MASARYKOVA UNIVERZITA
Přírodovědecká fakulta
Ústav matematiky a statistiky

Diplomová práce

Brno 2025 **Tomáš Petit**

MASARYKOVA
UNIVERZITA
PŘÍRODOVĚDECKÁ FAKULTA
ÚSTAV MATEMATIKY A STATISTIKY

Topological data analysis

Diplomová práce

Tomáš Petit

Vedoucí práce: prof. RNDr. Jan Slovák, DrSc. **Brno 2025**

Bibliografický záznam

Autor:	Bc. Tomáš Petit Přírodovědecká fakulta, Masarykova univerzita Ústav matematiky a statistiky
Název práce:	Topological data analysis
Studijní program:	Matematika
Studijní obor:	Matematika
Vedoucí práce:	prof. RNDr. Jan Slovák, DrSc.
Akademický rok:	2024/2025
Počet stran:	?? + ??
Klíčová slova:	Topologie; Algebraická Topologie; Homologie; Persistentní Homologie; Topologická analýza dat; TDA; Topologické Metody

Bibliographic Entry

Author:	Bc. Tomáš Petit Faculty of Science, Masaryk University Department of mathematics and statistics
Title of Thesis:	Topological data analysis
Degree Programme:	Mathematics
Field of Study:	Mathematics
Supervisor:	prof. RNDr. Jan Slovák, DrSc.
Academic Year:	2024/2025
Number of Pages:	?? + ??
Keywords:	Topology; Algebraic Topology; Homology; Persistent Homology; Topological data analysis; TDA; Topological Methods

Abstrakt

V této bakalářské/diplomové/rigorózní práci se věnujeme ...

Abstract

In this thesis we study ...

MUNI
SCI

MASARYKOVA UNIVERZITA
PŘÍRODOVĚDECKÁ FAKULTA
KOTLÁŘSKÁ 2, 611 37 BRNO
IČ: 00216224
DIČ: CZ00216224

ZADÁNÍ DIPLOMOVÉ PRÁCE

Akademický rok: 2024/2025

Ústav:	Přírodovědecká fakulta
Student:	Bc. Tomáš Petit
Program:	Matematika
Specializace:	Matematika

Ředitel *ústavu* PřF MU Vám ve smyslu Studijního a zkušebního řádu MU určuje diplomovou práci s názvem:

Název práce:	Topological data analysis
Název práce anglicky:	Topological data analysis
Jazyk závěrečné práce:	angličtina

Oficiální zadání:
Goal: The goal is to understand the concepts and tools of Topological Data Analysis, and to be ready to use them in practical tasks. Aim: Depending on the results of the initial period, the student will either focus on theoretical understanding and original research in Mathematics and Statistics, or the focus will be on smart use of advanced tools in solving practical problems, including the implementation issues. One of the resources for real data requiring sophisticated analysis will come form the project Machine Learning in Nanomaterial Biocompatibility Assessment (MUNI/G/1125/2022).

Literatura: RAÚL RABADÁN, ANDREW J. BLUMBERG, Topological Data Analysis for Genomics and Evolution, Cambridge University Press, 2020, DOI: 10.1017/9781316671665

Vedoucí práce:	prof. RNDr. Jan Slovák, DrSc.
Datum zadání práce:	20. 9. 2023
V Brně dne:	25. 7. 2024

Zadání bylo schváleno prostřednictvím IS MU.

Bc. Tomáš Petit, 16. 10. 2023
prof. RNDr. Jan Slovák, DrSc., 17. 10. 2023
RNDr. Jan Vondra, Ph.D., 18. 10. 2023

Poděkování

Na tomto místě bych chtěl(-a) poděkovat ...

Prohlášení

Prohlašuji, že jsem svoji bakalářskou / diplomovou práci vypracoval(-a) samostatně pod vedením vedoucího práce s využitím informačních zdrojů, které jsou v práci citovány.

Prohlašuji, že jsem svoji rigorózní práci vypracoval(-a) samostatně s využitím informačních zdrojů, které jsou v práci citovány.

Brno xx. měsíce 20xx

.....

Tomáš Petit

Contents

List of used notation

Pro snažší orientaci v textu zde čtenáři předkládáme přehled základního značení, které se v celé práci vyskytuje.

