Mapper

김지수 (Jisu KIM)

통계이론세미나 - 위상구조의 통계적 추정, 2023 가을학기

The lecture note is largely based on [5] and [1].

There are two directions for building covers and using their nerves to exhibit the topological structure of data. First is to cover data by balls, and then use distance function frameworks. This leads to geometric inference and providing a framework to establish various theoretical results in Topological Data Analysis. Second is to use a function defined on the data and use Mapper algorithm. This leads to exploratory data analysis and visualization. See Figure 1.

We first recall the cover and the Nerve Theorem.

Definition ([10, Section 26])**.** A collection $\mathcal{A}$ of subsets of a space X is said to cover X, or to be a covering of X, if the union of the elements of $\mathcal{A}$ is equal to X. It is called an open cover of X if its elements are open subsets of X.

We let $\mathcal{U} = \{U_i\}_{i\in I}$ be a cover of $\mathbb{X}$.

Definition. The nerve $Nrv_{\mathcal{U}}$ of $\mathcal{U}$ is the simplicial complex whose vertices are U_i's and

$$Nrv_{\mathcal{U}} := \left\{ \{U_0, \dots, U_k\} \in \mathcal{U} : \bigcap_{i=0}^{k} U_i \neq \emptyset \right\}. \tag{1}$$

Given a cover of a data set, where each set of the cover can be, for example, a local cluster or a grouping of data points sharing some common properties, its nerve provides a compact and global combinatorial description of the relationship between these sets through their intersection patterns. See Figure 2.

The topology of the nerve is linked to underlying continuous spaces via Nerve Theorem. Under some assumptions, the nerve of a cover is homotopic equivalent to the topology of the union of sets of the cover by the following Nerve Theorem.

Theorem (Nerve Theorem [9, Corollary 4G.3][8, Section III.2])**.** *Let $\mathcal{U} = \{U_i\}_{i\in I}$ be an open cover of a space $\mathbb{X}$ such that for any finite subset $\{U_0, \dots, U_k\} \subset \mathcal{U}$, the intersection $\bigcap_{i=0}^{k} U_i$ is either empty or contractible. Then, the nerve $Nrv_{\mathcal{U}}$ is homotopic equivalent to $\mathbb{X}$.*

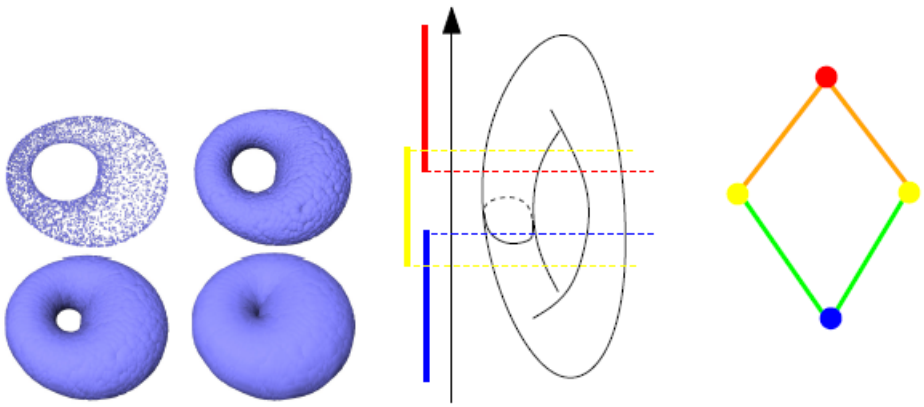

Figure 1: [1] Covering data by balls, and then use distance function frameworks (left), Using a function defined on the data and using Mapper algorithm (right).

