UNIVERSITÄT POTSDAM

Hans Gerhard Strohe (Hrsg.)

STATISTISCHE DISKUSSIONSBEITRÄGE

Nr. 53

Andreas Nastansky

Topologische Datenanalyse:

Eine Einführung in die Persistente Homologie und Mapper

Potsdam 2019

ISSN 0949-068X

STATISTISCHE DISKUSSIONSBEITRÄGE

Nr. 53

Andreas Nastansky

Topologische Datenanalyse: Eine Einführung in die Persistente Homologie und Mapper

Autoren: Prof. Dr. Andreas Nastansky, Hochschule für Wirtschaft und Recht (HWR) Berlin, Email: andreas.nastansky@hwr-berlin.de

Herausgeber: Prof. Dr. Hans Gerhard Strohe, ehemals Lehrstuhl für Statistik und Ökonometrie, Wirtschafts- und Sozialwissenschaftliche Fakultät der Universität Potsdam
Email: hgstrohe@uni-potsdam.de
2019

Danksagung: Ich danke Herrn M.Sc. Dariusz Lesniowski für seine Vorarbeiten und wertvolle Unterstützung.

Online veröffentlicht unter https://doi.org/10.25932/publishup-43142

Zusammenfassung

Bei der Analyse von höherdimensionalen Daten kann deren gegenseitige räumliche Anordnung im von den Variablen (Merkmalen) aufgespannten Raum wichtige Informationen über den Datensatz liefern. Bei einer gegebenen Punktwolke, die aus einem unbekannten topologischen Raum ausgewählt wurde, versucht die Topologische Datenanalyse (TDA) den ursprünglichen Raum zu rekonstruieren. Dieser Beitrag soll eine Einführung in die Topologische Datenanalyse geben und konzentriert sich dabei auf zwei wichtige Aspekte: die Persistente Homologie und den Mapper. Dabei werden zuerst die notwendigen theoretischen Grundlagen vorgestellt und anschließend wird die Methodik bei der Visualisierung von Daten eingesetzt.

Die Persistente Homologie ist eines der Standardwerkzeuge in der TDA. Sie findet ihre Anwendung beispielsweise in den Bereichen Formerkennung und -beschreibung. Der Mapper als zweites wichtiges Konzept der TDA wandelt umfangreiche, höherdimensionale Datensätze in Simplizialkomplexe um und kann dadurch geometrische und topologische Eigenschaften der Daten bestimmen. Des Weiteren ist die Mapper-Methode ein brauchbares Werkzeug zur Visualisierungen von mehrdimensionalen Daten, woran statistische Verfahren scheitern.

Abstract

In the analysis of higher-dimensional data their mutual spatial position within the variable space can provide important information about the dataset. Given a point cloud sampled from an unknown topological space, topological data analysis tries to reconstruct the original space. This paper provides an introduction into topological data analysis (TDA) and focuses on two important aspects: Persistent Homology and Mapper. First, the theoretical basics are introduced and then the methodology is applied to visualize data.

Persistent Homology represents a standard tools in the TDA. This approach is used, for example, in shape recognition and description. The Mapper is a second important concept of the TDA and converts wide, higher-dimensional datasets into simplicial complexes. By that it can determine geometric and topological properties of the data. Furthermore, the Mapper method is a useful tool for the visualization of multi-dimensional data, where statistical methods fail.

1 Einleitung

Gegenwärtig werden in Wirtschaft, Forschung, Medizin und Verwaltung zum Teil so große Datenmengen produziert, dass neue Methoden für deren Verarbeitung notwendig erscheinen. Für deren Auswertung werden vielfältige statistischen Methoden genutzt. Obwohl auf die Statistik bei der Analyse von Daten nicht verzichtet werden kann, ist es zum Teil sinnvoll, eine Methodik zu wählen, die einen komplett anderen Grundansatz verfolgt. Dadurch eröffnet sich unter Umständen ein alternativer Blickwinkel auf die Daten. Zu diesen Alternativen zählt die Topologische Datenanalyse (TDA). Die Topologische Datenanalyse repräsentiert dabei ein vergleichsweise junges Gebiet der Mathematik und ist mit ihren Anwendungen u.a. an den Bereichen Data Mining und Visualisierung von Daten ausgerichtet.

Die Topologie gehört zu den Grunddisziplinen der Mathematik. Sie liefert theoretische Grundlagen, die viele andere Gebiete der Mathematik - wie zum Beispiel die Geometrie, Funktionalanalysis oder komplexe Analysis - beeinflusst haben. Gleichzeitig wurde sie lange Zeit nur zur Behandlung abstrakter Objekte genutzt, da sie als ein Bereich der Mathematik angesehen wurde, der keine direkte Anwendung ermöglicht. In den vergangenen 15 Jahren hat sich diese Situation stark verändert. Carlsson hat den Einsatz der Topologie bei der Analyse von Daten popularisiert und die algorithmischen Grundlagen für die Auswertung und Visualisierung von höherdimensionalen und komplexen Datensätzen mit topologischen Instrumenten gelegt. In Zusammenarbeit mit anderen zeigte er die Anwendbarkeit der Persistenten Homologie und der Mapper-Methode als zwei wesentliche Verfahren der TDA in einer Reihe von Arbeiten [5]. Obwohl die TDA noch ein relativ junges Gebiet ist, hat sie sich bereits als nützliches Werkzeug bei der Analyse von komplexen Daten etabliert [4]. Sie findet erste Verwendung sowohl in der Wissenschaft als auch im kommerziellen Bereich.

Die Topologie beschäftigt u.a. mit der Form eines Objektes. Dadurch kann sie dem Nutzer quantitative Informationen über die Form des Datensatzes liefern. Die Topologie repräsentiert hierbei ein mathematisches Fachgebiet, das u.a. verschiedene topologische Räume zu klassifizieren versucht. Solche Räume besitzen innerhalb einer Klasse gleiche topologische Eigenschaften und lassen sich stetig aufeinander abbilden. Mit Hilfe der Topologie ist es möglich, gleiche Strukturen in verschiedenen Räumen zu entdecken. Bei der Arbeit mit den höherdimensionalen Daten ist es erstrebenswert, in den großen und häufig schwer überschaubaren Datensätzen Muster zu finden, die letztendlich wichtige Aussagen zu dem entsprechenden Gebiet liefern können. Dies führt zum Begriff der Topologischen Datenanalyse (TDA). Formal ausgedrückt versucht die TDA – gegeben eine endliche Menge von Punkten ausgewählt aus einem unbekannten topologischen Raum – die Topologie des Raumes wiederherzustellen. Bei der Analyse von komplexen, höherdimensionalen Daten kann deren Gestalt wichtige Informationen über den Datensatz liefern. Bei einer gegebenen Punktwolke, die aus einem unbekannten topologischen Raum ausgewählt wurde, versucht die Topologische Datenanalyse den ursprünglichen Raum zu rekonstruieren.

Der Beitrag ist wie folgt gegliedert: In den Kapiteln 2 und 3 werden zuerst die notwendigen theoretischen Grundlagen vorgestellt. Hierzu zählen u.a. die Begriffe topologische Räume und Simplizialkomplexe der algorithmischen Topologie. Anschließend werden in

den Abschnitten 4 und 5 die beiden Kernverfahren der TDA (Persistente Homologie und Mapper) dargelegt und anhand von Anwendungsbeispielen demonstriert. Speziell für Mapper werden Möglichkeiten der Visualisierung aufgezeigt. Der Beitrag endet mit einer kritischen Diskussion der Methodik.

2 Einführung in die Topologische Datenanalyse

In diesem Kapitel werden zunächst die theoretischen Grundlagen der Topologie dargelegt und anschließend Simplizialkomplexe und die dazugehörigen Konzepte erläutert [8, 10, 12, 19].

2.1 Topologische Räume

Wir beginnen mit einer Definition der Elemente topologischer Räume.

Definition 2.1 (Topologie)**:** Sei X eine Menge und $\mathcal{U} \subseteq \mathcal{P}(X)$ eine Familie von Teilmengen von X mit den Eigenschaften:

(i) $\emptyset, X \in \mathcal{U}$,

(ii) $U, V \in \mathcal{U} \Rightarrow U \cap V \in \mathcal{U}$,

(iii) $\{U_i\}_{i \in I} \in \mathcal{U} \Rightarrow \bigcup_{i \in I} U_i \in \mathcal{U}$.

Dann ist $\mathcal{U}$ eine **Topologie** auf X, $\mathbb{X} = (X, \mathcal{U})$ ist ein **topologischer Raum** und die Elemente von $\mathcal{U}$ werden **offene Mengen** genannt.

Typisches Beispiel für einen topologischen Raum ist der Euklidische Raum $\mathbb{R}^n$ mit der Topologie $\mathcal{U} = \{U \subseteq \mathbb{R}^n \mid \forall x \in U \ \exists \epsilon : B_\epsilon(x) \subset U\}$, wobei $B_\epsilon(x) := \{y \in \mathbb{R}^n \mid d(x, y) < \epsilon, \epsilon \in \mathbb{R}_+\}$. Der topologische Raum stellt an sich ein sehr allgemeines Konzept dar. Allgemein gilt, dass die metrischen Räume auch topologische Räume sind (eine Metrik induziert eine Topologie).

Für die Abbildungen zwischen den topologischen Räumen existiert auch der verallgemeinerte Begriff der Stetigkeit.

Definition 2.2 (Stetigkeit)**:** Seien $\mathbb{X}, \mathbb{Y}$ zwei topologische Räume. Eine Abbildung $f : \mathbb{X} \to \mathbb{Y}$ heißt **stetig**, falls für alle $U \in \mathcal{U}_\mathbb{Y}$ gilt $f^{-1}(U) \in \mathcal{U}_\mathbb{X}$.

Eine Abbildung ist demnach stetig, wenn die Urbilder der offenen Mengen selbst offen sind. In der Topologie spielen stetigen Abbildungen eine zentrale Rolle in der Klassifizierung von Objekten. Dies führt zum Begriff des Homöomorphismus.

Definition 2.3 (Homöomorphismus)**:** Seien $\mathbb{X}$ und $\mathbb{Y}$ topologische Räume. Eine Abbildung $f : \mathbb{X} \to \mathbb{Y}$ ist ein **Homöomorphismus**, falls folgendes gilt:

(i) f ist bijektiv,

(ii) f ist stetig,

(iii) die Umkehrfunktion f^{-1} ist stetig.

Zwei Räume $\mathbb{X}$ und $\mathbb{Y}$ heißen homöomorph, $\mathbb{X} \cong \mathbb{Y}$, wenn zwischen ihnen ein Homöomorphismus existiert.

Aus topologischer Sicht gibt es zwischen zwei homöomorphen Objekten keinen Unterschied. Ein Raum kann stetig in den anderen überführt werden. Eine wichtige Aufgabe der Topologie ist es zu zeigen, ob zwei Räume homöomorph sind oder nicht. Das kann belegt werden, wenn mindestens ein Homöomorphismus zwischen den Räumen gefunden werden kann.

Beispiel 2.1:
Für ein Quadrat $\mathbb{Y} = \{(x, y) \in \mathbb{R}^2 \mid \max\{|x|, |y|\} = 1\}$ und einen Kreis $\mathbb{X} = \{(x, y) \in \mathbb{R}^2 \mid x^2 + y^2 = 1\}$ gilt $\mathbb{X} \cong \mathbb{Y}$. Der dazu zugehörige Homöomorphismus lautet $f : \mathbb{Y} \to \mathbb{X}, (x, y) \mapsto \left(\frac{x}{\sqrt{x^2+y^2}}, \frac{y}{\sqrt{x^2+y^2}}\right)$.

Es ist im Allgemeinen schwieriger zu zeigen, dass zwei Räume nicht homöomorph sind. Hilfreich dabei sind sogenannte **topologische Invarianten**. Eine Invariante ist eine Abbildung f, die zwei homöomorphen Räumen gleiche Objekte zuordnet, d.h.

$$\mathbb{X} \cong \mathbb{Y} \Rightarrow f(\mathbb{X}) = f(\mathbb{Y}).$$

Die Kontraposition

$$f(\mathbb{X}) \neq f(\mathbb{Y}) \Rightarrow \mathbb{X} \ncong \mathbb{Y},$$

zeigt, ob die Räume $\mathbb{X}$ und $\mathbb{Y}$ homöomorph sind. Gilt $f : \mathbb{X} \to \mathbb{R}$, dann wird von einer numerischen Invariante gesprochen. Zielräume für f müssen aber keine reelle Zahlen sein. Es kann genauso $f : \mathbb{X} \to R$-Modul gelten (R-Modul ist eine Verallgemeinerung eines Vektorraums). Es gibt viele Invarianten mit sehr unterschiedlichem Charakter. Für die Computerwissenschaft kommt den abstrakten Invarianten geringe Bedeutung bei. Dagegen aber Invarianten, die sich algorithmisch berechnen lassen. Vor allem Letztere ist für die Topologische Datenanalyse nützlich, wie noch in Kapitel 4 näher ausgeführt wird.

Weitere wichtige Begriffe der Topologie sind die Homotopie und die Homotopieäquivalenz.

Definition 2.4 (Homotopie, homotop)**:** Zwei stetige Abbildungen $f, g : \mathbb{X} \to \mathbb{Y}$ zwischen topologischen Räumen heißen **homotop** (Notation: $f \simeq g$), wenn es eine **Homotopie** h zwischen ihnen gibt, d.h. eine stetige Abbildung $h : \mathbb{X} \times [0, 1] \to \mathbb{Y}$ mit $h(x, 0) = f(x)$ und $h(x, 1) = g(x)$, für alle $x \in \mathbb{X}$.

Definition 2.5 (Homotopieäquivalenz, homotopieäquivalent)**:** Eine stetige Abbildung $f : \mathbb{X} \to \mathbb{Y}$ heißt eine **Homotopieäquivalenz** zwischen $\mathbb{X}$ und $\mathbb{Y}$, wenn eine stetige Abbildung $g : \mathbb{Y} \to \mathbb{X}$ mit $g \circ f \simeq \mathbb{1}_{\mathbb{X}}$ und $f \circ g \simeq \mathbb{1}_{\mathbb{Y}}$ gibt (Homotopieinverse). Die Räume $\mathbb{X}$ und $\mathbb{Y}$ heißen dann **homotopieäquivalent** (Notation $\mathbb{X} \simeq \mathbb{Y}$).

Beispiel 2.2:
Seien $f : S^1 \to \mathbb{R}^2, x \mapsto x$ die Einbettung des Kreises in $\mathbb{R}^2$ und $g : S^1 \to \mathbb{R}^2, x \mapsto 0$ eine

Abbildung, die den Kreis auf den Ursprung abbildet. Die Abbildung $h : S^1 \times [0,1] \to \mathbb{R}^2, h(x,t) = (1-t)f(x)$ ist stetig. Außerdem gilt $h(x,0) = f(x)$ und $h(x,1) = 0 = g(x)$ (siehe Abbildung 1). Damit sind f und g homotop. Zu beachten wäre, dass obwohl eine Homotopie zwischen die Abbildungen f und g existiert, der Kreis und der Punkt nicht homotopieäquivalent sind.

Homotopieäquivalenz dagegen bedeutet, dass zwei Objekte aufeinander und wieder zurück stetig deformiert werden können. In Abbildung 2 wird die Homotopieäquivalenz zwischen einem Quadrat und einem Kreis anschaulich dargestellt. Die Homotopieäquivalenz folgt direkt aus dem Fakt, dass die beide Räume homöomorph zueinander sind. Demzufolge repräsentiert die Topologie ein mathematisches Fachgebiet, das u.a. verschiedene topologische Räume zu klassifizieren versucht. Solche Räume besitzen gleiche topologische Eigenschaften und lassen sich stetig aufeinander abbilden.

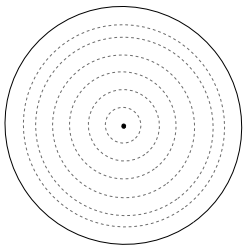


Abbildung 1: Veranschaulichung der Homotopie zwischen f und g aus dem Beispiel 2.2

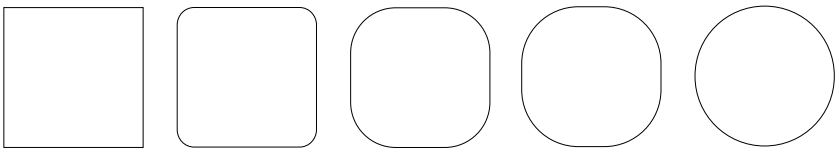


Abbildung 2: Homotopieäquivalenz zwischen einem Kreis und einem Quadrat

Mit Hilfe der Topologie ist es möglich, gleiche Strukturen in verschiedenen Räumen zu entdecken. Bei der Arbeit mit höherdimensionalen Daten ist es essentiell, in den großen und unüberschaubaren Datensätzen Muster zu finden, die uns letztendlich wichtige Aussagen zu dem entsprechenden Gebiet liefern können. Das führt zum Begriff der Topologischen Datenanalyse (TDA). Formal ausgedrückt, versucht die TDA, gegeben eine endliche Menge von Punkten S, ausgewählt aus einem unbekannten topologischen Raum $\mathbb{X}$, die Topologie des Raumes $\mathbb{X}$ wiederherzustellen. Gewöhnlich wird die TDA in zwei Schritten ausgeführt:

(i) Benutze S um eine geeignete kombinatorische Struktur K zu erzeugen. Beispiele solcher Strukturen und deren Konstruktion werden im Kapitel 3 vorgestellt.