$\mathbb{C}$	množina všech komplexních čísel
$\mathbb{R}$	množina všech reálných čísel
$\mathbb{Z}$	množina všech celých čísel
$\mathbb{N}$	množina všech přirozených čísel
$\mathbb{C}$	množina všech komplexních čísel
$\mathbb{R}$	množina všech reálných čísel
$\mathbb{Z}$	množina všech celých čísel
$\mathbb{N}$	množina všech přirozených čísel
$\mathbb{C}$	množina všech komplexních čísel
$\mathbb{R}$	množina všech reálných čísel
$\mathbb{Z}$	množina všech celých čísel
$\mathbb{N}$	množina všech přirozených čísel
$\mathbb{C}$	množina všech komplexních čísel
$\mathbb{R}$	množina všech reálných čísel
$\mathbb{Z}$	množina všech celých čísel
$\mathbb{N}$	množina všech přirozených čísel
$\mathbb{C}$	množina všech komplexních čísel
$\mathbb{R}$	množina všech reálných čísel
$\mathbb{Z}$	množina všech celých čísel

Introduction

To add later.

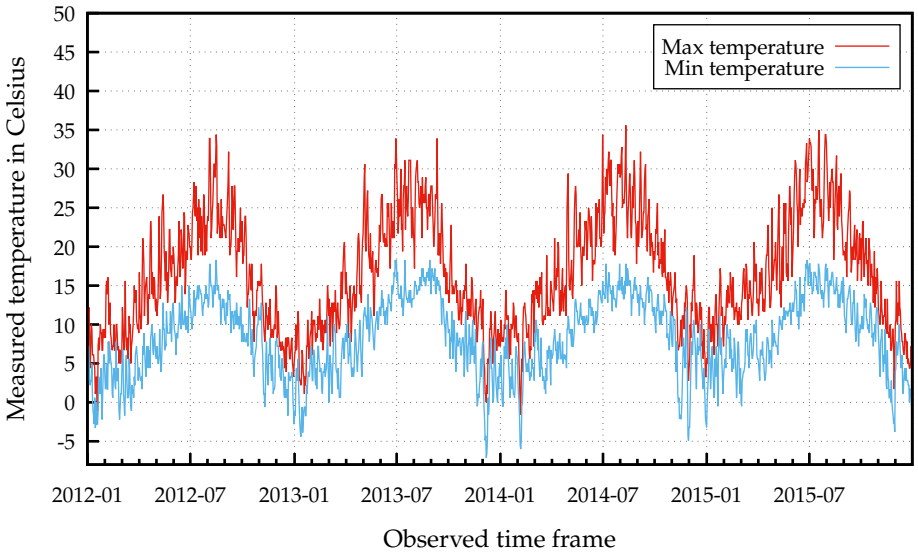
Chapter 1

Why Topology?

Data has shape. This is hardly a new or revolutionary idea in the realm of data analysis and statistics. It is an assumption that we make all the time, even if we do not say it out loud. Whenever one tries to construct a linear regression model, we all have the mental image of a straight line in our minds, which should roughly approximate the data. This is then generalized via hyperplanes in higher dimensions.

Another example would be periodic time series or signals – we all expect to see a "loop" of some sort, given a long enough time interval between the measurements, see for example 1.1. It isn't much of a stretch to imagine that we could use this loop to try to approximate the period of the time series (something that we will actually do in the following chapters, after introducing the necessary mechanisms).

Figure 1.1: Example plot of seasonal temperature changes in Seattle throughout the years.