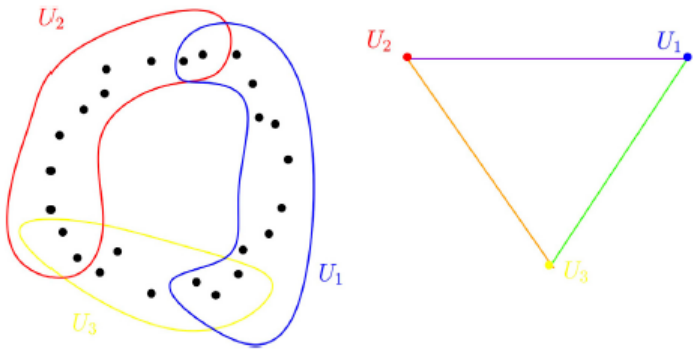

Figure 2: [5, Figure 3] Point cloud and an open cover (left), and the nerve of this cover (right).

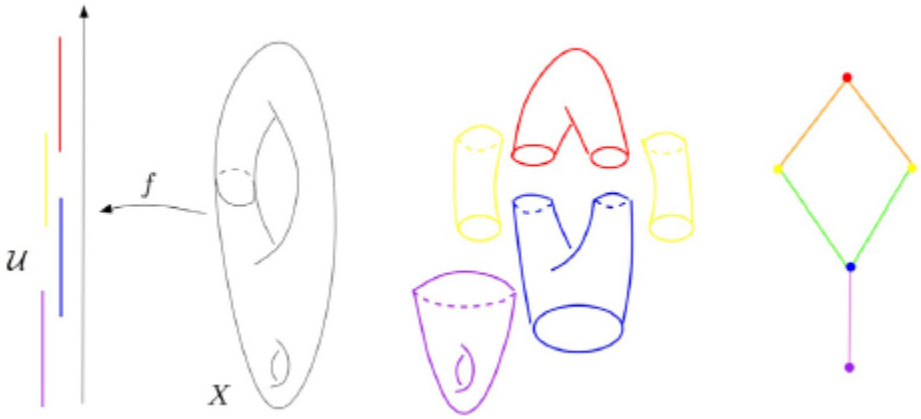

Figure 3: [5, Figure 4A] The pull-back cover of the height function f on a surface (left), The refined pull-back cover (middle), and its nerve (right).

Using Covers and Nerves for Exploratory Data Analysis and Visualization: the Mapper algorithm

Using the nerve of covers as a way to summarize, visualize, and explore data is a natural idea that was first proposed for TDA in the study by [13], giving rise to the so-called Mapper algorithm.

Definition. Let $f : X \to \mathbb{R}^d$, $d \geq 1$, be a continuous real valued function and let $\mathcal{U} = \{U_i\}_{i \in I}$ be a cover of $\mathbb{R}^d$. The pull-back cover of X induced by (f, U) is the collection of open sets $\{f^{-1}(U_i)\}_{i \in I}$. The refined pull-back cover is the collection of connected components of the open sets $f^{-1}(U_i)$, $i \in I$.

The idea of the Mapper algorithm is, given a data set $\mathcal{X}$ and a well-chosen real-valued function $f : \mathcal{X} \to \mathbb{R}^d$, to summarize $\mathcal{X}$ through the nerve of the refined pull-back of a cover $\mathcal{U}$ of $f(\mathcal{X})$. See Figure 3. For well-chosen covers $\mathcal{U}$, this nerve is a graph providing an easy and convenient way to visualize the summary of the data.

When $\mathcal{X}$ is indeed a finite set of points, each open set $f^{-1}(U_i)$ is also a finite set of points, so its connected components are one-point sets, and not quite useful. Hence, we use clusters instead of connected components. This algorithm is called Mapper, described in Algorithm 1 and illustrated in Figure 4.

The Mapper algorithm contains various choices that are left to the user.

The choice of lens/filter function

The choice of the lens or filter function f strongly depends on what to be highlighted. Some classical choices are:

- Density estimates: Mapper reveals the structure and connectivity of high-density areas (clusters).
- Principal Component Analysis (PCA) coordinates, Non-Linear Dimensionality Reduction (NLDR) coordinates, eigenfunctions of graph laplacians: Mapper revealssome ambiguity in the use of nonlinear dimensionality

Algorithm 1 The Mapper Algorithm.