(ii) Mit Hilfe der kombinatorischen Struktur K wird versucht, die Topologie von $\mathbb{X}$ zu finden. Das geschieht zum Beispiel mithilfe von topologischen Invarianten. Mehr dazu in Kapitel 4.

2.2 Simplizialkomplexe

Bis hierhin wurden hauptsächlich Konzepte aus der mengentheoretischen Topologie behandelt. Beim Lösen vieler topologischer Fragen ist die mengentheoretische Topologie nicht ausreichend. Mit Hilfe algebraischer Strukturen ist es möglich, topologischen Problemstellungen zu untersuchen und zu lösen. Das führt zum Gebiert der algebraischen Topologie. Simplizialkomplexe sind wegen deren Einfachheit und gleichzeitigen Vorteilen sehr beliebte topologische Objekte in der algebraischen Topologie und müssen als Nächstes definiert werden.

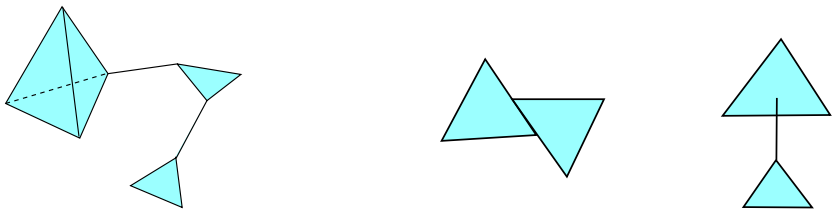
Definition 2.6 (Simplex)**:** Ein d**-Simplex** im $\mathbb{R}^n$ $(d \leq n)$ ist die konvexe Hülle $conv(v_0, \ldots, v_d)$ von $d+1$ Punkten $\{v_0, \ldots, v_d\}$ in allgemeiner Lage. Die konvexe Hülle von $\{v_0, \ldots, v_d\}$ ist die Menge $\{\sum_{i=0}^{d} \lambda_i v_i \mid \lambda_i \geq 0 \; und \; \lambda_0 + \cdots + \lambda_d = 1\}$ und die allgemeine Lage bedeutet, dass $\{v_1 - v_0, \ldots, v_d - v_0\}$ linear unabhängig ist.

Definition 2.7 (geometrischer Simplizialkomplex)**:** Eine Menge von endlich vielen Simplizes $K \subseteq \mathbb{R}^n$ heißt ein **(geometrischer) Simplizialkomplex**, wenn folgendes gilt:

(i) Sei $\sigma \in K$, $\tau \subseteq \sigma \Rightarrow \tau \in K$,

(ii) Für $\tau, \sigma \in K \Rightarrow \tau \cap \sigma \in K$ oder $\tau \cap \sigma = \varnothing$.

Die Vereinigung $|K| := \bigcup_{\sigma \in K} \sigma \subset \mathbb{R}^n$ aller Simplizes eines Komplexes K heißt der K zugrundeliegende topologische Raum.



(a) Simplizialkomplex

(b) Keine Simplizialkomplexe

Abbildung 3: Beispiel und Gegenbeispiele für Simplizialkomplexe

Beispiel 2.3:
Anschaulich ist ein 0-Simplex ein Punkt, ein 1-Simplex eine Strecke, ein 2-Simplex ein Dreieck, ein 3-Simplex ein Tetraeder, usw. In Abbildung 3(a) wird veranschaulicht, wie ein Simplizialkomplex aussehen kann. Abbildung 3(b) zeigt zwei Gegenbeispiele. Hier ist die Bedingung (ii) aus der Definition 2.7 verletzt.

Die in der Praxis zu analysierenden Datensätze können sehr umfangreich sein und setzen den Einsatz von computergestützter Datenananalyse voraus. Simplizialkomplexe besitzen eine recht einfache, diskrete Struktur und können dadurch auch gut mit Computerprogrammen behandelt werden. Gleichzeit erlauben sie Berechnungen von sehr mächtigen Invarianten, wie die Homologie. Die in der Definition 2.7 vorgestellten Komplexe werden in der Regel geometrische Simplizialkomplexe genannt. Simplizialkomplexe

können auch abstrakt definiert werden. Manchmal ist es hilfreicher, einen Komplex erst abstrakt zu beschreiben, ohne sich dabei Gedanken über seine geometrische Darstellung zu machen. In diesem Fall wird von abstrakten Simplizialkomplexen gesprochen.

Definition 2.8 (abstrakter Simplizialkomplex)**:** Für eine endliche Menge von Ecken $V = \{v_1, \ldots, v_d\}$ wird K ein **(abstrakter) Simplizialkomplex** genannt, wenn folgendes gilt:

(i) $K \subseteq \mathcal{P}(V)$

(ii) $\sigma \in K, \tau \subseteq \sigma \Rightarrow \tau \in K$.

Wenn für $\sigma \in K$ gilt $|\sigma| = n + 1$, dann heißt σ ein n-Simplex der Dimension $dim(\sigma) = n$. Wenn d die maximale Dimension eines Simplex in K ist, dann wird von einem d-dimensionalen Komplex gesprochen.

Bei abstrakten Simplizialkomplexen werden nur die Eckpunkte der Simplexe betrachtet. Aus einem geometrischen Simplizialkomplex kann ein abstrakter Simplizialkomplex konstruiert werden. Sei K ein geometrischer Simplizialkomplex, $\sigma \in K$ ein d-Simplex und $\sigma_{abst} = \{v_0, \ldots, v_d\}$ die dazugehörige Knoten-Menge, dann ist $K_{abst} = \bigcup_{\sigma \in K} \sigma_{abst}$ ein abstrakter Simplizialkomplex erzeugt aus K.

Andersherum kann gezeigt werden, dass die abstrakten Simplizialkomplexe immer eine geometrische Realisierung in $\mathbb{R}^n$ besitzen. Eine Grenze für n, für die diese Realisierung immer möglich ist, kann sogar genau angegeben werden.

Satz 2.1: Jeder abstrakte Simplizialkomplex der Dimension d besitzt eine geometrische Darstellung im $\mathbb{R}^{2d+1}$ [8].

Abschließend sollen noch die Konzepte der Überdeckung und des Nervs vorgestellt werden. Wie bereits angemerkt, ist S eine Punktwolke ausgewählt aus einem topologischen Raum $\mathbb{X}$. Hauptidee der TDA ist, $\mathbb{Y}$ lokal zu benutzen um $\mathbb{X}$ zu approximieren. Dazu werden die Begriffe Überdeckung und Nerven gebraucht.

Definition 2.9 (Überdeckung, Nerv)**:** Eine **offene Überdeckung** von S ist eine Menge $\mathcal{U} = \{U_i\}_{i \in I}$, $U_i \subseteq \mathbb{X}$, wobei $S \subseteq \bigcup_{i \in I} U_i$ und U_i offene Mengen sind. Eine Überdeckung wird **gute** Überdeckung genannt, wenn alle U_i und deren nicht leeren, endlichen Überschneidungen jeweils homotop zu einem Punkt sind. **Nerv** von $\mathcal{U}$ ist die Menge N, für die gilt:

(i) $\emptyset \in N$

(ii) $\bigcap_{j \in J} U_j \neq \emptyset$ für $J \subseteq I \Rightarrow J \in N$.

Ein Nerv ist ein Simplizialkomplex. Nerven sind eine kombinatorische Darstellungen von $\mathcal{U}$, mit denen die notwendigen Berechnungen, wie zum Beispiel die Bestimmung der Homologie, durchgeführt werden. Im nächsten Kapitel werden verschiedene Möglichkeiten vorgestellt, wie solche Überdeckungen und Nerven konstruiert werden.

3 Komplexe aus der algorithmischen Topologie

In diesem Kapitel werden einige der am häufigsten benutzten Strukturen vorgestellt, die zur Berechnung von der Persistenten Homologie angewandt werden [5, 8, 25]. Ziel ist es, mit Hilfe von kombinatorischen Strukturen aus einem vorgegebenen Datensatz einen Komplex zu erzeugen und mit diesem weitere topologische Analysen durchzuführen. Das entspricht dem ersten Schritt der TDA (siehe Kapitel 2.1, Seite 4). Da in der Praxis zum Teil sehr umfangreiche Datensätze analysiert werden, ist es notwendig, dass sich die erzeugten Komplexe programmiertechnisch elegant berechnen lassen.

In den folgenden Abschnitten seien $S \subseteq \mathbb{Y}$ eine endliche Punktmenge und $\mathbb{Y}$ der unbekannte metrische Raum mit Metrik $d : \mathbb{Y} \times \mathbb{Y} \to \mathbb{R}$ und K die kombinatorische Struktur erzeugt aus S. Mit der Hilfe von K wird versucht, die Topologie von $\mathbb{Y}$ zu rekonstruieren.

3.1 Čech-Komplex

Zu Beginn werden die Čech-Komplexe definiert. Die Idee hierfür ist intuitiv: Lege um alle Punkte offene Kugeln mit gleichem Radius und verbinde zwei Punkte, soweit der Schnitt von ihren zugehörigen offenen Kugeln nicht leer ist.

Definition 3.1 (Čech-Komplex)**:** Sei $B_\epsilon(x)$ eine offene Kugel mit dem Radius $\epsilon \in \mathbb{R}_+$ um den Punkt $x \in \mathbb{Y}$, definiert als:

$$B_\epsilon(x) := \{y \in \mathbb{Y} \mid d(x, y) < \epsilon\}.$$

Für $S \subset \mathbb{Y}$ gibt es dann eine Überdeckung, die in folgender Weise konstruiert wird

$$U_\epsilon = \{B_\epsilon(x) \mid x \in S\}.$$

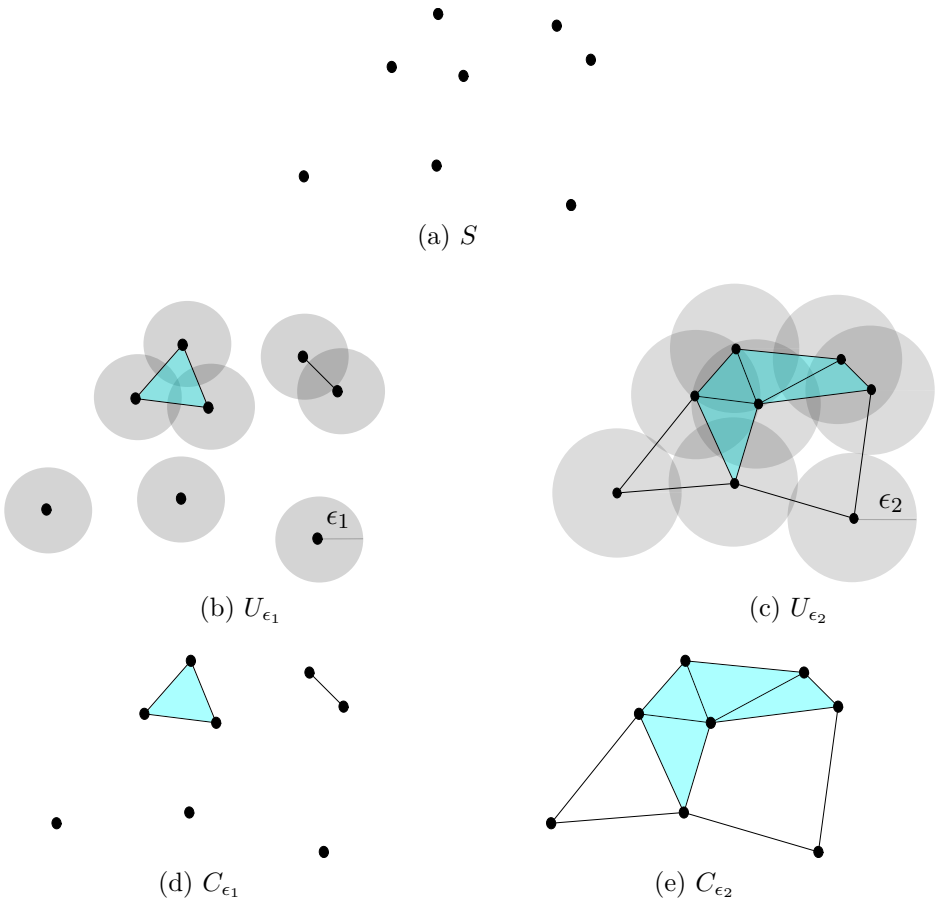
Der Nerv von U_ϵ wird **Čech-Komplex** C_ϵ genannt.

Beispiel 3.1:
Abbildung 4 zeigt zwei Überdeckungen U_{ϵ_1} und U_{ϵ_2} und die dazugehörigen Čech-Komplexe C_{ϵ_1} und C_{ϵ_2}.

Bei Čech-Komplexen wird für alle $x \in S$ der gleiche Radius genommen. Dies suggeriert, dass alle Punkte gleichverteilt in $\mathbb{X}$ liegen. In der Realität ist eher das Gegenteil der Fall. Es ist möglich, ein Čech-Komplex für ein beliebiges ϵ zu berechnen. Dabei gilt, dass $C_0 = \emptyset$ und C_∞ ein $(|S| - 1)$-Simplex ist. Folglich kann die Dimension des Čech-Komplexes größer als die Dimension von $\mathbb{Y}$ sein. Deswegen und weil der Berechnungsaufwand für Čech-Komplexe sehr groß sein kann, werden in der Praxis andere kombinatorische Konstruktionen bevorzugt. Die Idee dahinter ist, die Dimension des Komplexes zu reduzieren.

3.2 Alpha-Komplex

Um die Dimension des Komplexes zu reduzieren, bietet es sich an, die Dimension des Raumes $\mathbb{Y}$ zu benutzen. Für einen gegebenen Punkt $x \in S \subseteq \mathbb{Y}$ sei die Voronoi-Region



(a) S (b) U_{ϵ_1} (c) U_{ϵ_2} (d) C_{ϵ_1} (e) C_{ϵ_2}

Abbildung 4: Punktwolke S, Überdeckungen U_{ϵ_1}, U_{ϵ_2}, sowie dazugehörige Čech-Komplexe C_{ϵ_1} und C_{ϵ_2}

$R(x)$ definiert als

$$R(x) := \{y \in \mathbb{Y} \mid d(x,y) \leq d(x',y), \forall x' \in S, x' \neq x\}.$$

Die Vereinigung aller Voronoi-Regionen wird **Voronoi-Diagramm** genannt. Werden die Knoten aus S miteinander verbunden, die einer gemeinsamen Voronoi-Diagramm-Kante zugeordnet sind, entsteht dann die **Delaunay-Triangulierung**.

Beispiel 3.2:
Abbildung 5(a) zeigt ein Voronoi-Diagramm und eine Delaunay-Triangulierung für S. Die Delaunay-Triangulierung ist ein Simplizialkomplex mit der Dimension d. Die Dimension vom Voronoi-Diagramm und der Delaunay-Triangulierung ist gleich der Dimension des Raumes $\mathbb{Y}$.

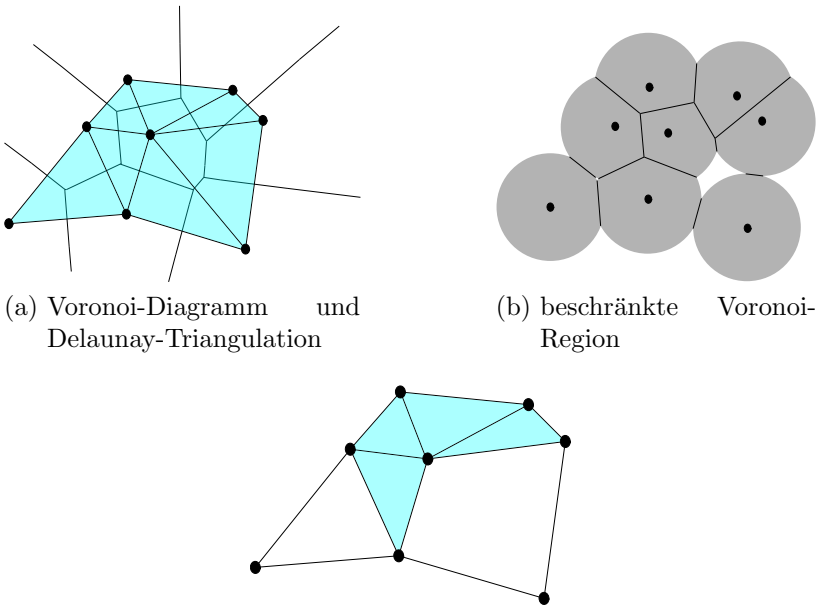
An dieser Stelle wird der Alpha-Komplex definiert, indem der Schnitt von der Voronoi-Region und den offenen Kugeln von $x \in S$ gebildet wird.