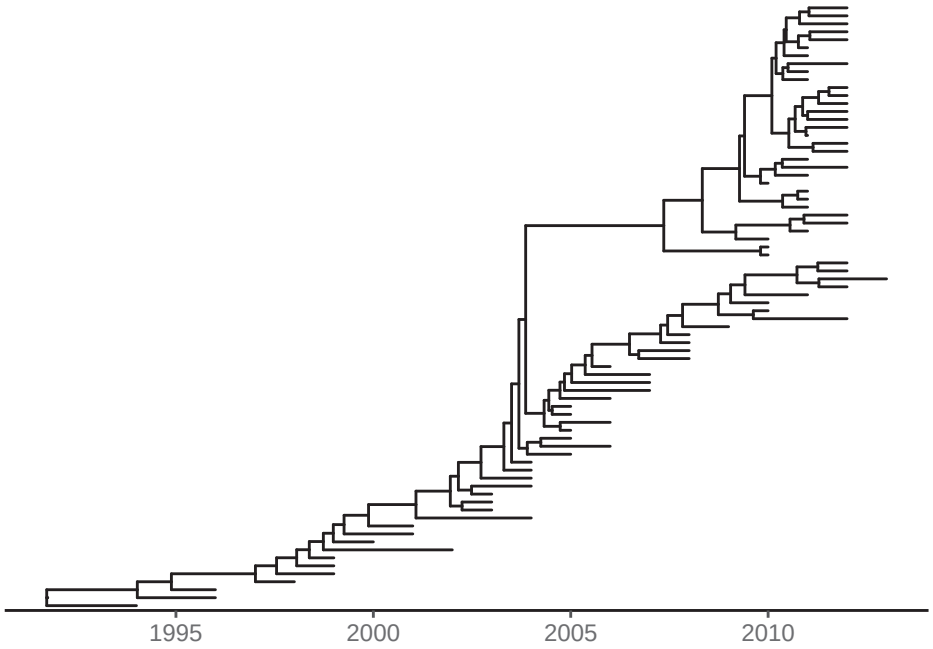


Clustering algorithms – be it k-means, hierarchical clustering and so on – all work with the shape and geometry of our data explicitly by partitioning the

available search space into the distinct clusters, once a measure of distance (which doesn't need to be a metric, per se) is chosen. The list could go on, but I believe the point was already made. Historically and traditionally, an appropriate analytic model was derived and constructed for each of the above-mentioned methods with a rich theoretical background to justify the results.

While this approach works for most simple cases the average data analyst will encounter on an Excel spreadsheet, thanks to the advancements made in computing power and software engineering, we're collecting massive amounts of complicated, high-dimensional data living faster than we did before, especially in the fields of biology and medical sciences.

Figure 1.2: Time-scaled phylogenetic tree of H3 influenza viruses inferred by BEAST using molecular clock model.



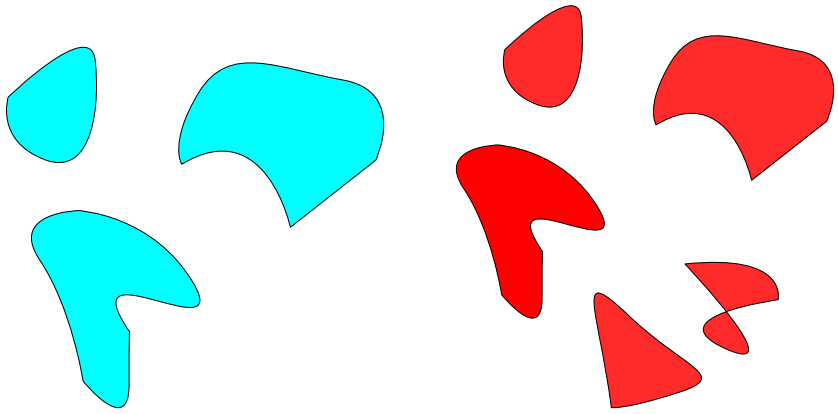
A good example of that would be phylogenetic and evolutionary trees, like the one in 1.2. We might be interested to know whether any mutation occurred, where did they happen and how far were they transmitted down the tree. We might look for re-combinations of the genomic material or both horizontal and vertical transfer of it. All those questions pertain to the shape of the phylogenetic tree and its branching. [1]

The situation only gets more complex with each passing day and batch of collected data. One *could* try to develop analytical models for each case, provide the rigorous theory and obtain the conclusions that they seek. But a far more reasonable and general method would be to instead study the shape itself and approximate the datasets by selecting the shape that describes it the "best". This is

[1] Data downloaded from the following link, visited on the 01.08.2024

where topology comes into play, as this gives us the tools and theory to do exactly that, with the help of algebraic topology.

Algebraic topology will help us to qualitatively distinguish the two situations in 1.3, where even on an intuitive level we can see that the difference lies in the number of "blobs" in each figure; or more rigorously in the number of connected components.