Input: a data set $\mathcal{X}$ with a metric or a dissimilarity measure between data points, a function $f : \mathcal{X} \to \mathbb{R}^d$ (called lens or filter), and a cover $\mathcal{U}$ of $f(\mathcal{X})$.

1. for each $U \in \mathcal{U}$, decompose $f^{-1}(U)$ into clusters $C_{U,1}, \ldots, C_{U,k_U}$.
2. Compute the nerve of the cover of $\mathcal{X}$ defined by the $C_{U,1}, \ldots, C_{U,k_U}$, $U \in \mathcal{U}$.

Output: the nerve (often a graph for well-chosen covers): a vertex $v_{U,i}$ for each cluster $C_{U,i}$, and an edge between $v_{U,i}$ and $v_{U',j}$ if $C_{U,i} \cap C_{U',j} \neq \emptyset$.

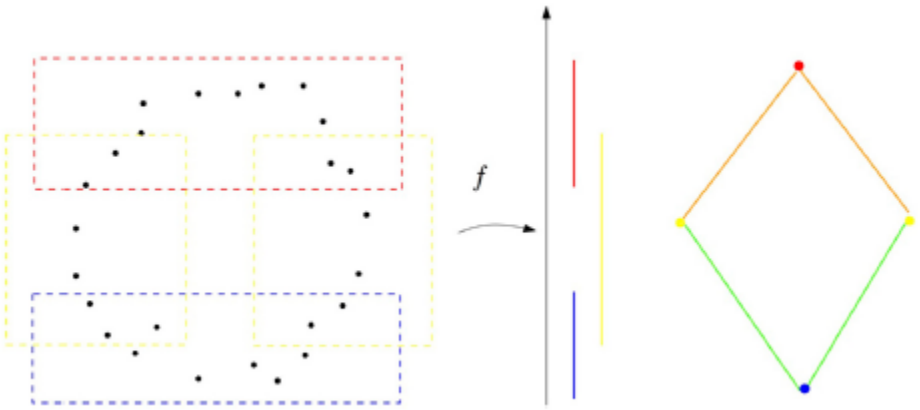

Figure 4: [5, Figure 4B] Mapper algorithm on a point cloud sampled around a circle and the height function. First, the pull-back cover of the height function n the point cloud is computed and refined via clustering (left). Second, the nerve of the refined pull-back cover is computed (right).

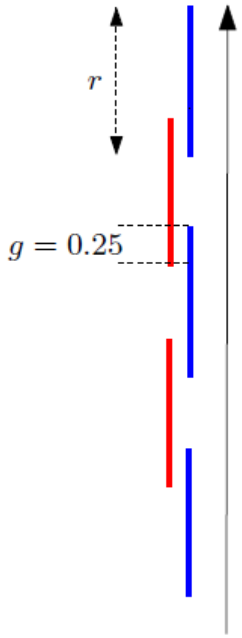

Figure 5: [1] The resolution r and the gain g.

reductions.

- The centrality function $f(x) = \sum_{y \in \mathcal{X}} d(x, y)$ and the eccentricity function $f(x) = \max_{y \in \mathcal{X}} d(x, y)$ do not require prior knowledge about the data.
- Distance to a root point for data that are sampled around one-dimensional filamentary structures: Mapper recovers the underlying topology of the filamentary structures [4].

The choice of the cover

Consider when f is a real-valued function $f : \mathcal{X} \to \mathbb{R}$, then $\mathcal{U}$ is usually a set of intervals. The resolution r is the maximum diameter of an interval in $\mathcal{U}$. The resolution may also be replaced by a number N of intervals in the cover. The gain g is the percentage of overlap between intervals, when they overlap. See Figure 5.