Definition 3.2 (Alpha-Komplex)**:** Für $S \subset \mathbb{Y}$ sei die Überdeckung definiert als:

$$U_\epsilon = \{B_\epsilon(x) \cap R(x) \mid x \in S\}.$$

U_ϵ wird beschränkte Voronoi-Region genannt. Der Nerv von U_ϵ ist dann der **Alpha-Komplex** A_ϵ.

Beispiel 3.3:
Abbildung 5(b) bildet eine beschränkte Voronoi-Region ab und Abbildung (c) A_ϵ zeigt einen Alpha-Komplex für die Punktwolke S. Der Radius ϵ entspricht dem Radius ϵ_2 aus dem Beispiel 3.1.



(a) Voronoi-Diagramm und Delaunay-Triangulation

(b) beschränkte Voronoi-Region

(c) Alpha-Komplex A_ϵ

Abbildung 5: Voronoi-Diagramm, Delaunay-Triangulierung sowie beschränkte Voronoi-Region und Alpha-Komplex für die Punktwolke S

Es gilt $U_\epsilon \subseteq B_\epsilon(x)$, was auch $A_\epsilon \subseteq C_\epsilon$ impliziert. Jedoch besitzen A_ϵ und C_ϵ den gleichen Homotopietyp [25]. Im Gegensatz zu C_ϵ besitzt aber A_ϵ die gleiche Dimension wie der Einbettungsraum $\mathbb{Y}$, was auch das Ziel war. Weiter gilt, dass $A_0 = \emptyset$ und A_∞ die Delaunay-Triangulierung ist.

In der Definition 3.2 wurde bis jetzt angenommen, dass alle Radien der offenen Kugeln gleich sind. Diese Definition kann „gelockert“ werden. Es existieren generalisierte Alpha-Komplexe, bei denen diese Annahme weggelassen wird [7]. Es sind effektive Algorithmen für die Konstruktion von Delaunay-Triangulierung (und damit auch für Alpha-Komplexe) in den Dimensionen 2 und 3 vorhanden [25]. Für höhere Dimensionen ist die Theorie noch nicht komplett entwickelt und es gibt bisher noch keine allgemeinen effektiven Konstruktionsverfahren für Alpha-Komplexe in den Dimensionen $d > 3$. Das Problem des Mangels an effektiven Algorithmen wird sich jedoch mit den weiteren Simplizialkomplexen ändern.

3.3 Vietoris-Rips-Komplex

Der Vorteil des Vietoris-Rips-Komplex besteht in deren einfachen und effizienten Konstruierbarkeit. Deshalb werden sie präferiert in der Topologischen Datenanalyse genutzt.

Definition 3.3 (Vietoris-Rips-Komplexe)**:** Sei $x_i \in S$. Ein **Vietoris-Rips-Komplex** V_ϵ wird konstruiert, indem aus $\{x_0,\ x_1,\ \ldots,\ x_n\}$ immer dann ein n-Simplex erzeugt wird, wenn $d(x_i, x_j) < \epsilon, 0 \leq i, j \leq n$ ist.

Beispiel 3.4:
Die Abbildung 6 zeigt S und dazu den konstruierten Vietoris-Rips-Komplex V_ϵ. Der Radius ϵ entspricht dem $2\epsilon_2$ aus Beispiel 3.1 und kann damit direkt mit C_{ϵ_2} verglichen werden.

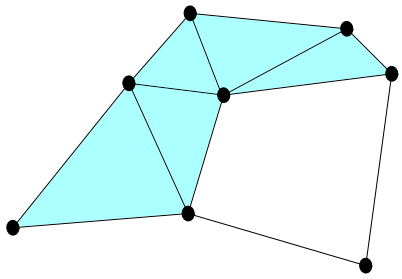


Abbildung 6: Vietoris-Rips-Komplex $V_{2\epsilon_2}$

Das Beispiel 3.4 verdeutlicht zudem, dass der Čech-Komplex und der Vietoris-Rips-Komplex nicht homotopiäequivalent sein müssen. $V_{2\epsilon}$ beinhaltet die gleichen Kanten, die auch im C_ϵ zu finden sind. $V_{2\epsilon}$ beinhaltet aber auch alle Simplexe, die mit den Kanten erzeugt werden können. Es gilt also $C_\epsilon \subseteq V_{2\epsilon}$. Für zwei beliebige Punkte eines Simplex aus $V_{2\epsilon}$ gilt: $d(x, y) < 2\epsilon$. $C_{2\epsilon}$ wird demzufolge mindestens alle Simplexe aus $V_{2\epsilon}$ beinhalten. Das bedeutet dass $V_{2\epsilon} \subseteq C_{2\epsilon}$ ist. Damit gilt die Relation:

$$C_\epsilon \subseteq V_{2\epsilon} \subseteq C_{2\epsilon}.$$

Wie bei Čech-Komplexen gilt, dass für $\epsilon = \infty$ der Vietoris-Rips-Komplex V_ϵ ein $(|S| - 1)$-Simplex ist. In die Praxis werden V_ϵ nur für ein endliches ϵ berechnet. Es existieren schnelle Algorithmen für die Konstruktion von Vietoris-Rips-Komplexen in höheren Dimensionen, die u.a. in den folgenden Softwareprogrammen (für C++, MATLAB, R) enthalten sind.

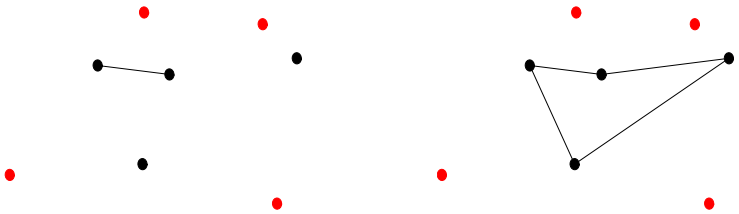
3.4 Witness-Komplex

Vietoris-Rips-Komplexe sind zwar schnell zu konstruieren, können aber leider auch sehr große Dimensionen erreichen. Um dies zu umgehen, wird versucht, mit einer kleineren Anzahl von Knoten als in S vorhanden ist, eine Struktur K zu konstruieren. Das geschieht mit Hilfe von sogenannten Witnesses (dt. Zeugen).

Definition 3.4 (Witness-Komplex)**:** Sei $L \subseteq S$, genannt Menge der **Landmarken**, $W = S - L$ die Menge der Witnesses, $w \in W$, $x_i \in L$ und $\epsilon > 0$.

Ein **Witness-Komplex** W_ϵ wird konstruiert, indem aus $\{x_0,\ x_1,\ \ldots,\ x_n\}$ immer dann ein n-Simplex erzeugt wird, wenn $d(x_i, w) < \epsilon, 0 \leq i \leq n$.

Beispiel 3.5:
In den Abbildungen 7(b) und 7(c) werden zwei Witness-Komplexe $W_{2\epsilon_1}$ und $W_{2\epsilon_2}$ dargestellt. Dabei sind die roten Punkte die Witnesses und die schwarzen Punkte die Landmarken.



(a) $W_{2\epsilon_1}$ (b) $W_{2\epsilon_2}$

Abbildung 7: Witness-Komplexe $W_{2\epsilon_1}$ und $W_{2\epsilon_2}$

Für $\epsilon < \epsilon'$ gilt die Inklusion

$$W_\epsilon \hookrightarrow W_{\epsilon'}.$$

Es ist zu erkennen, dass das Aussehen des Witness-Komplex abhängig von der Wahl der Landmarken ist. In der Praxis werden dazu u.a. Bootstrap-Methoden benutzt. Es existieren aber auch andere Verfahren, um die Landmarken auszuwählen. Der Witness-Komplex gehört jedoch zu den praktikableren Methoden. Wie im Fall des Vietoris-Rips-Komplex existieren Programme für die Konstruktion von Witness-Komplexen.

In diesem Kapitel wurden einige kombinatorische Strukturen vorgestellt, die auch tatsächlich in der Praxis Anwendung finden (vgl. [5]). Je nach Struktur des Datensatzes oder den zu untersuchenden Eigenschaften werden bestimmte Konstruktionen präferiert. Damit wurde ein Fundament für die weitere Analyse geschaffen. Die nächsten Schritte bestehen darin, die definierten Komplexe mit Hilfe von topologischen Werkzeugen zu analysieren und eine Schlussfolgerung auf den ursprünglichen Datensatz zu treffen. Dabei können die Datenmengen sehr umfangreich sein. Ein Werkzeug aus der abstrakten algebraischen Topologie, das sich auch gut mit Computerprogrammen behandeln lässt, ist die Persistente Homologie. Damit setzt sich das nächste Kapitel auseinander.

4 Persistente Homologie

In den vorangegangen Abschnitten wurden verschiedene algorithmische Strukturen K (bzw. Simplizialkomplexe) vorgestellt. Die nächste Aufgabe wäre das Lösen des Homöomorphismus-Problems, d.h. zu welchen topologischen Räumen sind die erzeugten Strukturen homöomorph. Das Ergebnis des Problems ist leider negativ, da dieses Problem algorithmisch nicht entscheidbar ist! Für zwei d-Mannigfaltigkeiten M und N (topologischer Raum, der sich lokal wie $\mathbb{R}^d$ verhält) bzw. Polyeder, die in der Form von endlichen

Simplizialkomplexen repräsentiert werden, ist es unentscheidbar, ob $M \cong N$ (für $d > 3$) ist. Es gibt auch keinen Algorithmus, der entscheiden kann, ob ein endlicher Simplizialkomplex homöomorph zu einer Mannigfaltigkeit ist [14, 16]. Trotzdem soll das Ziel sein, auch Aussagen über höherdimensionale Datensätze treffen zu können. Eine Teillösung liefern die topologischen Invarianten. So können zumindest die topologischen Eigenschaften der topologischen Räume bestimmt sowie Homöomorphismen ausgeschlossen werden.

4.1 Euler-Charakteristik

Im Kapitel 3 wurde erläutert, wie aus einer Punktwolke ein Simplizialkomplex K erzeugt werden kann. Damit ist der erste Schritt der Datenanalyse abgeschlossen (siehe Kapitel 2.1, Seite 4). Der nächste Schritt soll sich mit der Topologie von $\mathbb{X}$ auseinandersetzen. Das erste Beispiel für eine Invariante war die Homotopieäquivalenz (Def. 2.5). Leider eignet sie sich mehr für die theoretischen Untersuchungen als für die praktische Datenanalyse. Es werden alternative Invarianten benötigt. Eine Invariante, die sich sehr praktisch auf Simplizialkomplexe anwenden lässt, ist die Euler-Charakteristik [19].

Definition 4.1 (Euler-Charakteristik)**:** Sei K ein d-dimensionaler Simplizialkomplex und sei c_k die Anzahl der k-dimensionalen Simplizes in K. Die **Euler-Charakteristik** χ von K ist definiert durch

$$\chi(K) := \sum_{k=0}^{d} (-1)^k c_k.$$

Beispiel 4.1:
Für den Vietoris-Rips Komplex $V_{2\epsilon_2}$ aus Abbildung 6(b) gilt

$$\chi(V_{2\epsilon_2}) = 8 - 13 + 5 = 0.$$

Und da χ eine Invariante ist [19], haben alle zu $V_{2\epsilon_2}$ homöomorphen topologischen Räume die gleiche Charakteristik χ.

Analog wie für Simplizialkomplexe kann die Euler-Charakteristik auch für kubische Komplexe definiert werden. Sie lässt sich auch allgemein mit Hilfe von sogenannten Bettizahlen definieren (siehe Definition 4.5). Die Euler-Ckarakteristik von endlichen Polyedern ist vermutlich die älteste bekannte Invariante. χ hängt mit dem Geschlecht (Anzahl der Henkel bzw. Löcher) einer geschlossenen Fläche F zusammen und mit ihrer Hilfe ist es möglich, geschlossene Flächen komplett zu klassifizieren [15][19]. Eine geschlossene Fläche F lässt sich immer triangulieren [18]. Dann gilt für die Charakteristik χ:

$$\chi(F) = E - K + D,$$

wobei E die Anzahl der Ecken, K die Anzahl der Kanten und D die Anzahl der Dreiecke in der Triangulierung sind. Für orientierbare F gilt dann

$$\chi(F) = 2 - 2g$$

und für nicht orientierbare F gilt

$$\chi(F) = 2 - g,$$

wobei g das Geschlecht von F ist. Der Hauptsatz der Flächentopologie besagt außerdem, dass zwei geschlossene Flächen dann und nur dann homöomorph sind, wenn sie in Charakteristik und Orientierbarkeitscharakter übereinstimmen [19].

Die Euler-Charakteristik lässt sich relativ einfach und schnell für Simplizialkomplexe berechnen. Zudem lassen sich mit ihrer Hilfe schon die ersten Aussagen über Homöomorphie treffen. Die Euler-Charakteristik alleine ist nicht ausreichend, um präzise Informationen über $\mathbb{X}$ zu erhalten. Es sind weitere Invarianten notwendig. Eine klassische und sehr mächtige Invariante, die sich gut für Simplizialekomplexe berechnen lässt, ist die Homologie.

4.2 Simpliziale Homologie

Eine ausführliche Beschreibung der simplizialen Homologie ist in [10] dargelegt.

Als Nächstes soll für die Simplizialkomplexe eine neue topologische Invariante definiert werden - die Homologie. Hierzu werden die dafür notwendigen Begriffe eingeführt. Sei σ ein n-Simplex mit den Knoten $v_0, \ldots, v_n$. Für $n \geq 0$ ist die Orientierung von σ eine Äquivalenzklasse von Reihenfolgen der Knoten, mit $(v_0, v_1, \ldots, v_n) \sim (v_{\tau(0)}, v_{\tau(1)}, \ldots, v_{\tau(n)})$, wenn τ eine grade Permutation ist. Ein orientierter n-Simplex wird notiert als $[v_0, \ldots, v_n]$.

Sei K ein Simplizialkomplex. Für K können die Kettenkomplexe und eine Randabbildung definiert werden.

Definition 4.2 (Freie abelsche Gruppe)**:** Sei S eine Menge. Dann ist die **freie abelsche Gruppe**, oder das freie $\mathbb{Z}$**-Modul**, über S, definiert als

$$Z[S] := \left\{ \sum_{s \in S} n_s \cdot s \mid n_s \in \mathbb{Z}, \{n_s \neq 0\} \; endlich \right\}.$$

Definition 4.3 (Randabbildung, Kettenkomplex)**:** Sei K ein d-dimensionaler Simplizialkomplex und $C_n(K)$ die freie abelsche Gruppe, welche von den n-Simplexe von K erzeugt wird. Die Elemente $c \in C_n(K)$ heißen ***n*-Ketten** und können als Summe $c = \sum_k c_k[\sigma_k]$ mit $\sigma_k \in K,\ c_k \in \mathbb{Z}$ geschrieben werden. Für $c \in C_n$ wird die **Randabbildung** $\partial_n : C_n \to C_{n-1}$ definiert. Für ein n-Simplex in c ist sie definiert als

$$\partial_n[v_0, \ldots, v_n] = \sum_k (-1)^k [v_0, \ldots, \hat{v}_k, \ldots, v_n] = \sum_k (-1)^k [v_0, \ldots, v_{k-1}, v_{k+1}, \ldots, v_n],$$

wobei die Schreibweise $\hat{v}_k$ bedeutet, dass der Knoten v_k aus dem Simplex gelöscht wurde. Zusammen mit der Randabbildung erzeugen die n-Ketten einen **Kettenkomplex** C_*:

$$\cdots \longrightarrow C_{n+1} \xrightarrow{\partial_{n+1}} C_n \xrightarrow{\partial_n} C_{n-1} \longrightarrow \cdots \longrightarrow C_1 \xrightarrow{\partial_1} C_0 \xrightarrow{\partial_0} 0.$$

Die Randabbildung ist ein Homomorphismus und besitzt außerdem folgende wichtige Eigenschaft:

Satz 4.1: Die Komposition von aufeinander folgenden Randabbildungen ist gleich Null, d.h. $\partial_{n-1} \circ \partial_n = 0$ [10].

Beispiel 4.2:
$\partial_1 \circ \partial_2$ angewandt auf den 2-Simplex σ in der Abbildung 8 ergibt

$$(\partial_1 \circ \partial_2)[v_0, v_1, v_2] = \partial_1 \left([v_1, v_2] - [v_0, v_2] + [v_0, v_1]\right) = (v_2 - v_1) - (v_2 - v_0) + (v_1 - v_0) = 0$$

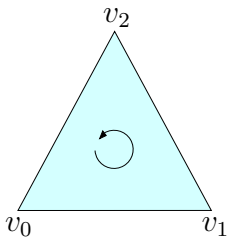


Abbildung 8: 2-Simplex σ mit der Orientierung $[v_0, v_1, v_2]$.