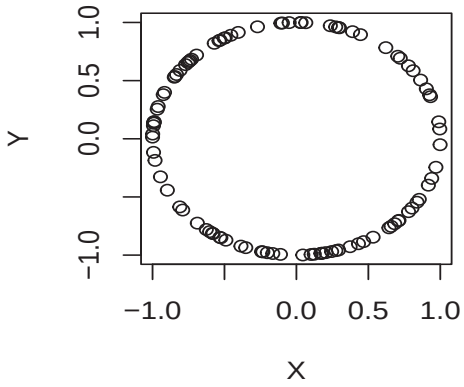


(a) 3 blobs (b) 5 blobs

Figure 1.3: Counting the number of blobs in each figure.

Likewise, using the already established machinery, we will be able to describe and quantify the fact that in 1.4 there is a hole in the middle of it.

Figure 1.4: Data sampled from a unit circle characterized by the distinct hole in the middle.



Both of these can be thought of as counting $n-$dimensional holes: connected components being of dimension 0, while the hole in 1.4 is a $1-$dimensional one. We will see how this approach and its extension will be the foundation of what we call TDA – Topological Data Analysis.

Chapter 2

Topology without tears

Unfortunately, we cannot go deeper into TDA without establishing a basic topology dictionary and theoretical background. This chapter will be a brief summary and introduction to the necessary definitions and theorems we will use. We assume that the reader already has rudimentary understanding of metric and topological spaces but for the sake of establishing notation and terminology, we're going to go through them anyway. (For more details, see Appendix A?)

2.1 Metric Spaces

Definition 2.1.1. A *metric space* is a tuple (X, ∂_X), where X is a set and $\partial_X : X \times X \to \mathbb{R}$ is a function satisfying the following:

$$1)\ \partial_X(x,y) = 0 \iff x = y.$$

$$2)\ \forall x, y \in X, \quad \partial_X(x,y) = \partial_X(y,x).$$

$$3)\ \forall x, y, z \in X, \quad \partial_X(x,z) \leq \partial_X(x,y) + \partial_X(y,z).$$

The last property is known as the *triangle inequality*.

Remark. Later on, we'll encounter *dissimilarity measures*, i.e., metrics that fail to satisfy one of the three points in the definition above. The most prominent of these will be the Kullback-Leibler divergence and the Gromov-Hausdorff distance.

Typical examples of known metrics include:

Example 2.1.1. On Euclidean spaces $\mathbb{R}^n$, for two n-tuples we have the standard distance metric given by

$$\partial_{\mathbb{R}^n}((x_1, \ldots, x_n), (y_1, \ldots, y_n)) = \sqrt{(x_1 - y_1)^2 + \ldots (x_n - y_n)^2}.$$

Example 2.1.2. Fix an alphabet Σ. Let x and y be words of length n with letters in Σ. The *Hamming distance* is then defined as the number of positions at which the letters of x and y are different:

$$\partial_H(x,y) = \{i \mid x_i \neq y_i\}.$$

Another important notion that we'll use are *open* and *closed* balls of radius ε and center $x \in X$:

$$B_\varepsilon(x) = \{z \in X \mid \partial_X(z, x) < \varepsilon\} \quad \text{and} \quad \overline{B}_\varepsilon(x) = \{z \in X \mid \partial_X(z, x) \leq \varepsilon\}$$

Chapter 3

TDA pipeline

The goal of this chapter is to establish and set up the pipeline for extracting the algebraic invariants of our data. Usually, we can only work with sampled and discrete data coming from some set of measurements. As such we can't directly use methods of algebraic topology since we won't usually be working with discrete topological spaces and to properly use these methods, we would need an uncountable amount of data; something that isn't feasible from a computational point of view.

This forces us to develop new methods to somehow approximate and recover the topology of the ambient space given only a finite set of points. Secondly, we also need to consider the *scale* of the data - some interesting properties may be more apparent only after we "zoom" in closely on them, some may not become apparent at all. All in all, we have to first construct the following pipeline:

and repeat this step for all scales at once, effectively measuring the evolution of the algebraic invariants through the changes in the feature scale.

3.1 Associated Čech and VR complexes

As mentioned above, most of the time in practice we work with a *finite metric space*, that is, a metric space (X, ∂_X) with only finitely many points. Most often, this space is $\mathbb{R}^n$ with a collection of measurements $\{x_1, \ldots, x_n\}$ and the euclidean metric.

Summary

Summary of the work.

Appendix A

Bibliography and sources