Multiparameter Persistence (for AMAT 840)*. 2023. URL: https://www.albany.edu/~ML644186/840_2022/Math840_Notes_22.pdf.

108. M. Lesnick and K. McCabe. *Nerve Models of Subdivision Bifiltrations*. 2024. arXiv: 2406.07679 [cs, math].

109. M. Lesnick and M. Wright. *Interactive Visualization of 2-D Persistence Modules*. 2015. arXiv: 1512.00180 [cs, math].

110. H. Li, T. Liu, X. Wu, and Q. Chen. "A Bearing Fault Diagnosis Method Based on Enhanced Singular Value Decomposition". *IEEE Trans. Ind. Informat.* 17:5, 2021, pp. 3220–3230. DOI: 10.1109/TII.2020.3001376.

111. Y. Li, J. Zhou, H. Li, G. Meng, and J. Bian. "A Fast and Adaptive Empirical Mode Decomposition Method and Its Application in Rolling Bearing Fault Diagnosis". *IEEE Sens. J.* 23:1, 2023, pp. 567–576. DOI: 10.1109/JSEN.2022.3223980.

112. D. Liu, L. Cui, and W. Cheng. "Flexible Generalized Demodulation for Intelligent Bearing Fault Diagnosis under Nonstationary Conditions". *IEEE Trans. Ind. Informat.*, 2022, pp. 1–12. DOI: 10.1109/TII.2022.3192597.

113. X. Liu, H. Feng, J. Wu, and K. Xia. "Dowker Complex Based Machine Learning (DCML) Models for Protein-Ligand Binding Affinity Prediction". *PLOS Computational Biology* 18:4, 2022, e1009943. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1009943.

114. X. Liu and K. Xia. "Neighborhood Complex Based Machine Learning (NCML) Models for Drug Design". In: *Interpretability of Machine Intelligence in Medical Image Computing, and Topological Data Analysis and Its Applications for Medical Data*. Ed. by M. Reyes, P. Henriques Abreu, J. Cardoso, M. Hajij, G. Zamzmi, P. Rahul, and L. Thakur. Lecture Notes in Computer Science. Springer International Publishing, Cham, 2021, pp. 87–97. ISBN: 978-3-030-87444-5. DOI: 10.1007/978-3-030-87444-5_9.

115. L. Lovász. "Kneser's Conjecture, Chromatic Number, and Homotopy". *Journal of Combinatorial Theory, Series A* 25:3, 1978, pp. 319–324. ISSN: 0097-3165. DOI: 10.1016/0097-3165(78)90022-5.

116. A. Maqsood, D. Oslebo, K. Corzine, L. Parsa, and Y. Ma. "STFT Cluster Analysis for DC Pulsed Load Monitoring and Fault Detection on Naval Shipboard Power Systems". *IEEE Trans. Transp. Electrification* 6:2, 2020, pp. 821–831. DOI: 10.1109/TTE.2020.2981880.

117. C. Maria, J.-D. Boissonnat, M. Glisse, and M. Yvinec. "The Gudhi Library: Simplicial Complexes and Persistent Homology". In: *Mathematical Software – ICMS 2014*. Ed. by H. Hong and C. Yap. Lecture Notes in Computer Science. Springer Berlin, Heidelberg, 2014, pp. 167–174. ISBN: 978-3-662-44199-2. DOI: 10.1007/978-3-662-44199-2_28.

118. L. McInnes, J. Healy, and J. Melville. *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*. arXiv, 2020. DOI: 10.48550/arXiv.1802.03426. arXiv: 1802.03426 [cs, stat].

119. Y. Mileyko, S. Mukherjee, and J. Harer. "Probability Measures on the Space of Persistence Diagrams". *Inverse Probl.* 27:12, 2011, p. 22. ISSN: 0266-5611. DOI: 10.1088/0266-5611/27/12/124007.

120. D. W. Muller and G. Sawitzki. "Excess Mass Estimates and Tests for Multimodality". *Journal of the American Statistical Association* 86:415, 1991, pp. 738–746. ISSN: 01621459.