Der Satz 4.1 besagt, dass das Bild $\operatorname{im} \partial_{n+1}$ stets in dem Kern $\ker \partial_n$ enthalten ist. Das motiviert die nächste Definition.

Definition 4.4 (Homologie)**:** Für einen Kettenkomplex C_* wird die n-te **Homologiegruppe** als Quotient

$$H_n(K) := \ker \partial_n / \operatorname{im} \partial_{n+1}$$

definiert. Die Elemente von $\ker \partial_n$ werden ***n*-Zyklen** und die von $\operatorname{im} \partial_{n+1}$ ***n*-Rändern** genannt. Die Folge der Homologiegruppen wird als **Homologie** $H_*(K)$ des Simplizialkomplexes K bezeichnet.

Umgangssprachlich besagt die Homologie die Existenz n-dimensionale Löcher eines topologischen Raumes.

Satz 4.2: Für die Homologie eines Simplizialkomplexes K gilt [10]:

(i) Sei K eine disjunkte Vereinigung

$$K = \bigsqcup_{i \in I} K_i,$$

dann gilt für die Homologiegruppen $H_n(K)$ und für alle $n \in \mathbb{Z}$

$$H_n(K) = \bigoplus_{i \in I} H_n(K_i).$$

(ii) Für einen nicht leeren und wegzusammenhängenden Simplizialkomplex K gilt $H_0(K) \cong \mathbb{Z}$. Nach (i) zählt daher $H_0(K)$ die Wegzusammenhangskomponente.

(iii) Sind zwei Simplizialkomplexe K und K' homotopieäquivalent, dann sind deren Homologiegruppen $H_n(K)$ und $H_n(K')$ zueinander isomorph.

Der Satz 4.2(iii) besagt, dass für zwei homöomorphe Simplizialkomplexe, die damit auch homotopieäquivalent sind, deren Homologiegruppen bis auf Isomorphie gleich sind und damit stellt die Homologie eine topologische Invariante dar.

In der Definition 4.3 ist vom $\mathbb{Z}$-Modul die Rede. Homologie kann aber für beliebige R-Moduln definiert werden [13]. Die R-Moduln über einem Hauptidealbereich (zum Beispiel $\mathbb{Z}$) können komplett klassifiziert werden.

Satz 4.3: Sei R ein Hauptidealring. Ein endlich erzeugtes R-Modul M kann als direkte Summe

$$M = \bigoplus_{i=1}^{\beta_n} R \oplus \bigoplus_{j=1}^{m} R/r_j R,$$

mit $\beta_n \in \mathbb{Z}, r_j \in R, r_j | r_{j+1}$, geschrieben werden. Die Ideale $r_j R = (r_j)$ sind dann eindeutig bestimmt [11].

Die ganze Zahl β_n wird auch die n-te Bettizahl genannt. Damit kann eine alternative Definition der Euler-Charakteristik angegeben werden.

Definition 4.5 (Bettizahl, Euler-Charakteristik)**:** Sei K ein Simplizialkomplex und $H_n(K)$ seine Homologie. Der Rang von $H_n(K)$ heißt die n-te **Bettizahl**. Die Euler-Charakteristik χ von K ist dann definiert als

$$\chi(K) := \sum_{n \in \mathbb{N}} (-1)^n \beta_n.$$

Das ist eine Verallgemeinerung der Definition 4.1. Nach Definition 4.5 wird die Charakteristik $\chi(K)$ komplett durch die Homologie beschrieben.

Beispiel 4.3:
In Tabelle 1 werden die Homologiegruppen für die verschiedenen Komplexe aus Kapitel 3 berechnet. Für die aufgelisteten Strukturen gilt

$$C_{\epsilon_2} \cong A_\epsilon, \qquad C_{\epsilon_1} \simeq Q_{\epsilon_1'}, \qquad V_{2\epsilon_2} \simeq Q_{\epsilon_2'} \simeq W_{2\epsilon_2}$$

und für homöomorphe bzw. homotopisch äquivalente Räume müssen die Homologiegruppen und Charakteristik χ gleich sein, was auch mit den Ergebnissen aus Tabelle 1 übereinstimmt.

Simplizialkomplex K	$H_0(K)$	$H_1(K)$	$H_2(K)$	χ
Punkt p	$\mathbb{Z}$	0	0	1
C_{ϵ_1}	$\mathbb{Z}^5$	0	0	5
C_{ϵ_2}	$\mathbb{Z}$	$\mathbb{Z} \oplus \mathbb{Z}$	0	-1
A_ϵ	$\mathbb{Z}$	$\mathbb{Z} \oplus \mathbb{Z}$	0	-1
$V_{2\epsilon_2}$	$\mathbb{Z}$	$\mathbb{Z}$	0	0

Tabelle 1: Homologie und Euler-Charakteristik für die in Kapitel 3 gezeigten Simplizialkomplexe.

Sei K ein Simplizialkomplex. Dann kann die n-te Randabbildung ∂_n in Form einer **Randmatrix** dargestellt werden; mit n-Simplexe repräsentiert als Spalten und $(n-1)$-Simplexe als Zeilen. Für eine fixe Reihenfolge der Simplexe gilt für die Matrix $\partial_n = [c_i^j]$, dass $c_i^j = 1$, wenn für das i-te $(n-1)$-Simplex gilt, dass er eine Teilmenge vom j-te n-Simplex ist, also $\sigma_{n-1}^i \subset \sigma_n^j$. Sonst ist $c_i^j = 0$. Damit gilt

$$\partial_n c = \begin{bmatrix} c_1^1 & c_1^2 & \dots & c_1^{m_n} \\ c_2^1 & c_2^2 & \dots & c_2^{m_n} \\ \vdots & \vdots & \ddots & \vdots \\ c_{m_{n-1}}^1 & c_{m_{n-1}}^2 & \dots & c_{m_{n-1}}^{m_n} \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_{m_n} \end{bmatrix}.$$

Es gibt Algorithmen, um eine Randmatrix in die Smith-Normalform zu bringen und um die Homologiegruppen zu bestimmen [8]. Für einen Simplizialkomplex (und damit auch für die algorithmischen Strukturen aus Kapitel 3) kann demzufolge eine Homologie algorithmisch berechnet werden.

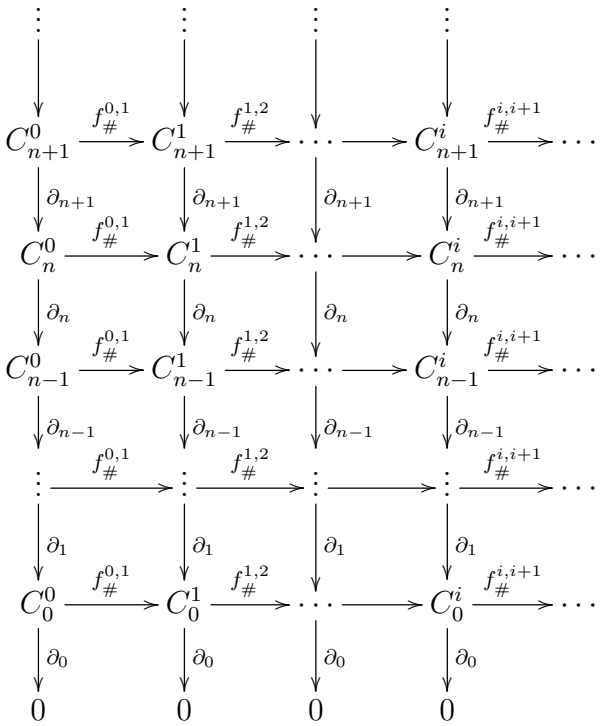
4.3 Persistente Homologie

Bei der Betrachtung von C_{ϵ_1} und C_{ϵ_2} (siehe Abbildung 4 und Tabelle 1) wird schnell klar, dass die Änderung des Parameters ϵ zu unterschiedlichen Ergebnissen bei der Berechnung der Homologie führen kann. Ein falsch gewähltes ϵ kann ein irreführendes Resultat liefern. Sinnvoller erscheint die Berechnung der Homologie für verschiedenen ϵ und die Betrachtung ihrer Entwicklung.

Ein weiterer Aspekt bei der Analyse von Punktwolken ist, dass so ein Datensatz wegen verschiedener Faktoren Rauschen beinhalten kann. Durch das Variieren von ϵ soll das Rauschen aus dem Datensatz herausgefiltert werden. Das ist die Idee, die sich hinter der Persistenten Homologie verbirgt [5, 8, 25, 26].

Definition 4.6 (Filtration)**:** Sei K ein Simplizialkomplex und $f : K \to \mathbb{R}$. Sei außerdem f monoton, d. h. wenn $\tau, \sigma \in K$ und $\tau \subseteq \sigma$, dann $f(\tau) \leq f(\sigma)$. Seien $a_1 < a_2 < \ldots < a_n$ die Funktionswerte der Simplexe in K, $a_0 = -\infty$ und $K_i = K(a_i) = f^{-1}(-\infty, a_i]$. Eine Sequenz der Komplexe $\emptyset = K_0 \subseteq K_1 \subseteq \ldots \subseteq K_n = K$ wird **Filtration** genannt.

Alle im Kapitel 3 vorgestellten Komplexe, mit Ausnahme der kubischen Komplexe, besitzen für ein wachsendes ϵ eine Filtration [25]. Sei $C_n^i := C_n(K_i)$. Für einen gefilterten Komplex und $i < j$ existieren die Inklusionabbildung $f^{i,j} : K_i \hookrightarrow K_j$. Nach Satz 4.2 induziert $f^{i,j}$ ein Homomorphismus $f_n^{i,j} : H_n(K_i) \to H_n(K_j)$ sowie folgendes Kettennetz:



Für einen gefilterten Komplex existiert damit eine Sequenz

$$0 = H_n(K_0) \xrightarrow{f_n^{0,1}} H_n(K_1) \xrightarrow{f_n^{1,2}} \cdots \longrightarrow H_n(K_{n-1}) \xrightarrow{f_n^{n-1,n}} H_n(K_n) = H_n(K).$$

Bei dem Übergang von K_i zu K_{i+1} können neue Homologieklassen entstehen oder alte verschwinden bzw. zusammenschmelzen. Diese Änderungen stellen den Mittelpunkt der Forschung auf dem Gebiet der Persistenten Homologie dar.

Definition 4.7 (Persistente Homologie)**:** Für $0 \leq i \leq j \leq m$ sei $f_n^{i,j}$ der Homomorphismus zwischen $H_n(K_i)$ und $H_n(K_j)$.

$$H_n^{i,j} := \operatorname{im} f_n^{i,j}$$

wird bezeichnet als die ***n*-te persistente Homologiegruppe**. Die dazugehörenden **Bettizahlen** sind der Rang der persistenten Homologiegruppen, d.h. $\beta_n^{i,j} := \operatorname{rank} H_n^{i,j}$.

Die Persistente Homologie beschreibt demnach die Beziehung zwischen Homologiegruppen einer Filtration. Es ist $H_n^{i,i} = H_n(K_i)$. Außerdem gilt für die Persistente Homologie: $H_n^{i,j} \simeq \ker \partial_n(K_i) \;/\; (\operatorname{im} \partial_{n+1}(K_j) \cap \ker \partial_n(K_i))$ [26]. Anders ausgedrückt beinhaltet die Persistente Homologie die Homologieklasse von K_i, welche in K_j weiter vorhanden ist. Um die persistente Homologie besser zu verstehen und zu analysieren, werden sogenannte Persistenz-Diagramme eingesetzt. Dabei werden die Bettizahlen der persistenten Homologiegruppen in der erweiterten euklidischen Ebene $\overline{\mathbb{R}}^2$, mit $\overline{\mathbb{R}} = \mathbb{R} \cup \{+\infty, -\infty\}$, dargestellt.

Definition 4.8 (Persistenz-Diagramm)**:** Sei $\gamma \in H_n(K_i)$. Wenn $\gamma \notin H_n^{i-1,i}$ ist, dann bedeutet dies, dass die Klasse γ in K_i geboren ist. Überdies gilt, dass γ in K_j verschwindet, wenn $f_n^{i,j-1}(\gamma) \notin H_n^{i-1,j-1}$ und $f_n^{i,j}(\gamma) \in H_n^{i-1,j}$ gilt, d. h. γ ist mit einer anderen Klasse verschmolzen. Sei $\mu_n^{i,j}$ die Anzahl der n-dimensionalen Klassen geboren in K_i und gestorben in K_j. Es gilt

$$\mu_n^{i,j} := (\beta_n^{i,j-1} - \beta_n^{i,j}) - (\beta_n^{i-1,j-1} - \beta_n^{i-1,j})$$

, wobei für alle n gilt $i < j$. Die erste Differenz auf der rechten Seite zählt die Anzahl der Klassen, die in K_i oder früher geboren sind und in K_j sterben. Die zweite Differenz erzeugt das gleiche für K_{i-1} und K_j.

Das **n-te Persistenz-Diagramm** $D_n(f) \subset \overline{\mathbb{R}}^2$ einer Filtration ist die Menge der Punkte (a_i, a_j) zusammen mit der **Vielfachheit** $\mu_n^{i,j}$. Der Ausdruck $pers(\gamma) = a_j - a_i$ wird als **Persistenz** bezeichnet und $j - i$ ist der dazugehörige **Persistenz-Index**.

Werden die Punkte (a_i, a_j) in einem Koordinatensystem abgebildet, dann liegen alle Punkte auf oder über der Winkelhalbierenden. Aus einem Persistenz-Diagramm können die Bettizahlen sofort abgelesen werden: $\beta_n^{k,l}$ ist die Anzahl der Punkte im oberen linken Quadranten $(-\infty; a_j] \times (a_k; \infty]$. Homologieklassen, die in K_i geboren werden und in K_l sterben, werden nur gezählt, wenn gilt: $a_i \leq a_k$ und $a_j > a_l$. Damit ist die horizontale Seite von $(-\infty; a_j] \times (a_k; \infty]$ offen und die vertikale abgeschlossen. Dadurch beschreibt das Persistenz-Diagramm $D_n(f)$ die persistente Homologie vollständig. Genauer gesagt:

Satz 4.4: Sei $\emptyset = K_0 \subseteq K_1 \subseteq \ldots \subseteq K_m = K$ eine Filtration. Für $0 \leq k \leq l \leq m$ und die Bettizahl $\beta_n^{k,l}$ der n-ten persistenten Homologiegruppe gilt: [8]

$$\beta_n^{k,l} = \sum_{j>l} \sum_{i \leq k} \mu_n^{i,j}.$$

Die Analyse von Persistenz-Diagrammen stellt einen wichtigen Bestandteil der Topologischen Datenanalyse dar und ist sinnvoll, um Informationen über den Raum $\mathbb{X}$ zu bekommen.

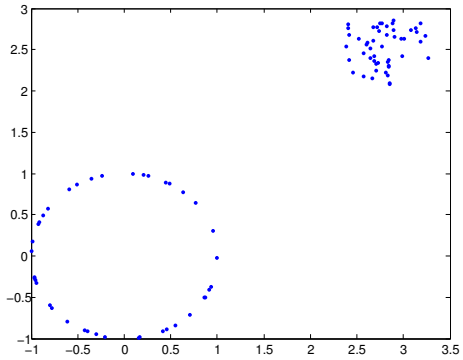
Beispiel 4.4:
Bei großen Datensätzen ist die Berechnung der persistenten Homologie sehr rechenintensiv. Dazu wird in diesem Beitrag die Open-Source-Software JavaPlex [1] genutzt. Die Software erlaubt, die Persistente Homologie zu bestimmen und beinhaltet viele Routinen zu topologischen Berechnungen.