- small r (large N) implies finer resolution and more nodes.
- large r (small N) implies rougher resolution and less nodes.
- small g implies less connectivity. Note that if g is below 50%, then every point of the real line is covered by at most 2 open sets of $\mathcal{U}$, and the output nerve is a graph.
- large g implies more connectivity and the dimensionality of the nerve increases.

The output of Mapper is very sensitive to the choice of $\mathcal{U}$, and small changes in the resolution and gain parameters may result in very large changes in the output, making the method very unstable. A classical strategy explores some range of parameters and selects the ones that provide the most informative output from the user perspective.

The choice of the clusters

There are two strategies to compute the clusters of the preimage of the open sets $U \in \mathcal{U}$.

1. (local) Apply for each $U \in \mathcal{U}$, a cluster algorithm chosen by the user, to the preimage $f^{-1}(U)$. See Figure 6.
2. (global) Build a neighboring graph on top of the data set $\mathcal{X}$, for example, a k-NN graph or a Vietoris-Rips graph, and for each $U \in \mathcal{U}$, take the connected components of the subgraph with the vertex set $f^{-1}(U)$. See Figure 7.

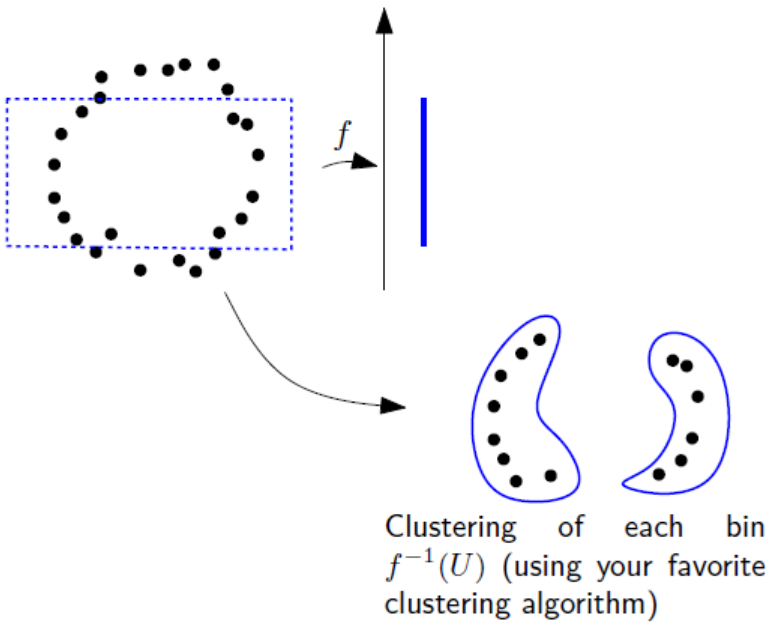

Figure 6: [1] Local approach to compute the clusters of the preimage of the open sets $U \in \mathcal{U}$.