121. J. R. Munkres. *Topology*. 2nd ed. Upper Saddle River, NJ: Prentice Hall, 2000. ISBN: 0-13-181629-2.

122. J. A. Peacock. "Two-Dimensional Goodness-of-Fit Testing in Astronomy". *Monthly Notices of the Royal Astronomical Society* 202:3, 1983, pp. 615–627. ISSN: 0035-8711. DOI: 10.1093/mnras/202.3.615.

123. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. "Scikit-Learn: Machine Learning in Python". *Journal of Machine Learning Research* 12, 2011, pp. 2825–2830.

124. M. Penrose. *Random Geometric Graphs*. Oxford Studies in Probability. Oxford University Press, Oxford, 2003. DOI: 10.1093/acprof:oso/9780198506263.001.0001.

125. J. A. Perea and J. Harer. "Sliding Windows and Persistence: An Application of Topological Methods to Signal Analysis". *Found Comput Math* 15:3, 2015, pp. 799–838. ISSN: 1615-3383. DOI: 10.1007/s10208-014-9206-z. URL: https://doi.org/10.1007/s10208-014-9206-z.

126. W. Polonik. "Measuring Mass Concentrations and Estimating Density Contour Clusters-an Excess Mass Approach". *The Annals of Statistics* 23:3, 1995, pp. 855–881. ISSN: 00905364. DOI: 10.1214/aos/1176324626.

127. C. Puritz, E. Ness-Cohn, and R. Braun. *Fasano.Franceschini.Test: An Implementation of a Multidimensional KS Test in R*. 2022. arXiv: 2106.10539 [stat.ME].

128. S. T. Rachev. *Probability Metrics and the Stability of Stochastic Models*. Chichester etc.: John Wiley & Sons Ltd., 1991, pp. xiv + 494. ISBN: 0-471-92877-1.

129. B. Ramsundar, P. Eastman, P. Walters, and V. Pande. *Deep Learning for the Life Sciences: Applying Deep Learning to Genomics, Microscopy, Drug Discovery, and More*. "O'Reilly Media, Inc.", 2019. ISBN: 978-1-4920-3980-8.

130. E. Richardson and M. Werman. "Efficient Classification Using the Euler Characteristic". *Pattern Recognition Letters* 49, 2014, pp. 99–106. ISSN: 0167-8655. DOI: 10.1016/j.patrec.2014.07.001.

131. V. Robins. "Betti Number Signatures of Homogeneous Poisson Point Processes". *Phys. Rev. E* 74:6, 2006, p. 061107. ISSN: 1539-3755, 1550-2376. DOI: 10.1103/PhysRevE.74.061107. URL: https://link.aps.org/doi/10.1103/PhysRevE.74.061107.

132. V. Robins. "Computational Topology at Multiple Resolutions: Foundations and Applications to Fractals and Dynamics". PhD thesis. Boulder: University of Colorado, 2000.

133. V. Robins. "Towards Computing Homology from Finite Approximations". In: *Topology Proceedings*. Vol. 24. 1. 1999, pp. 503–532.

134. M. Robinson. "Cosheaf Representations of Relations and Dowker Complexes". *J Appl. and Comput. Topology* 6:1, 2022, pp. 27–63. ISSN: 2367-1726, 2367-1734. DOI: 10.1007/s41468-021-00078-y.

135. A. Rolle. "The Degree-Rips Complexes of an Annulus with Outliers", 2022, 14 pages, 886629 bytes. ISSN: 1868-8969. DOI: 10.4230/LIPICS.SOCG.2022.58.

136. A. Rolle and L. Scoccola. *Stable and Consistent Density-Based Clustering via Multiparameter Persistence*. 2023. arXiv: 2005.09048 [math.ST].

137. M. Rucco, E. Concettoni, C. Cristalli, A. Ferrante, and E. Merelli. "Topological Classification of Small DC Motors". In: *2015 IEEE 1st International Forum on Research and Technologies for Society and Industry Leveraging a Better Tomorrow (RTSI)*. 2015, pp. 192–197. DOI: 10.1109/RTSI.2015.7325097.

138. G. Samoradnitsky. *Stable Non-Gaussian Random Processes: Stochastic Models with Infinite Variance*. Routledge, New York, 1994. ISBN: 978-0-203-73881-8. DOI: 10.1201/9780203738818.