Sei $\mathbb{X} = S^1 \sqcup B_{0.5}(x)$. Mit der persistenten Homologie soll die Homologie von $\mathbb{X}$ rekonstruiert werden. Die vorgegebene Punktwolke T entsteht aus 50 Punkten, gezogen aus einer Kugel, und 40 Punkten gezogen aus einem Kreis (Abbildung 9(a)). Für $\mathbb{X} = S^1 \sqcup B_{0.5}(x)$ gilt

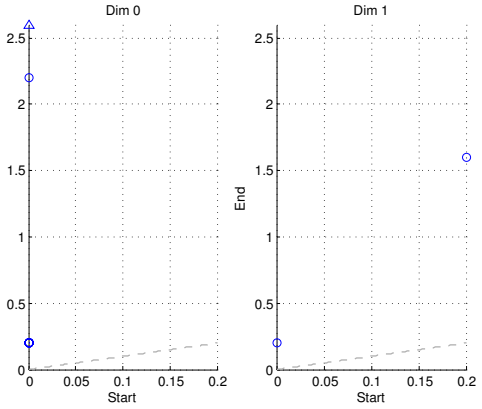
$$H_n(S^1 \sqcup B_{0.5}(x)) = \begin{cases} \mathbb{Z} \oplus \mathbb{Z}, & n = 0, \\ \mathbb{Z}, & n = 1. \end{cases}$$

Um die Homologie von $\mathbb{X}$ zu bestimmen, wurden Vietoris-Rips-Komplexe benutzt und für die Berechnungen wurde die Filtration $V_{0.2} \subseteq V_{0.4} \subseteq \ldots \subseteq V_{3.8} \subseteq V_4$ eingesetzt. Die

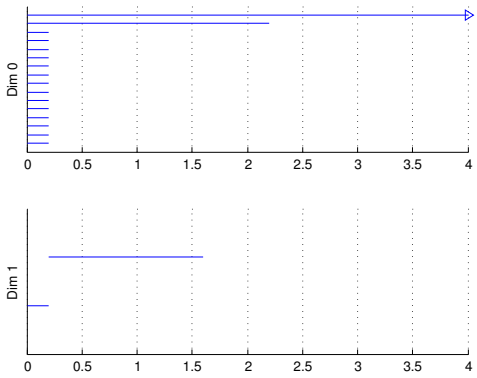
Ergebnisse wurden graphisch dargestellt. Ein Persistenz-Diagramm kann direkt in einem Koordinatensystem abgebildet werden (siehe Abbildung 9(b)). Auf der Abszisse ist die Zeit der Geburten und auf der Ordinate der Zeitpunkt, an dem die Homologieklassen verschwinden, abgebildet. Ein Dreieck in dem Diagramm bedeutet, dass die Homologieklassen nie verschwanden. Die zweite graphische Methode für die Darstellung von $D_n(f)$ ist der sogenannte Barcode (siehe Abbildung 9(c)). Dabei werden alle Homologieklassen untereinander abgebildet. Die x-Achse entspricht dem Zeitverlauf vom Radius ϵ. Die Pfeile symbolisieren das Bestehen einer Klasse bis zum Ende. Im Barcode ist die Vielfachheit direkt ablesbar.



(a) Punktwolke T



(b) $D_n(f)$ in dem Koordinatensystem



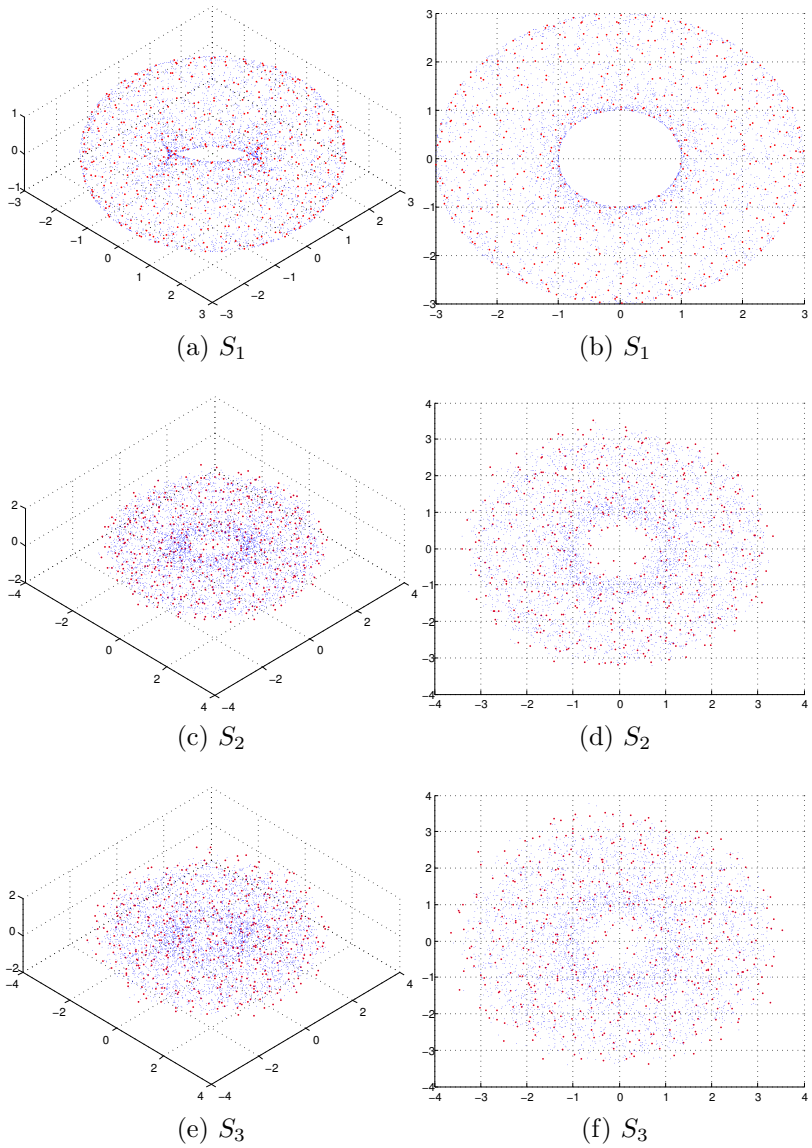
(c) Barcodes von $D_n(f)$

Abbildung 9: Datensatz T und die dazugehörige persistente Homologie

Die Betrachtung von Persistenz-Diagrammen liefert nützliche Informationen. Zu Beginn werden sehr viele Homologiegruppen in der Dimension 0 geboren (siehe Abbildung 9(c)). Für kleine ϵ ist das ein normales Verhalten, da $H_0(K)$ wegzusammenhängende Komponenten zählt. Die meisten davon verschwinden sehr schnell und für lange Zeit bleiben nur noch zwei Klassen erhalten. Irgendwann verschmelzen auch die beiden Klassen. Der Verlauf deutet auf die Existenz von zwei Komponenten in K hin, was sich auch mit $H_0(S^1 \sqcup B_{0.5}(x)) = \mathbb{Z} \times \mathbb{Z}$ überdeckt. Im Diagramm $D_1(f)$ ist auch die Gruppe $H_1(S^1 \sqcup B_{0.5}(x)) = \mathbb{Z}$ zu finden. Zuerst wird ein kleines eindimensionales Loch im Datensatz T erkannt. Die dazugehörende Homologiegruppe verschwindt genau so schnell, wie sie entstanden ist. Das Verhalten deutet auf ein Rauschen in den Daten. Später

entsteht noch eine Homologiegruppe in $H_1(K)$ und bleibt lange am Leben. Damit liefert die persistente Homologie von T die erwarteten Ergebnisse.

Beispiel 4.5:
Die Punktwolke S_1 entsteht aus 5000 Punkten, gezogen aus einem Torus (siehe Abbildung 10(a) und (b)). In dem Datensatz soll ein weißes Rauschen simuliert werden. Das geschieht, indem zu allen drei Koordinaten eine normalverteilte Zufallszahl addiert wird, d.h. $s = (x + \mathcal{N}(0;\sigma^2), y + \mathcal{N}(0;\sigma^2), z + \mathcal{N}(0;\sigma^2))^T$, $s \in S_1$. Damit entstehen $S_2 = S_1 + \mathcal{N}(0; 0.2^2)$ und $S_3 = S_1 + \mathcal{N}(0; 0.3^2)$ (siehe Abbildung 10(c),(d), (e) und (f)).



(a) S_1 (b) S_1

(c) S_2 (d) S_2

(e) S_3 (f) S_3

Abbildung 10: Drei Datensätze gezogen aus einem Torus mit eingebautem Rauschen. Die roten Punkte sind die Landmarken und die blauen die Witnesses.

Wegen der großen Menge an Punkten in den Datensätzen werden zur Berechnung der persistenten Homologie statt den rechenintensiven Vietoris-Rips-Komplexen die

Witness-Komplexe benutzt. Aus den vorgegebenen Mengen werden zufällig 500 Landmarken L gewählt (rote Punkte in der Abbildung 10). Sei $\epsilon = \max(w, l),\ w \in W,\ l \in L$, also die größte Distanz zwischen einem Witness und einer Landmarke und $\delta = \frac{3}{4}\epsilon$, dann hat die Filtration folgende Form: $W_{\frac{1}{10}\delta} \subseteq W_{\frac{2}{10}\delta} \subseteq \ldots \subseteq W_{\frac{9}{10}\delta} \subseteq W_\delta$. Die Homologie des Torus T lautet

$$H_n(T) = \begin{cases} \mathbb{Z}, & n = 0, \\ \mathbb{Z} \oplus \mathbb{Z}, & n = 1, \\ \mathbb{Z}, & n = 2. \end{cases}$$

Für S_1 findet die Persistente Homologie alle Homologiegruppen des Torus problemlos (siehe Abbildung 11). Auch im verrauschten Datensatz S_2 ist die erwartete Homologie zu finden (siehe Abbildung 12). In diesem Fall erscheint es aber weniger eindeutig, dass es tatsächlich die Homologie des Torus ist. Die zwei größten Klassen in Dimension eins verschmelzen und in Dimension zwei einsteht am Ende eine Klasse zu viel. Die Punktwolke S_3 ist zu stark verrauscht (siehe Abbildung 13). Die Homologie des Torus wird nicht mehr erkennbar.

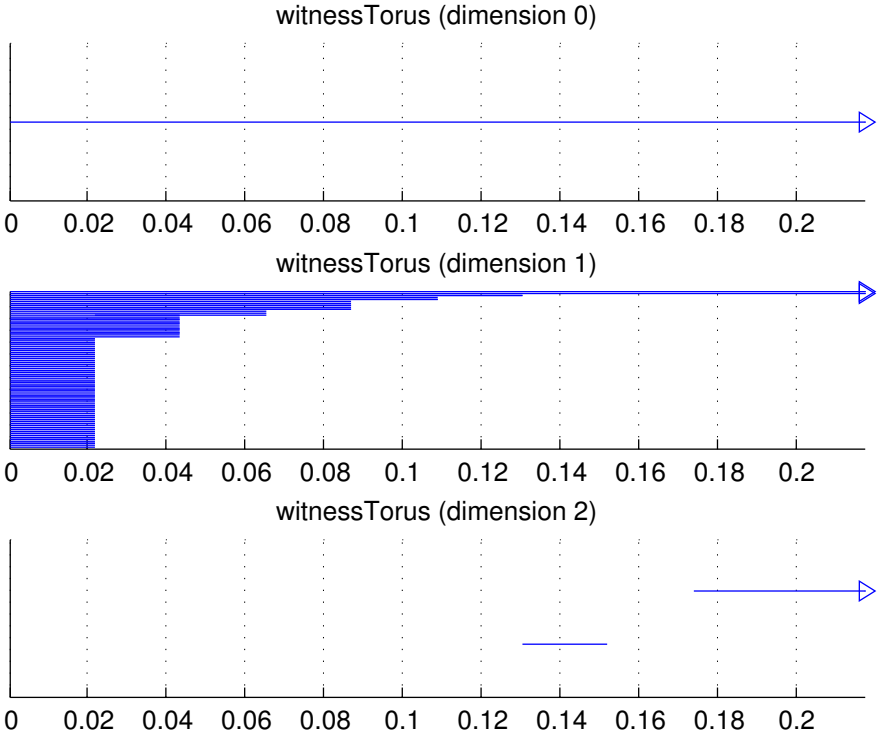


Abbildung 11: Persistente Homologie von S_1

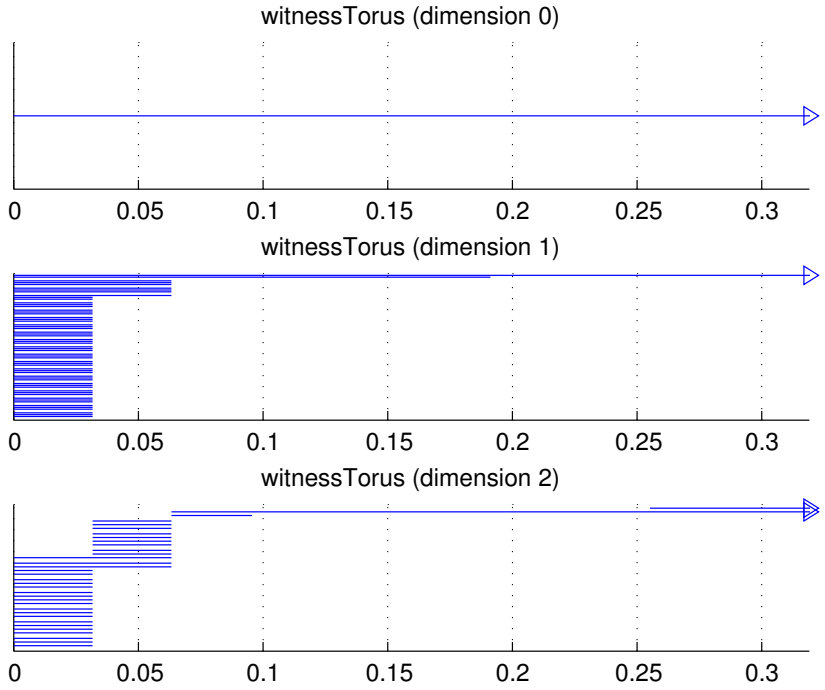


Abbildung 12: Persistente Homologie von S_2

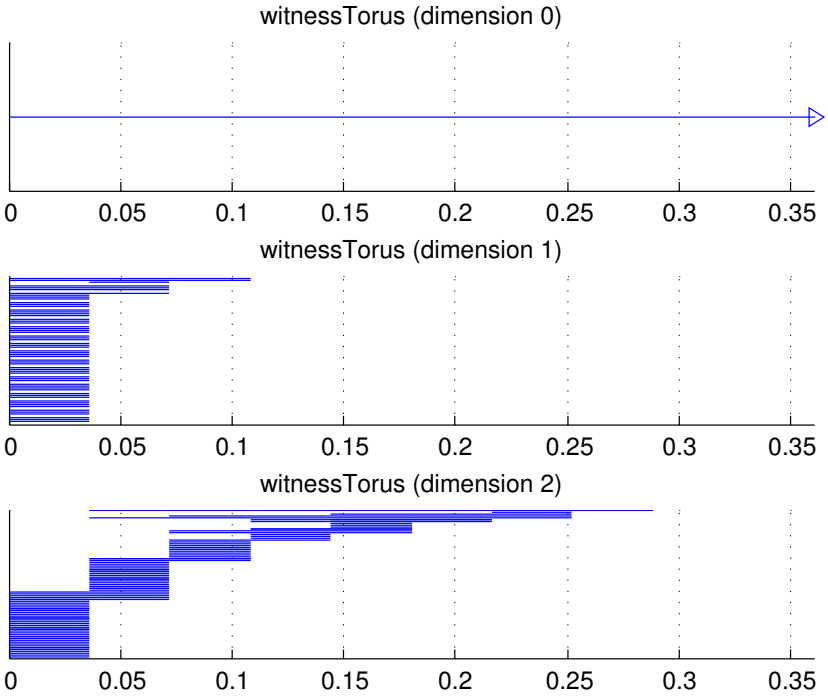


Abbildung 13: Persistente Homologie von S_3

Um zwei Persistenz-Diagramme zu vergleichen und die Ähnlichkeit bzw. Unterschiede zwischen diesen zu messen, wird ein Maß benötigt. Das führt zur sogenannten Bottleneck-Distanz. Ein Persistenz-Diagramm beinhaltet endlich viele Punkte, die über der Diago-

nalen liegen. Zu dieser Menge werden unendlich viele Punkte auf der Diagonalen hinzugefügt. Um die Distanz zwischen zwei Persistenz-Diagrammen X und Y zu messen, wird das Infimum über alle Bijektionen $\eta : X \to Y$ betrachtet.

Definition 4.9 (Bottleneck-Distanz)**:** Seien X und Y zwei Persistenz-Diagramme. Für zwei Punkte $x = (x_1, x_2)$ und $y = (y_1, y_2)$ sei $||x - y||_\infty = max\{|x_1 - y_2|, |x_2 - y_2|\}$. Die **Bottleneck-Distanz** ist definiert als

$$W_\infty(X, Y) = \inf_{\eta: X \to Y} \sup_{x \in X} ||x - \eta(x)||_\infty.$$

Es gilt für drei Persistenz Diagramme X, Y, Z, dass

1. $W_\infty(X, Y) = 0 \Leftrightarrow X = Y$,
2. $W_\infty(X, Y) = W_\infty(Y, X)$,
3. $W_\infty(X, Z) \leq W_\infty(X, Y) + W_\infty(Y, Z)$,

womit die Bottleneck-Distanz alle Metrik-Axiome erfüllt [8]. Sei K ein Simplizialkomplex und $f, g : K \to \mathbb{R}$ zwei monotone Funktionen. Für die Persistenz-Diagramme $X = D_n(f)$ und $Y = D_n(g)$ gilt $W_\infty(X, Y) \leq ||f - g||_\infty$. Damit bleiben die Persistenz-Diagramme unter Einfluss von kleinen Störungen stabil. Persistenz liefert für ähnliche Funktionen ähnliche Resultate und ist damit ein nützliches und gutes Mittel zum Messen von topologischen Eigenschaften eines Datensatzes.