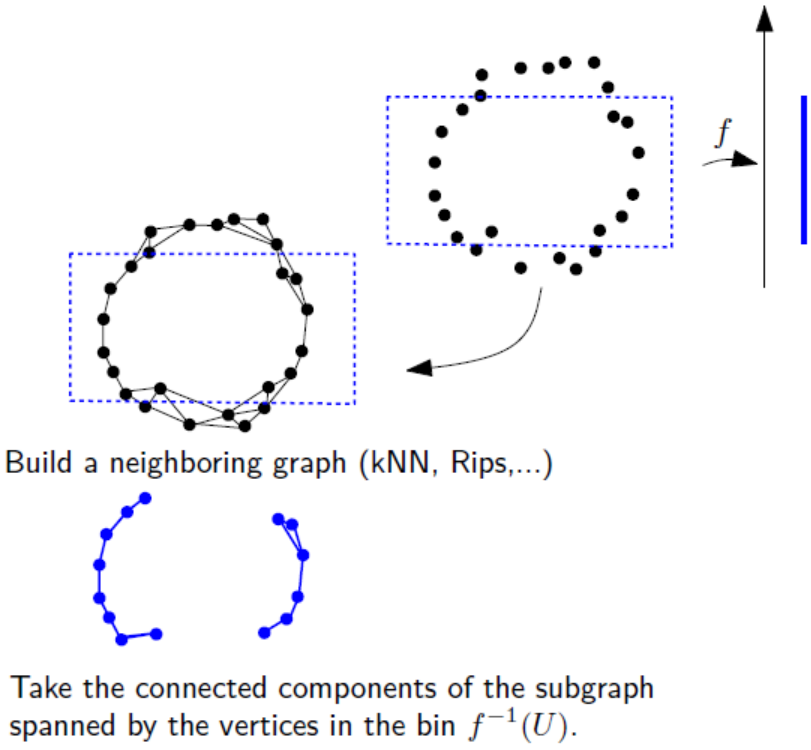

Figure 7: [1] Global approach to compute the clusters of the preimage of the open sets $U \in \mathcal{U}$.

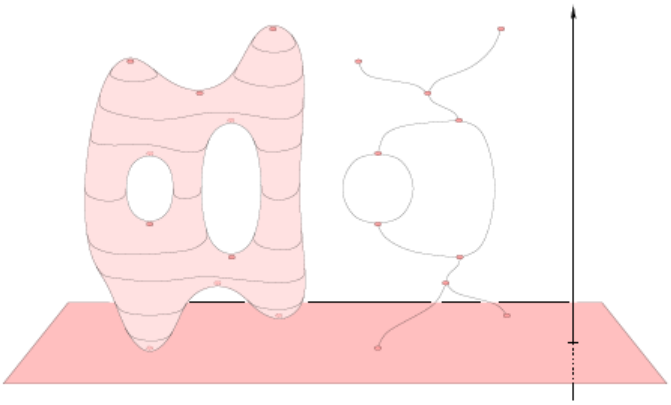

Figure 8: [8, Figure VI.13] For a Morse function, its Reeb graph is indeed a finite graph where critical points of a Morse function correspond to nodes of a finite graph.

Theoretical and Statistical Aspects of Mapper

The output of the Mapper algorithm can be seen as a discretized version of the Reeb graph [11], or Reeb space. Given a topological space X and a continuous function $f : X \to Z$, where (Z, d_Z) is a metric space, the Reeb space of X is an approximation of X that preserves its connectivity structures. When $f : X \to \mathbb{R}$ is scalar-valued, it is usually called the Reeb graph.

Definition. [2, Definition 2.1] Let X be a topological space and $f : X \to Z$ be a continuous function defined on it. The Reeb space of X is the quotient space endowed with a quotient topology,

$$R_f(X) \coloneqq X/_{\sim_f},$$

where, for all $x, x' \in X$, one has $x \sim_f x'$ if and only if $f(x) = f(x_0)$ and x, x' belong to the same connected component of $f^{-1}(f(x)) = f^{-1}(f(x'))$.

When X is a manifold (with boundary) and $f : X \to \mathbb{R}$ is a Morse function, Reeb graph is indeed a finite graph (finite simplicial complex with dimension ≤ 1): see [12, Lemma 2.1]. The nodes of Reeb graph R_f correspond to the critical points of X, and the rest of the Reeb graph is partitioned into arcs connecting the nodes. See Figure 8.