139. S. Schmidt, A. Mauricio, P. S. Heyns, and K. C. Gryllias. "A Methodology for Identifying Information Rich Frequency Bands for Diagnostics of Mechanical Components-of-Interest under Time-Varying Operating Conditions". *Mech. Syst. Signal Process.* 142, 2020, p. 106739. DOI: 10.1016/j.ymssp.2020.106739.

140. E. Schubert. *Kmedoids 0.1.5-Dev : K-Medoids Clustering with the FasterPAM Algorithm*. URL: https://github.com/kno10/python-kmedoids.

141. E. Schubert and P. J. Rousseeuw. "Fast and Eager K-Medoids Clustering: O(k) Runtime Improvement of the PAM, CLARA, and CLARANS Algorithms". *Information Systems* 101, 2021, p. 101804. ISSN: 0306-4379. DOI: 10.1016/j.is.2021.101804. URL: https://www.sciencedirect.com/science/article/pii/S0306437921000557.

142. L. Scoccola. "Locally Persistent Categories And Metric Properties Of Interleaving Distances". PhD thesis. London, Ontario, Canada: Western University, 2020.

143. D. R. Sheehy. "A Multicover Nerve for Geometric Inference". In: *CCCG: Canadian Conference in Computational Geometry*. 2012, pp. 309–314.

144. P. Skraba and K. Turner. *Wasserstein Stability for Persistence Diagrams*. 20, 2023. arXiv: 2006.16824 [math]. (Visited on 05/17/2024). preprint.

145. E. H. Spanier. *Algebraic Topology*. Springer, New York, NY, 1981. DOI: 10.1007/978-1-4684-9322-1.

146. The RIVET Developers. *RIVET*. Version 1.1.0. 2020. URL: https://github.com/rivetTDA/rivet/.

147. A. M. Thomas. "Stochastic Process Limits for Topological Functionals of Geometric Complexes". PhD thesis. West Lafayette: Purdue University, 2021.

148. A. M. Thomas and T. Owada. "Functional Limit Theorems for the Euler Characteristic Process in the Critical Regime". *Advances in Applied Probability* 53:1, 2021, pp. 57–80. ISSN: 0001-8678, 1475-6064. DOI: 10.1017/apr.2020.46.

149. K. Turner, Y. Mileyko, S. Mukherjee, and J. Harer. "Fréchet Means for Distributions of Persistence Diagrams". *Discrete Comput Geom* 52:1, 2014, pp. 44–70. ISSN: 1432-0444. DOI: 10.1007/s00454-014-9604-7.

150. Y. Umeda. "Time Series Classification via Topological Data Analysis". *Trans. Jpn. Soc. Artif. Intell.* 32:3, 2017, D–G72_1–12. DOI: 10.1527/tjsai.D-G72.

151. M. Uray, B. Giunti, M. Kerber, and S. Huber. *Topological Data Analysis in Smart Manufacturing Processes – A Survey on the State of the Art*. arXiv, 2023. URL: http://arxiv.org/abs/2310.09319.

152. M. Vaupel and B. Dunn. *The Bifiltration of a Relation and Extended Dowker Duality*. 17, 2023. DOI: 10.48550/arXiv.2310.11529. arXiv: 2310.11529 [math].

153. L. Vietoris. "Über Den Höheren Zusammenhang Kompakter Räume Und Eine Klasse von Zusammenhangstreuen Abbildungen". *Math. Ann.* 97:1, 1927, pp. 454–472. ISSN: 1432-1807. DOI: 10.1007/BF01447877. URL: https://doi.org/10.1007/BF01447877.

154. C. Villani. *Optimal Transport*. Ed. by M. Berger, B. Eckmann, P. De La Harpe, F. Hirzebruch, N. Hitchin, L. Hörmander, A. Kupiainen, G. Lebeau, M. Ratner, D. Serre, Ya. G. Sinai, N. J. A. Sloane, A. M. Vershik, and M. Waldschmidt. Vol. 338. Grundlehren Der Mathematischen Wissenschaften. Springer, Berlin, Heidelberg, 2009. ISBN: 978-3-540-71049-3. DOI: 10.1007/978-3-540-71050-9.

155. C. Villani. *Topics in Optimal Transportation*. Vol. 58. American Mathematical Society (AMS), Providence, RI, 2003. xvi + 370 p. ISBN: 978-0-8218-3312-4.