5 Mapper

Im Folgenden wird der Mapper-Algorithmus als zweites wichtiges Konzept der Topologischen Datenanalyse vorgestellt. Der Mapper wandelt höherdimensionale Datensätze in Simplizialekomplexe um und kann dadurch geometrische und topologische Eigenschaften der Daten bestimmen [6]. Des Weiteren ist der Mapper ein brauchbares Werkzeug für Visualisierungen. Mit deren Hilfe können mehrdimensionale Daten grafisch veranschaulicht werden, woran zum Beispiel statistische Methoden scheitern.

5.1 Topologischer und statistischer Mapper

Sei $\mathbb{X}$ ein topologischer Raum und $f : \mathbb{X} \to \mathbb{Y}$ eine stetige Abbildung eines bekannten topologischen Raumes $\mathbb{Y}$. Mit f und $\mathbb{Y}$ sollen die geometrischen und topologischen Eigenschaften von $\mathbb{X}$ rekonstruiert werden. Genauer gesagt werden die Urbilder der offenen Überdeckung von $\mathbb{Y}$ und dessen Nerv benutzt, um Informationen über $\mathbb{X}$ zu erhalten.

Als Motivation seien $\mathbb{X} = S^1$ und $\mathbb{Y} = [-1, 1]$. Außerdem wird für $f : S^1 \to [-1, 1]$ die Funktion genommen, die Kreispunkte auf ihrer y-Koordinate projiziert, d.h. es gilt $f(x, y) = y$. Die Überdeckung von $\mathbb{Y}$ ist dann

$$V_1 = [-1, -0.33), V_2 = (0.33, 1], V_3 = (-0.5, 0.5).$$

Da f stetig ist, induziert dies die folgende

$$\mathcal{U} = \{U_1 = f^{-1}(V_1), U_2 = f^{-1}(V_2), U_3 = f^{-1}(V_3)\}$$

von $\mathbb{X}$ (siehe Abbildung 14).

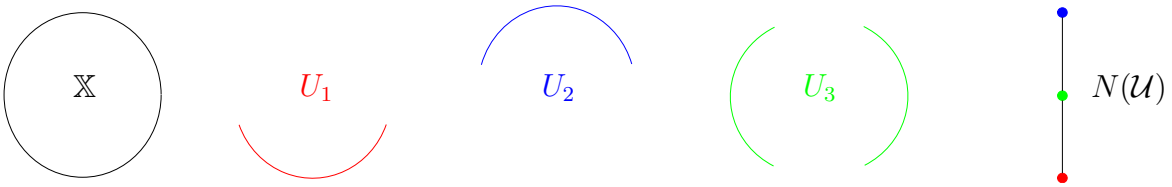


Abbildung 14: Überdeckung $\mathcal{U}$ von S^1 und dessen Nerv

Der Nerv $N(\mathcal{U})$ hat einen anderen Homotopietyp als S^1. Damit ist $N(\mathcal{U})$ eine unpräzise Darstellung für $\mathbb{X}$. Sie kann verbessert werden, indem U_3 weiter in seine Wegzusammenhangskomponente geteilt wird. Es entsteht eine neue Überdeckung $\mathcal{U}^{wz} = \{U_1, U_2, U_{3_1}, U_{3_2}\}$ (siehe Abbildung 15).

Die Idee, die hinter dem Mapper steht, ergibt sich aus der Beobachtung, dass die Filterfunktion $f : \mathbb{X} \to \mathbb{Y}$ eine Überdeckung von $\mathbb{X}$ liefert, die wiederum in seine Wegzusammenhangskomponente zerlegt werden kann. Die Nerven dieser Überdeckung werden benutzt, um $\mathbb{X}$ zu analysieren.

Der Mapper soll zur Analyse von großen und komplexen Datenmengen angewendet werden. Die grade vorgestellte topologische Version eignet sich jedoch nicht dafür. Das Schema lässt sich aber auf eine Punktwolke S übertragen. Dabei werden die Wegzusammenhangskomponente mit Clustern von Punkten ersetzt. Eine Filterfunkion liefert eine

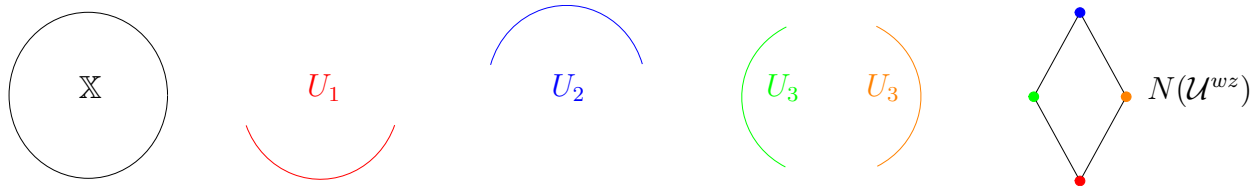


Abbildung 15: Überdeckung $\mathcal{U}^{wz}$ von S^1 und dessen Nerv

Überdeckung von S. Mit der Clusteranalyse werden anschließend Cluster gefunden, um die neue Überdeckung $\mathcal{U}^C$ zu erhalten (siehe Abbildung 16).

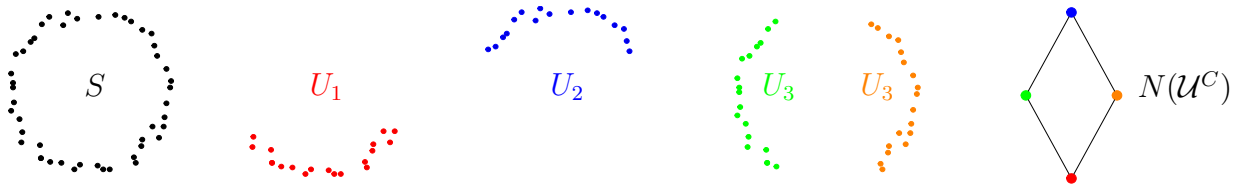


Abbildung 16: Überdeckung $\mathcal{U}^C$ von S und deren Nerv

An dieser Stelle erfolgt eine Formalisierung der topologischen und statistischen Mapper. Der Ansatz geht auf Carlsson zurück [6][21].

Definition 5.1 (topologischer Mapper)**:** Seien $\mathbb{X}$ und $\mathbb{Y}$ topologische Räume und $f : \mathbb{X} \to \mathbb{Y}$ eine stetige Abbildung. $\mathbb{Y}$, zusammen mit seiner Überdeckung $\mathcal{V} = \{V_i\}_{i \in I}$, heißt **Parameterraum** und die Abbildung f wird **Filterfunktion** genannt. f induziert eine Überdeckung $\mathcal{U} = \{U_i\}_{i \in I}$ von $\mathbb{X}$. Sei n_i hier die Anzahl der Wegzusammenhangskomponente von U_i und $U_{i,1}, \ldots, U_{i,n_i}$ die Wegzusammenhangskomponente von U_i, dann ist $\mathcal{U}^{wz}$ definiert als

$$\mathcal{U}^{wz} := f^{-1}(\mathcal{V})^{wz} = \{U_{i,k} \mid i \in I \text{ und } k = 1, \ldots, n_i\}.$$

Der Nerv $N(\mathcal{U}^{wz}) = N(f^{-1}(\mathcal{V})^{wz})$ ist der **Output des Mapper-Algorithmus**. Für den Raum $\mathbb{X}$, Filter f und die Überdeckung $\mathcal{V}$ wird die Notation $Mapper(f^{-1}(\mathcal{V}))$ benutzt.

Die Anwendung von topologischen Mappern auf eine diskrete Punktwolke S ist problematisch, da jedes Element von S eine Wegzusammenhangskomponente darstellt und damit wird der Ansatz aus 15 nicht funktionieren. Das Konzept aus Definition 5.1 muss angepasst werden. Analog zu Wegzusammenhangskomponente in der Überdeckung von $\mathbb{X}$ werden aus den ähnlichen Elementen von S Cluster erzeugt. Dieses Vorgehen wird auch als statistischer Mapper bezeichnet. Hierfür bieten sich verschiedene Verfahren der Clusteranalyse an (siehe z.B. [2]). Der Sinn der Clusteranalyse ist, mit Hilfe der Informationen aus einem Datensatz S Gruppen von Daten mit ähnlichen Eigenschaften zu finden. Die Clusteranalyse unterteilt einen Datensatz in homogene Gruppen und liefert damit eine Klassifikation einzelner Elemente. Dadurch können Aussagen über die Struktur der Daten gemacht werden.

Definition 5.2 (statistischer Mapper)**:** Sei $\mathbb{X}$ ein metrischer Raum, $S \subseteq \mathbb{X}$ eine Punktwolke und $f : \mathbb{X} \to \mathbb{R}^n$. Sei außerdem $\mathcal{V} = \{V_i\}_{i \in I}$ eine Überdeckung von $\mathbb{R}^n$. U_i ist

dann definiert als $U_i := \{x \in S \mid f(x) \in V_i\}$. Sei jetzt n_i die Anzahl der Cluster von U_i, gefunden mit einem Clusteralgorithmus C, und $U_{i,1}, \ldots, U_{i,n_j}$ die Cluster in U_i, dann ist $\mathcal{U}^C$ definiert als

$$\mathcal{U}^C := f^{-1}(\mathcal{V})^C = \{U_{i,k} \mid i \in I \text{ und } k = 1, \ldots, n_i\}.$$

Der Nerv $N(\mathcal{U}^C) = N(f^{-1}(\mathcal{V})^C)$ ist der **Output des statistischen Mapper-Algorithmus**. Für die Punktwolke S, Filter f, Clusteralgorithmus C und die Überdeckung $\mathcal{V}$ wird die Notation $Mapper(f^{-1}(\mathcal{V}), C)$ benutzt.

Beispiel 5.1:
Abbildung 15 demonstriert die Funktionsweise des topologischen Mappers und Abbildung 16 analog die Funktionsweise des statistischen Mappers.

Beispiel 5.2:
Sei S eine Menge von standardnormalverteilten Punkten. Als Filter $f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ wird die standardisierte Gauß-Kurve genutzt. Außerdem wird folgende Überdeckung $\mathcal{V} = \{[0, 0.125], [0.075, 0.225], [0.175, 0.325], [0.275, 0.4]\}$ für $\mathbb{R}$ gewählt. Da Außerhalb vom Bereich $[0, 0.4]$ keine Punkte liegen, muss nicht der komplette Raum $\mathbb{R}$ überdeckt werden. Abbildung 17 zeigt den erwarteten Output des Mappers bei einem großen Datensatz S. In einem Output des Mappers werden noch folgende nützliche Informationen kodiert: Die Größe eines Knotens entspricht der Anzahl der Punkte in einem Cluster. Die Farben sollen mit den Werten der Funktion $f(x)$ korrespondieren (rot - hohe Dichte, blau - geringe Dichte). Um die Farben zu bestimmen, kann zum Beispiel die durchschnittliche Dichte aller Punkte herangezogen werden.

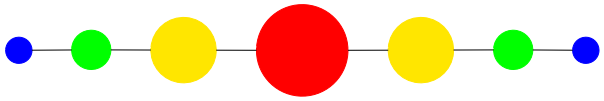


Abbildung 17: Output des Mappers aus Beispiel 5.2

Es wird deutlich, dass der Output des Mappers stark von der Wahl der Filterfunktion und der Überdeckung von $\mathbb{R}^n$ abhängt. Die in der Realität nützlichen Filter basieren auf der Möglichkeit eine Distanz zwischen Punkten in S zu berechnen. Aus diesem Grund wird vorausgesetzt, dass $\mathbb{X}$ ein metrischer Raum ist, d.h. eine Metrik $d : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$ besitzt. Als nächstes werden einige in der Praxis gebräuchliche Filter, die auch wichtige geometrische Informationen über den Datensatz liefern, vorgestellt.

5.2 Filter

Dichte

Dichteschätzer messen, wie nah ein Punkt aus S an anderen Punkten liegt. In der Statistik spielen Dichterschätzer eine große Rolle. Im Folgenden werden die gebräuchlichsten Dichteschätzer kurz skizziert. Dabei sind die Kerndichteschätzer hervorzuheben. Für

Mapper wird üblicherweise der Gaußkern genutzt, definiert als

$$dens_\epsilon(x) = C_\epsilon \sum_{y \in S} \exp\left(\frac{-d(x,y)^2}{\epsilon}\right),$$

mit $x, y \in S$, $\epsilon > 0$ und C_ϵ einer Konstante, sodass $\int_{-\infty}^{\infty} dens_\epsilon(x)dx = 1$. Dabei steuert ϵ die „Glattheit“ der Funktion (je größer desto glatter). Ein Anderer wichtiger Dichteschätzer ist der sogenannte k-Nächster Nachbarn-Schätzer:

$$dens_k(x) = \frac{k}{2Nd_k(x)},$$

wobei N die Anzahl der Punkte (Beobachtungen) im Datensatz S und $d_k(x)$ die Distanz zu den k-ten Nachbarn ist.

Exzentrizität

Die Exzentrizität ist ein weiterer Filter zum Messen von geometrischen Eigenschaften einer Punktwolke. Dabei sollen die Punkte identifiziert werden, die weit von der Mitte eines Datensatzes liegen, ohne den Mittelpunkt selbst zu kennen. Für die Exzentrizität gilt:

$$E_p(x) = \left(\frac{\sum_{y \in S} d(x,y)^p}{N}\right)^{1/p},$$

mit $1 \leq p < \infty$. Die Definition kann auch für $p = \infty$ wie folgt erweitert werden:

$$E_\infty(x) = \max_{y \in S} d(x,y).$$

Dadurch werden den Punkten, die weit vom Zentrum liegen, größere Werte zugeordnet als den Punkten nahe der Mitte.

Gemischte Filter

Als letztes werden gemischte Filter betrachtet. Die Dichte sowie die Exzentrizität liefern Informationen über einen Datensatz. Oft ist sinnvoll, nicht nur einen Filter davon zu wählen. Als geeigneter Filter kann das Paar $(dens_\epsilon, E_p) : \mathbb{X} \to \mathbb{R}^2$ genommen werden. Jedoch können auch andere Filterfunktionen f_i gewählt werden, die interessanten Eigenschaften des Datensatzes messen, sodass am Ende ein n-Tupel Filter $(f_1, \ldots, f_n) : \mathbb{X} \to \mathbb{R}^n$ entsteht.

Weiterhin ist es möglich, verschiedene Filter mit arithmetischen Operationen zu verknüpfen. Als Beispiel kann eine Addition erwähnt werden, wodurch der Filter

$$f_+ = dens_\epsilon(x) + E_p(x)$$

entsteht. Das Gleiche ist auch mit $-, \cdot, \div$ sowie mit mehreren Funktionen möglich. Damit können Einflüsse verschiedener Filter aufeinander untersucht werden. Beispielweise kann der Datensatz in Bezug auf unterschiedliche Kombinationen von hoher/niedriger Exzentrizität und hoher/geringer Dichte gefiltert werden. Bei normalverteilten Werten sind Dichte und Exzentrizität negativ miteinander korreliert. Derartige Filter kommt eine wichtige Bedeutung bei der Analyse bestimmter Daten bei.

5.3 Mapper-Algorithmus

Für den statistischen Mapper ist zusätzlich noch eine Überdeckung von $\mathbb{R}^m$ notwendig. Für $\mathbb{R}$ wird diese normalerweise auf folgende Weise konstruiert: Da S endlich viele Punkte besitzt, reicht es aus, nur das Intervall $[a, b] = [\min f, \max f] \subseteq \mathbb{R}$ zu betrachten. Die Überdeckung entsteht aus n Intervallen mit der Länge $l = \frac{b-a}{n}$ und der prozentualen Überlappung p. Diese konstruierte Überdeckung wird mit $\mathcal{V}(n, p)$ bezeichnet.

Beispiel 5.3:
Sei $[a, b] = [0, 2],\ n = 4,\ p = 0.5,\ l = 0.5$. Für die Überdeckung von $[0, 2]$ gilt damit $\mathcal{V}(4, 0.5) = \{[0, 0.75], [0.25, 1.25], [0.75, 1.75], [1.25, 2]\}$.

Für die Überdeckung von $\prod_{i=1}^{m}[a_i, b_i] \subseteq \mathbb{R}^m$ wird eine ähnliche Konstruktion benutzt. Sei $\mathcal{V}^i(n_i, p_i) = \{V_1^i, \ldots, V_{n_i}^i\}$ eine Überdeckung von $[a_i, b_i]$. Die Überdeckung von $\prod_{i=1}^{m}[a_i, b_i]$ wird dann definiert als

$$\left\{V_{k_1}^1 \times \ldots \times V_{k_i}^i \times \ldots \times V_{k_m}^m \mid k_{1,\ldots,m} = 1, \ldots, n_i\right\}.$$

Ein Filter f und eine Überdeckung $[a, b]$ reichen schon, um einen Datensatz S mit dem Mapper untersuchen zu können. Eine solche Analyse verläuft normalerweise in folgenden Schritten ab:

1. Ein Filter f wird gewählt und die zugehörigen Werte $f(x)$ werden für alle $x \in S$ berechnet.
2. Für $[a, b] = [\min f, \max f]$ werden die Werte n und p ausgewählt und die dazugehörige Überdeckung $\mathcal{V}(n, p)$ wird bestimmt.
3. Für alle $V \in \mathcal{V}(n, p)$ werden alle Werte x bestimmt, sodass $f(x) \in V$ gilt. Finde mit einem Algorithmus C die Cluster und interpretiere jedes Cluster als ein 0-Simplex. Bestimme $\mathcal{U}^C = f^{-1}(\mathcal{V}(n, p)^C)$.
4. Bestimme $Mapper(f^{-1}(\mathcal{V}(n, p), C)$. Interpretiere das Ergebnis.
5. Sollte der Output zu ungenau oder zu klumpig sein, können die Schritte 2 bis 5 mit verschiedenen p und n wiederholt werden.