There is no standard connection between the Mapper and the Reeb graph, relatively compared to, for e.g., the stability theorem of Persistent Homology. Stability, consistency, and confidence set between the Mapper and the Reeb graph has been studied in [3]:

Definition. The (exact) modulus of continuity of f is defined as

$$\omega_f(\delta) = \sup_{\|x - x'\| \leq \delta} |f(x) - f(x')| .$$

Theorem. *[3, Theorem 7] Suppose $\mathbb{X}$ is a smooth and compact Euclidean submanifold with reach $\tau > 0$ and convexity radius $\rho > 0$. Let $\mathcal{X} \subset \mathbb{X}$ be a finite set of points. Assume that the filter function $f : \mathbb{X} \to \mathbb{R}$ is Morse. Let ω_f be the modulus of continuity for f. Let r, g be the resolution parameter and the gain parameter for Mapper, and suppose we use Vietoris-Rips graph $\mathrm{Rips}(\mathcal{X}, \delta)$ for clusters for Mapper. Then, if δ and $d_H(\mathcal{X}, \mathbb{X})$ is small enough compared to τ, ρ, g, r, then the bottleneck distance $d_\triangle$ between extended persistence diagram of the Reeb graph $R_f(\mathbb{X})$ and the mapper $M(\mathcal{X})$ is bounded as*

$$d_\triangle(R_f(\mathbb{X}), M(\mathcal{X})) \leq r + 2\omega_f(\delta).$$

Stability and consistency is extended to between the Mapper and the Reeb space in [2]. Other approaches have been focused on how the perturbation of Mapper algorithm affects in [6, 7].

Data Analysis with Mapper

As an exploratory data analysis tool, Mapper has been used for visualizing the topological shape of data, detecting clusters, and feature selection. After Mapper graph is computed from data, we find interesting topological structures

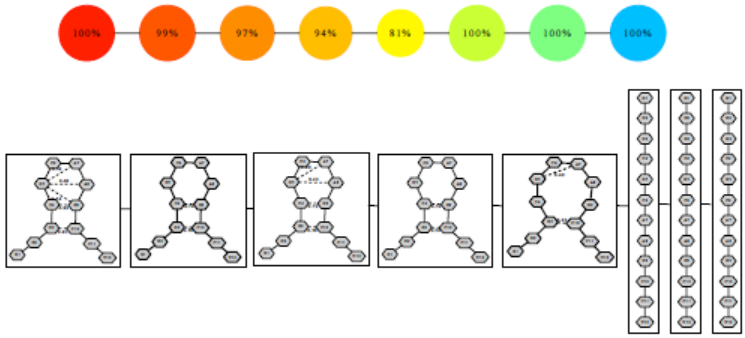

Figure 9: [14, Figure 3(a)] Unfolding pathway.

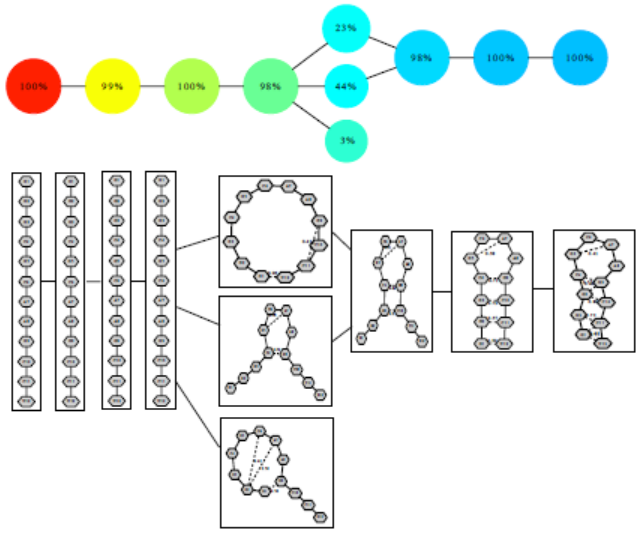

Figure 10: [14, Figure 3(b)] Refolding pathway.

(loops, clusters). For visualization, these topological structures directly represents the topological shape of data. For clustering, these topological structres are used to exhibit interesting clusters. For feature selection, we select the features/variables that best discriminate the data in these topological structures.