156. Ž. Virk. "Rips Complexes as Nerves and a Functorial Dowker-Nerve Diagram". *Mediterr. J. Math.* 18:2, 2021, p. 58. ISSN: 1660-5454. DOI: 10.1007/s00009-021-01699-4. URL: https://doi.org/10.1007/s00009-021-01699-4.

157. S. Vishwanath, K. Fukumizu, S. Kuriki, and B. Sriperumbudur. *On the Limits of Topological Data Analysis for Statistical Inference*. DOI: 10.48550/arXiv.2001.00220. arXiv: 2001.00220 [math, stat].

158. B. Wang, C. Lin, H. Inoue, and M. Kanemaru. "Topological Data Analysis for Electric Motor Eccentricity Fault Detection". In: *IECON 2022 – 48th Annual Conference of the IEEE Industrial Electronics Society*. 2022, pp. 1–6. DOI: 10.1109/IECON49645.2022.9968912.

159. D. Wang. "Some Further Thoughts about Spectral Kurtosis, Spectral L2/L1 Norm, Spectral Smoothness Index and Spectral Gini Index for Characterizing Repetitive Transients". *Mech. Syst. Signal Process.* 108, 2018, pp. 360–368. DOI: 10.1016/j.ymssp.2018.02.034.

160. K. J. Worsley. "The Geometry of Random Images". *CHANCE* 9:1, 1996, pp. 27–40. ISSN: 0933-2480. DOI: 10.1080/09332480.1996.10542483. URL: https://doi.org/10.1080/09332480.1996.10542483.

161. A. Wyłomańska, R. Zimroz, J. Janczura, and J. Obuchowski. "Impulsive Noise Cancellation Method for Copper Ore Crusher Vibration Signals Enhancement". *IEEE Trans. Ind. Electron.* 63:9, 2016, pp. 5612–5621. DOI: 10.1109/TIE.2016.2564342.

162. M. C. Yesilli, F. A. Khasawneh, and B. P. Mann. "Transfer Learning for Autonomous Chatter Detection in Machining". *Journal of Manufacturing Processes* 80, 2022, pp. 1–27. DOI: 10.1016/j.jmapro.2022.05.037.

163. M. C. Yesilli, F. A. Khasawneh, and A. Otto. "Topological Feature Vectors for Chatter Detection in Turning Processes". *The International Journal of Advanced Manufacturing Technology*, 2022, pp. 1–27.

164. M. C. Yesilli, S. Tymochko, F. A. Khasawneh, and E. Munch. "Chatter Diagnosis in Milling Using Supervised Learning and Topological Features Vector". In: *2019 18th IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE, 2019, pp. 1211–1218. DOI: 10.1109/ICMLA.2019.00200.

165. D. Yogeshwaran, E. Subag, and R. J. Adler. "Random Geometric Complexes in the Thermodynamic Regime". *Probab. Theory Relat. Fields* 167:1, 2017, pp. 107–142. ISSN: 1432-2064. DOI: 10.1007/s00440-015-0678-9.

166. I. H. R. Yoon, R. Jenkins, E. Colliver, H. Zhang, D. Novo, D. Moore, Z. Ramsden, A. Rullan, X. Fu, Y. Yuan, H. A. Harrington, C. Swanton, H. M. Byrne, and E. Sahai. *Deciphering the Diversity and Sequence of Extracellular Matrix and Cellular Spatial Patterns in Lung Adenocarcinoma Using Topological Data Analysis*. bioRxiv, 2024. DOI: 10.1101/2024.01.05.574362. bioRxiv: 2024.01.05.574362.

167. A. Zomorodian. "Fast Construction of the Vietoris-Rips Complex". *Computers & Graphics*. Shape Modelling International (SMI) Conference 2010 34:3, 2010, pp. 263–271. ISSN: 0097-8493. DOI: 10.1016/j.cag.2010.03.007. URL: https://www.sciencedirect.com/science/article/pii/S0097849310000464.

168. A. Zomorodian and G. Carlsson. "Computing Persistent Homology". *Discrete Comput Geom* 33:2, 2005, pp. 249–274. ISSN: 0179-5376, 1432-0444. DOI: 10.1007/s00454-004-1146-y. URL: http://link.springer.com/10.1007/s00454-004-1146-y.