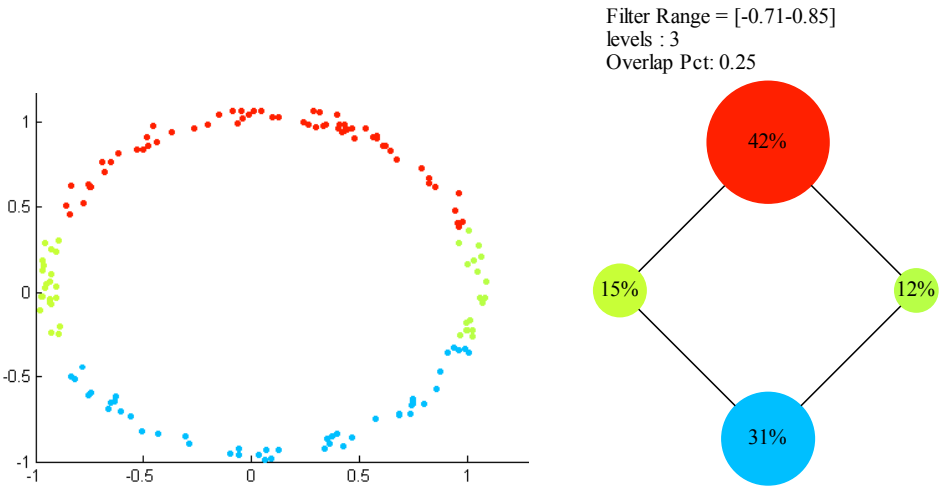
Kleine p bedeuten weniger Überlappung und dadurch mehr getrennte Elemente. Kleine n bedeuten wiederum größere Überdeckungsintervalle und damit weniger 0-Simplizia im Output. Es ist sinnvoll, verschiedene n und p auszuprobieren, die Ergebnisse zu vergleichen und Auffälligkeiten zu interpretieren.

6 Anwendung

Bis hierhin lag der Schwerpunkt auf der methodischen Darstellung der beiden Kernverfahren der Topologischen Datenanalyse: Persistente Homologie und Mapper. In diesem Kapitel soll es darum gehen, die entwickelte Maschinerie anzuwenden und zu schauen wie sie zur Analyse von Daten gebraucht werden kann. Vor allem die Stärken von Mapper, d.h. die Visualisierung von mehrdimensionalen Datensätzen wird in den folgenden Abschnitten präsentiert.

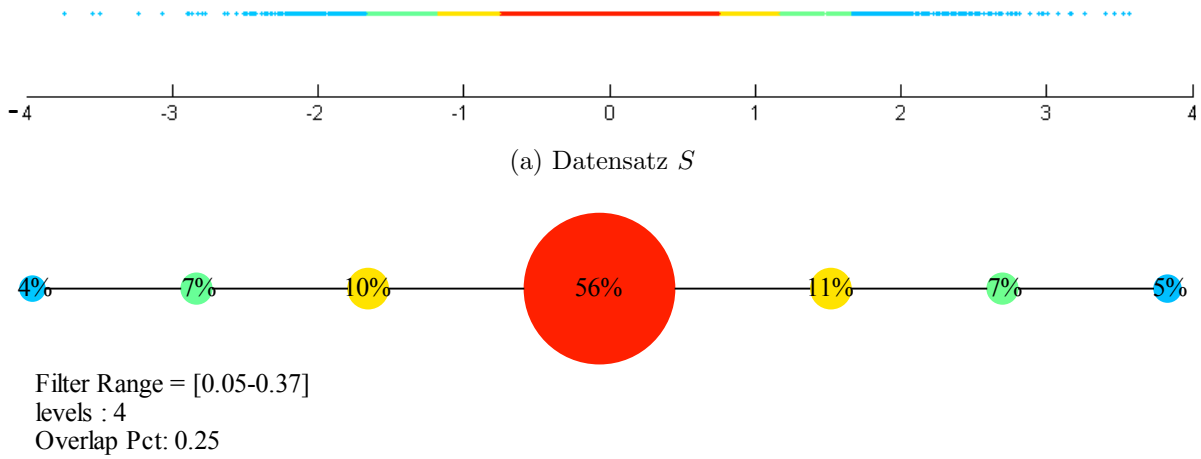
6.1 Software

Für Mapper existiert keine komplett freie vorgefertigte Software-Lösung. Auf Basis der Arbeit von [3] für sogenannte BioMapper wurden in R Modifizierungen, Fehlerbehebung und Anpassungen zur Berechnung von $Mapper(f^{-1}(\mathcal{V}), C)$ vorgenommen. Zur Visualisierung wird die freie Software Graphviz angewandt, die zur Darstellung von gerichteten und ungerichteten Graphen entwickelt wurde [9]. Die Mapper-Software wird getestet, indem die Abbildungen 16 und 17 reproduziert werden. Die Ergebnisse sind in der Abbildung 18 bzw. 19 zu sehen. Es wird nur der 1-Skeleton des Outputs ausgegeben. Wie im vorherigen Kapitel angeführt, sind im Mapper-Output nützliche Informationen codiert. Die Größe der Knoten entspricht der Anzahl der zugeordneten Punkte. Die Farbe spiegelt den Mittelwert des benutzen Filters wieder: rot für den höchsten Wert der Dichte und blau für den niedrigsten Wert. Außerdem können den Knoten auch Beschriftungen zugeordnet werden. In diesem Fall wird ein prozentualer Anteil der Punkte in Knoten ausgegeben. Je nach Analyse und Anwendung kann eine andere Beschriftung sinnvoller sein. Die angepasste Mapper-Software funktioniert und die ersten Ergebnisse sehen wie erhofft aus (siehe Abbildung 18 und 19).



(a) Datensatz S (b) Mapper-Output für S

Abbildung 18: Reproduktion der Abbildung 16 mit der Mapper-Software. Als Filter dient die Funktion $f(x, y) = y$. (a) zeigt den Datensatz S. Die Punkte wurden nach der Mapper-Zuordnung gefärbt. In (b) ist der Mapper-Output zu sehen. Auch der Filterbereich und die Informationen über die Überdeckung $\mathcal{V}(n, p)$ werden ausgegeben.



(a) Datensatz S

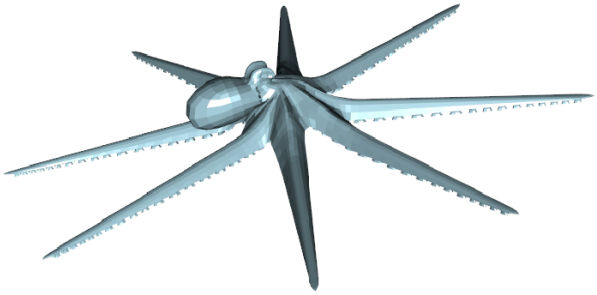
(b) Mapper-Output für S

Abbildung 19: Reproduktion der Abbildung 17 mit der Mapper-Software. Als Filter dient hier die Standardnormalverteilung $f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$. In (a) ist wieder der Datensatz S und in (b) der Mapper-Output zu sehen.

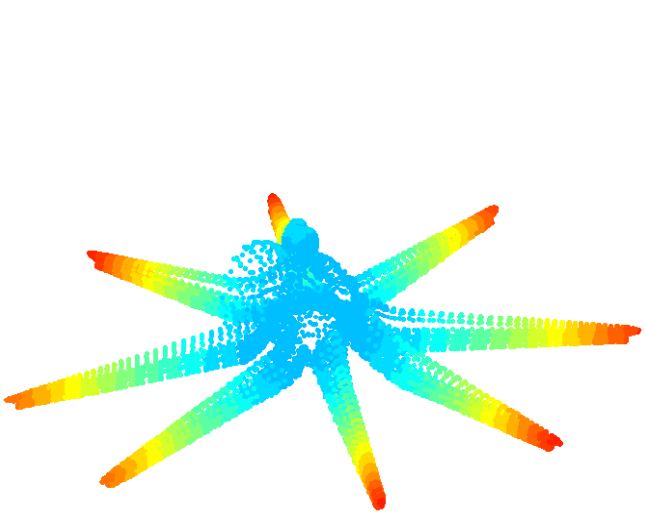
6.2 Visualisierung mit dem Mapper-Algorithmus anhand von 3D-Objekten

Bei der Arbeit mit Daten ist es oft hilfreich, diese erst zu visualisieren. Bei mehrdimensionalen Daten kann das recht problematisch sein. Mit dem Mapper ist es möglich einen Datensatz zu einem 1-Skeleton eines Simplizialkomplexes zu reduzieren. Der 1-Sekeleton stellt zwar nur eine Vereinfachung dar, beinhaltet aber immer noch große Menge an nützlichen Informationen, wie zum Beispiel die Anzahl der getrennten Komponenten.

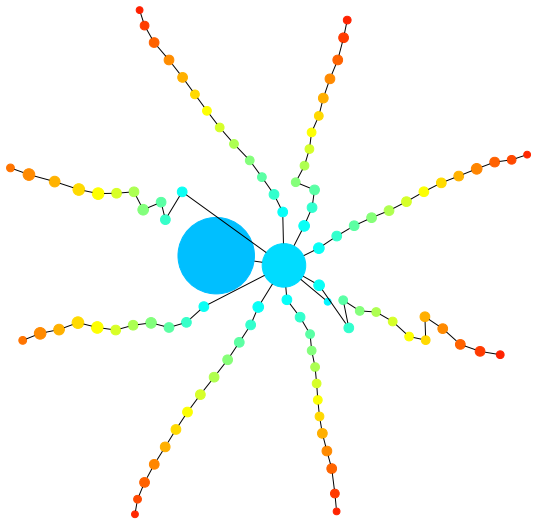
Die Visualisierung mit Mapper wird am Beispiel von dreidimensionalen Daten durchgeführt. Die 3D-Modelle, die hier benutzt wurden, können auch unter [17, 23] gefunden werden. Als Filter wurde die Exzentrizität $E_1(x)$ genommen (siehe Seite 27). Aus einem vorhandenem 3D-Modell werden nur die Knoten benutzt und mit der Mapper-Routine analysiert - mit der Hoffnung, dass die Software eine gute Visualisierung der Punktwolke liefert. Abbildung 20 zeigt den Verlauf des Prozesses.



(a) 3D-Modell Krake

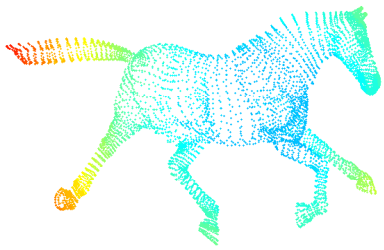


(b) Krake als Punktwolke

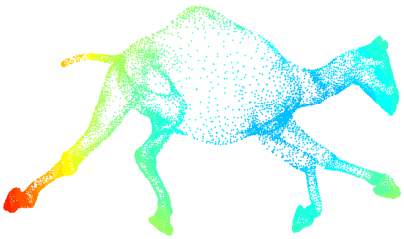


(c) Mapper-Output für Kraken-Modell

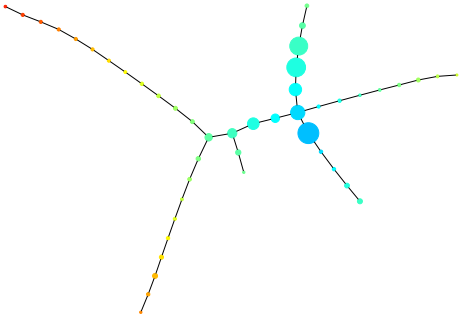
Abbildung 20: Visualisierung mit Mapper



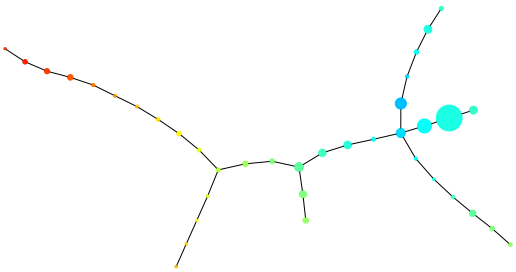
(a) Pferd in Position 1



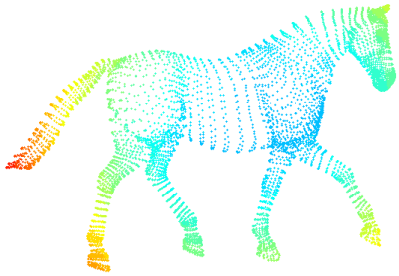
(b) Kamel in Position 1



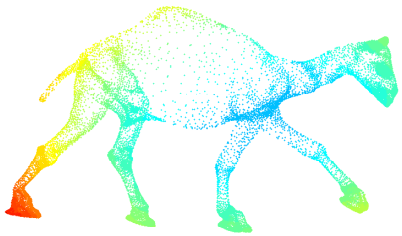
(c) Mapper für Pferd in Position 1



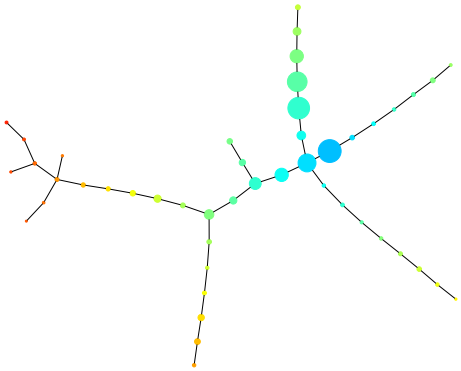
(d) Mapper für Kamel in Position 1



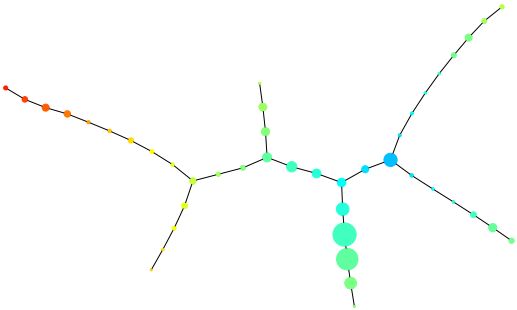
(e) Pferd in Position 2



(f) Kamel in Position 2



(g) Mapper für Pferd in Position 2



(h) Mapper für Kamel in Position 2

Abbildung 21: Visualisierungen von Tier-Modellen (siehe [6])

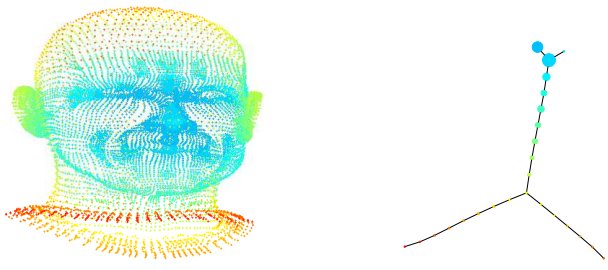
Die Betrachtung des Mapper-Outputs für die 3D-Modelle liefert insgesamt adäquate Resultate. Das Modell der Krake und das dazugehörige 1-Skeleton (siehe Abbildung 20(a) und (c)) korrespondieren sehr präzise. Auch in Abbildung 21 kann der Mapper die Tier-Modelle gut widerspiegeln. Alle sechs Äste des Modells werden immer gefunden. Darüber hinaus ist erkennbar, dass ähnliche Modelle (in dem Fall ähnliche Tiere: Pferd und Kamel) gleichartige Resultate liefern. Auch bei den verschiedenen Positionen des gleichen Tieres werden gute und fast gleiche Bilder erzeugt.

Faktoren, die Einfluss auf den Mapper nehmen, sind die Dichte und die Entfernung von einzelnen Punkten. Der Mapper-Output für das zweite Pferd-Modell (zu sehen in (g)) kann irreführend sein. Die Abzweigung, die der Schwanz des Pferdes repräsentiert, wird auf viele kleine Zweige gespalten. Die Kamel-Modelle werden dafür jedes mal gut visualisiert. Die Anzahl der Punkte in den Pferd- und Kamel-Modellen ist aber sehr unterschiedlich (Pferd-Modelle haben $n = 8431$ und Kamel-Modelle $n = 21884$) und damit liegen die Punkte in den Kamel-Modellen näher beieinander. Bei den Mensch-Modellen ist erneut zu erkennen, dass Mapper teilweise eine sehr gute (siehe Abbildung 22(a) und (b)), aber auch eine eher wenig anschaubare (siehe Abbildung 22(c) und (d)) Visualisierung eines Datensatzes geben kann.