Example ([14])**.** Mapper is used for detecting multiple intermediate states on folding pathways. Data are conformations of molecules, 760 from unfolding events and 550 from folding events. The density is used as a filter function, and for unfolding events, the filter function reflects distance to folded states, and for folding events, the filter function reflects distance to extended states. Figure 9 and 10 show Mapper results for unfolding events events and folding events, respectively. Mapper of unfolding events has one unfolding pathway, but mapper of folding events has two refolding pathways. So this indicates that Refolding pathway has two different pahways to follow.

Software

There are many software options for computing mapper: Ayasdi, giotto-tda, Mapper Interactive, Scikit-TDA: Keppler Mapper, TDA Mapper, Python Mapper. For example, you can find a tutorial for using TDA Mapper at http://bertrand.michel.perso.math.cnrs.fr/Enseignements/TDA/Mapper.html.

References

[1] Jean-Daniel Boissonnat, Frédéric Cazals, Frédéric Chazal, and Julien Tierny. https://geometrica.saclay.inria.fr/team/fred.chazal/sophia2017/tdasophia2017.html, 2017.

[2] Mathieu Carrière and Bertrand Michel. Statistical analysis of Mapper for stochastic and multivariate filters. *J. Appl. Comput. Topol.*, 6(3):331–369, 2022.

[3] Mathieu Carrière, Bertrand Michel, and Steve Oudot. Statistical analysis and parameter selection for Mapper. *J. Mach. Learn. Res.*, 19:Paper No. 12, 39, 2018.

[4] Frédéric Chazal, Ruqi Huang, and Jian Sun. Gromov-hausdorff approximation of filamentary structures using reeb-type graphs. *Discret. Comput. Geom.*, 53(3):621–649, 2015.

[5] Frédéric Chazal and Bertrand Michel. An introduction to topological data analysis: Fundamental and practical aspects for data scientists. *Frontiers Artif. Intell.*, 4:667963, 2021.

[6] Tamal K. Dey, Facundo Mémoli, and Yusu Wang. Multiscale mapper: Topological summarization via codomain covers. In Robert Krauthgamer, editor, *Proceedings of the Twenty-Seventh Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2016, Arlington, VA, USA, January 10-12, 2016*, pages 997–1013. SIAM, 2016.

[7] Tamal K. Dey, Facundo Mémoli, and Yusu Wang. Topological analysis of nerves, reeb spaces, mappers, and multiscale mappers. *CoRR*, abs/1703.07387, 2017.

[8] Herbert Edelsbrunner and John L. Harer. *Computational topology.* American Mathematical Society, Providence, RI, 2010. An introduction.

[9] Allen Hatcher. *Algebraic topology.* Cambridge University Press, Cambridge, 2002.

[10] James R. Munkres. *Topology.* Prentice Hall, Inc., Upper Saddle River, NJ, 2000. Second edition of [MR0464128].

[11] Georges Reeb. Sur les points singuliers d'une forme de Pfaff complètement intégrable ou d'une fonction numérique. *C. R. Acad. Sci. Paris*, 222:847–849, 1946.

[12] V. V. Sharko. About Kronrod-Reeb graph of a function on a manifold. *Methods Funct. Anal. Topology*, 12(4):389–396, 2006.

[13] Gurjeet Singh, Facundo Mémoli, and Gunnar E. Carlsson. Topological methods for the analysis of high dimensional data sets and 3d object recognition. In Mario Botsch, Renato Pajarola, Baoquan Chen, and Matthias Zwicker, editors, *4th Symposium on Point Based Graphics, PBG@Eurographics 2007, Prague, Czech Republic, September 2-3, 2007*, pages 91–100. Eurographics Association, 2007.

[14] Yuan Yao, Jian Sun, Xuhui Huang, Gregory R. Bowman, Gurjeet Singh, Michael Lesnick, Leonidas J. Guibas, Vijay S. Pande, and Gunnar Carlsson. Topological methods for exploring low-density states in biomolecular folding pathways. *The Journal of Chemical Physics*, 130(14):144115, 04 2009.