Als Fazit bleibt festzuhalten, dass Mapper eine gute Hilfe bei der Visualisierung von mehrdimensionalen Daten sein kann. Weitere Anwendungen mit alternativen Modellen und verschiedenen Filtern sind notwendig, um eine stabile Aussage zur Eignung treffen zu können.

(a) Ganzkörper-Modell

(b) Mapper-Output Körper



(c) Kopf-Modell

(d) Mapper-Output Kopf

Abbildung 22: Visualisierung von Mensch-Modellen

7 Fazit

Bei der Analyse von höherdimensionalen Daten kann deren gegenseitige räumliche Anordnung im von den Variablen (Merkmalen) aufgespannten Raum wichtige Informationen über den Datensatz liefern. Mit Hilfe der Topologie ist es möglich, gleiche Strukturen in verschiedenen Räumen zu entdecken. Die Topologische Datenanalyse, d.h. ein Gebiet der Mathematik das topologische Methoden zur Untersuchung reeller Daten nutzt, hat an Bedeutung gewonnen und wird weiter entwickelt. Insbesondere die effiziente programmiertechnische Umsetzung der Verfahren und Algorithmen der TDA – auch für Open Source Software wie R – ist entwicklungsfähig. Dieser Beitrag konzentrierte sich darauf, die beiden wichtigsten Methoden der TDA (Persistente Homologie und der Mapper) formal und anhand von Beispieldaten grafisch vorzustellen. Hierfür wurde der Begriff Homologie definiert, die Euler-Charakteristik vorgestellt und der Mapper-Algorithmus beschrieben. Es wurde die Vorteile von Mapper bei der die Visualisierung von mehrdimensionalen Datensätzen demonstriert, woran klassische statistische Verfahren scheitern.

Die grafischen Anwendung des Mapper auf Daten war beschränkte, indem die Ergebnisse des Mappers nur als 1-Skeleton ausgegeben wurden. Der $Mapper(f^{-1}(\mathcal{V}), C)$ kann aber beliebige n-Simplizia beinhalten. Um dreidimensionale Simplizialkomplexe des Mapper-Algorithmus graphisch zu präsentieren, müsste eine Software für 3D-Graphik entwickelt werden. Dies ist mit hohem Aufwand verbunden. Aus diesem Grund hat sich dieser Beitrag auf 1-Skeletons konzentriert. Interpretation sowie Präsentation der von Mapper produzierten n-Skeletons könnten Gegenstand weiterer Forschung sein. Darüber hinaus ergibt sich Forschungsbedarf in der Anwendung der beschriebenen Methodik auf Wirtschaftsdaten. Inwieweit kann die Topologische Datenanalyse zur Analyse von komplexen Datensätzen praktikabel genutzt werden. Vor allem die Vorteile von Mapper, d.h. die Erkennung von Gruppen in Daten sollte untersucht werden. Dies muss einen Vergleich mit statistischen Verfahren der Clusteranalyse beinhalten.

Die Topologische Datenanalyse ist noch ein vergleichsweise junges Gebiet der Mathematik mit großem Entwicklungspotenzial für die Analyse von höherdimensionalen Daten. Weil aber die Methodik noch nicht so fortgeschritten ist, besteht noch weiterer Bedarf an theoretischer Arbeit. Dies sollte von einen effizienten Umsetzung in Analysesoftware begleitet werden, um die Verbreitung dieses neuen Ansatzes der Datenanalyse zu erleichtern.

Literatur und Quellen

[1] H. Adams, A. Tausz, M. Vejdemo-Johansson, *JavaPlex: A research software package for persistent (co)homology.*
`http://appliedtopology.github.io/javaplex/`

[2] J. Bacher, *Clusteranalyse. Anwendungsorientierte Einführung in Klassifikationsverfahren.* 2. Auflage 2, Oldenbourg Wissenschaftsverlag, München-Wien-Oldenbourg, 1996.

[3] G. R. Bowman, G. Carlsson, L. J. Guibas, X. Huang, M. Lesnick, V. S. Pande, G. Singh, J. Sun und Y. Yao, *Topological methods for exploring low-density states in biomolecular folding pathways.* The Journal of Chemical Physics 130, 2009.

[4] P. Y. Lum, G. Singh, A. Lehman, T. Ishkanov, M. Vejdemo-Johansson, M. Alagappan, J. Carlsson und G. Carlsson, *Extracting insights from the shape of complex data using topology.* Scientific Reports 3, no. 1236, 2013.

[5] G. Carlsson, *Topology and data.* Bulletin of the American Mathematical Society (New Series) vol. 46, no. 2, s. 255–308, 2009.
`http://www.ams.org/journals/bull/2009-46-02/S0273-0979-09-01249-X/S0273-0979-09-01249-X.pdf`

[6] G. Carlsson, F. Mémoli, G. Singh, *Topological Methods for the Analysis of High Dimensional Data Sets and 3D Object Recognition.* In Eurographics Symposium on Point-Based Graphics (2007), M. Botsch, R. Pajarola (Editors)

[7] F. Cazals, J. Giesen, M. Pauly und A. Zomorodian, *The conformal alpha shape filtration.* The Visual Computer vol. 22, 2006.
`http://lgg.epfl.ch/publications/2006/cazals_2006_CAS.pdf`

[8] H. Edelsbrunner, J.L. Harer, *Computational Topology: An Introduction.* American Mathematical Society, 2010.

[9] E. R. Gansner and S.C. North, *GraphViz: An open graph visualization system and its applications to software engineering.*
`http://www.graphviz.org/`

[10] A. Hatcher, *Algebraic Topology.* Cambridge University Press, New York, NY, 2002.
`http://www.math.cornell.edu/~hatcher/AT/AT.pdf`

[11] T. W. Hungerford, *Algebra.* Springer-Verlag, New York, 1974.

[12] K. Jänich, *Topologie.* 8. Auflage, Springer-Verlag, Berlin-Heidelberg, 2005.

[13] W. Lück, *Algebraische Topologie: Homologie und Mannigfaltigkeiten.* 1. Auflage, Vieweg+Teubner Verlag, 2005.

[14] A. A. Markov, *Insolubility of the Problem of Homeomorphy.* Doklady Akademii nauk SSSR, 01/1958.

[15] W. S. Massey, *Algebraic Topology: An Introduction*. Springer-Verlag, New York, 1977.

[16] B. Poonen, *Undecidable problems: a sampler*. arXiv: 1204.0299, 2012. `http://arxiv.org/abs/1204.0299`

[17] J. Popovic, R. R. Summer, *Mesh Data from Deformation Transfer for Triangle Meshes*. `http://people.csail.mit.edu/sumner/research/deftransfer/data.html`

[18] T. Rado, *Über den Begriff der Riemannschen Fläche*. Acta Szeged 2: 101–121, 1925.

[19] H. Seifert, W. Threlfall, *Lehrbuch der Topologie*. American Mathematical Society (Reprinted 2003), 1934.

[20] G. Singh, Y. Yao, *Bio Mapper v1*. `https://simtk.org/project/xml/downloads.xml?group_id=362`

[21] R. B. Stovner, *On the Mapper Algorithm: A study of a new topological method for data analysis*. NTNU - Institutt for matematiske fag, Norway, 2012.

[22] M. Steinbach, P.-N. Tan, V. Kumar, *Introduction to Data Mining*. Pearson Higher Ed USA, 2005.

[23] *TurboSquid, the world's source for professional 3D models*. `http://www.turbosquid.com/`

[24] A. Zomorodian, *Fast construction of the Vietoris-Rips complex*. Computers & Graphics vol. 34, Issue 3, 2010.

[25] A. Zomorodian, *Topological Data Analysis*. Advances in Applied and Computational Topology, Proc. Symp. Applied Math vol. 70, 2012.

[26] A. Zomorodian, *Topology for Computing*. Cambridge Monographs on Applied and Computational Mathematics (No. 16), 2005.

UNIVERSITÄT POTSDAM

STATISTISCHE DISKUSSIONSBEITRÄGE

Herausgeber: Hans Gerhard Strohe

Nr. 1	1995	Strohe, Hans Gerhard: Dynamic Latent Variables Path Models - An Alternative PLS Estimation -
Nr. 2	1996	Kempe, Wolfram. Das Arbeitsangebot verheirateter Frauen in den neuen und alten Bundesländern - Eine semiparametrische Regressionsanalyse
Nr. 3	1996	Strohe, Hans Gerhard: Statistik im DDR-Wirtschaftsstudium zwischen Ideologie und Wissenschaft
Nr. 4	1996	Berger, Ursula: Die Landwirtschaft in den drei neuen EU-Mitgliedsstaaten Finnland, Schweden und Österreich - Ein statistischer Überblick
Nr. 5	1996	Betzin, Jörg: Ein korrespondenzanalytischer Ansatz für Pfadmodelle mit kategorialen Daten
Nr. 6	1996	Berger, Ursula: Die Methoden der EU zur Messung der Einkommenssituation in der Landwirtschaft - Am Beispiel der Bundesrepublik Deutschland
Nr. 7	1997	Strohe, Hans Gerhard / Geppert, Frank: Algorithmus und Computerprogramm für dynamische Partial Least Squares Modelle
Nr. 8	1997	Rambert, Laurence / Strohe, Hans Gerhard: Statistische Darstellung transformationsbedingter Veränderungen der Wirtschafts- und Beschäftigungsstruktur in Ostdeutschland
Nr. 9	1997	Faber, Cathleen: Die Statistik der Verbraucherpreise in Rußland - Am Beispiel der Erhebung für die Stadt St. Petersburg
Nr. 10	1998	Nosova, Olga: The Atractiveness of Foreign Direct Investment in Russia and Ukraine - A Statistical Analysis
Nr. 11	1999	Gelaschwili, Simon: Anwendung der Spieltheorie bei der Prognose von Marktprozessen
Nr. 12	1999	Strohe, Hans Gerhard / Faber, Cathleen: Statistik der Transformation - Transformation der Statistik. Preisstatistik in Ostdeutschland und Rußland
Nr. 13	1999	Müller, Claus: Kleine und mittelgroße Unternehmen in einer hoch konzentrierten Branche am Beispiel der Elektrotechnik. Eine statistische Langzeitanalyse der Gewerbezählungen seit 1882
Nr. 14	1999	Faber, Cathleen: The Measurement and Development of Georgian Consumer Prices
Nr. 15	1999	Geppert, Frank / Hübner, Roland: Korrelation oder Kointegration – Eignung für Portfoliostrategien am Beispiel verbriefter Immobilenanlagen
Nr. 16	2000	Achsani, Noer Azam / Strohe, Hans Gerhard: Statistischer Überblick über die indonesische Wirtschaft
Nr. 17	2000	Bartels, Knut: Testen der Spezifikation von multinominalen Logit-Modellen
Nr. 18	2002	Achsani, Noer Azam / Strohe, Hans Gerhard: Dynamische Zusammenhänge zwischen den Kapitalmärkten der Region Pazifisches Becken vor und nach der Asiatischen Krise 1997
Nr. 19	2002	Nosova, Olga: Modellierung der ausländischen Investitionstätigkeit in der Ukraine
Nr. 20	2003	Gelaschwili, Simon / Kurtanidse, Zurab: Statistische Analyse des Handels zwischen Georgien und Deutschland
Nr. 21	2004	Nastansky, Andreas: Kurz- und langfristiger statistischer Zusammenhang zwischen Geldmengen- und Preisentwicklung: Analyse einer kointegrierenden Beziehung
Nr. 22	2006	Kauffmann, Albrecht / Nastansky, Andreas: Ein kubischer Spline zur temporalen Disaggregation von Stromgrößen und seine Anwendbarkeit auf Immobilienindizes
Nr. 23	2006	Mangelsdorf, Stefan: Empirische Analyse der Investitions- und Exportentwicklung des Verarbeitenden Gewerbes in Berlin und Brandenburg
Nr. 24	2006	Reilich, Julia: Return to Schooling in Germany
Nr. 25	2006	Nosova, Olga / Bartels, Knut: Statistical Analysis of the Corporate Governance System in the Ukraine: Problems and Development Perspectives
Nr. 26	2007	Gelaschwili, Simon: Einführung in die statistische Modellierung und Prognose
Nr. 27	2007	Nastansky, Andreas: Modellierung und Schätzung von Vermögenseffekten im Konsum
Nr. 28	2008	Nastansky, Andreas: Schätzung vermögenspreisinduzierter Investitionseffekte in Deutschland

UNIVERSITÄT POTSDAM

STATISTISCHE DISKUSSIONSBEITRÄGE

Herausgeber: Hans Gerhard Strohe

Nr. 29	2008	Ruge, Marcus / Strohe, Hans Gerhard: Analyse von Erwartungen in der Volkswirtschaft mit Partial-Least-Squares-Modellen
Nr. 30	2009	Newiak, Monique: Prüfungsurteile mit Dollar Unit Sampling – Ein Vergleich von Fehlerschätzmethoden für Zwecke der Wirtschaftsprüfung: Praxis, Theorie, Simulation –
Nr. 31	2009	Ruge, Marcus: Modellierung von Stimmungen und Erwartungen in der deutschen Wirtschaft
Nr. 32	2009	Nosova, Olga: Statistical Analysis of Regional Integration Effects
Nr. 33	2009	Mangelsdorf, Stefan: Persistenz im Exportverhalten – Kann punktuelle Exportförderung langfristige Auswirkungen haben? -
Nr. 34	2009	Kbiladze, David: Einige historische und gesetzgeberische Faktoren der Reformierung der georgischen Statistik
Nr. 35	2009	Nastansky, Andreas / Strohe, Hans Gerhard: Die Ursachen der Finanz- und Bankenkrise im Lichte der Statistik
Nr. 36	2009	Gelaschwili, Simon / Nastansky, Andreas: Development of the Banking Sector in Georgia
Nr. 37	2010	Kunze, Karl-Kuno / Strohe, Hans Gerhard: Time Varying Persistence in the German Stock Market
Nr. 38	2010	Nastansky, Andreas / Strohe, Hans Gerhard: The Impact of Changes in Asset Prices on Real Economic Activity: A Cointegration Analysis for Germany
Nr. 39	2010	Kunze, Karl-Kuno / Strohe, Hans Gerhard: Antipersistence in German Stock Returns
Nr. 40	2010	Dietrich, Irina / Strohe, Hans Gerhard: Die Vielfalt öffentlicher Unternehmen aus der Sicht der Statistik - Ein Versuch, das Unstrukturierte zu strukturieren
Nr. 41	2010	Nastansky, Andreas / Lanz, Ramona: Bonuszahlungen in der Kreditwirtschaft: Analyse, Regulierung und Entwicklungstendenzen
Nr. 42	2010	Dietrich, Irina / Strohe, Hans Gerhard: Die Vermögenslage öffentlicher Unternehmen in Deutschland - Statistische Analyse anhand von amtlichen Mikrodaten der Jahresabschlüsse.
Nr. 43	2010	Ulbrich, Hannes-Friedrich: Höherdimensionale Kompositionsdaten – Gedanken zur grafischen Darstellung und Analyse -
Nr. 44	2011	Dietrich, Irina / Strohe, Hans Gerhard: Statistik der öffentlichen Unternehmen in Deutschland – Die Datenbasis
Nr. 45	2011	Nastansky, Andreas: Orthogonale und verallgemeinerte Impuls-Antwort-Funktionen in Vektor-Fehlerkorrekturmodellen
Nr. 46	2011	Dietrich, Irina / Strohe, Hans Gerhard: Die Finanzlage öffentlicher Unternehmen in Deutschland - Statistische Analyse amtlicher Mikrodaten der Jahresabschlüsse -
Nr. 47	2011	Teitge, Jonas / Nastansky, Andreas: Interdependenzen in den Renditen DAX-notierter Unternehmen nach Branchen
Nr. 48	2011	Dietrich, Irina: Die Ertragslage öffentlicher Unternehmen in Deutschland - Statistische Analyse amtlicher Mikrodaten der Jahresabschlüsse -
Nr. 49	2011	Kauper, Benjamin / Kunze, Karl-Kuno: Modellierung von Aktienkursen im Lichte der Komplexitätsforschung
Nr. 50	2011	Nastansky, Andreas / Strohe, Hans Gerhard: Konsumausgaben und Aktienmarktentwicklung in Deutschland: Ein kointegriertes vektorautoregressives Modell
Nr. 51	2014	Nastansky, Andreas / Mehnert, Alexander / Strohe, Hans Gerhard: A Vector Error Correction Model for the Relationship between Public Debt and Inflation in Germany
Nr. 52	2019	Kauffmann, Albrecht / Nastansky, Andreas: Explorative Analyse der Preise von Einfamilienhäusern und Eigentumswohnungen in Deutschland
Nr. 53	2019	Nastansky, Andreas: Topologische Datenanalyse: Eine Einführung in die Persistente Homologie und Mapper