Journal of Machine Learning Research 21 (2020) 1-45 Submitted 6/19; Revised 7/20; Published 8/20

A Numerical Measure of the Instability of Mapper-Type Algorithms

Francisco Belchí FRBEGU@GMAIL.COM
Jacek Brodzki J.BRODZKI@SOTON.AC.UK
Matthew Burfitt M.I.BURFITT@SOTON.AC.UK
School of Mathematical Sciences
University of Southampton
Highfield, Southampton, SO17 1BJ, UK

Mahesan Niranjan MN@ECS.SOTON.AC.UK
Department of Electronics and Computer science
University of Southampton
Highfield, Southampton, SO17 1BJ, UK

Editor: Ohad Shamir

Abstract

Mapper is an unsupervised machine learning algorithm generalising the notion of clustering to obtain a geometric description of a dataset. The procedure splits the data into possibly overlapping bins which are then clustered. The output of the algorithm is a graph where nodes represent clusters and edges represent the sharing of data points between two clusters. However, several parameters must be selected before applying Mapper and the resulting graph may vary dramatically with the choice of parameters.

We define an intrinsic notion of Mapper instability that measures the variability of the output as a function of the choice of parameters required to construct a Mapper output. Our results and discussion are general and apply to all Mapper-type algorithms. We derive theoretical results that provide estimates for the instability and suggest practical ways to control it. We provide also experiments to illustrate our results and in particular we demonstrate that a reliable candidate Mapper output can be identified as a local minimum of instability regarded as a function of Mapper input parameters.

Keywords: topological data analysis, Mapper, clustering stability, parameter selection, sub-sampling

1. Introduction

The success of topological data analysis rests on the discovery, demonstrated in many groundbreaking results, that methods from algebraic topology can provide insight into the structure and meaning of complex, multidimensional data (Carlsson, 2009). Mapper is a very important tool in any practical implementation of the central philosophy of topological data analysis and has been used with great success in many contexts. The list is very long and diverse, and includes breakthrough results in medical applications such as cancer research (Cecco et al., 2015; Monica et al., 2011; Romano et al., 2014), the study of asthma (Hinks et al., 2016; Torres et al., 2016; Schofield et al., 2019; Hinks et al., 2015), diabetes (Sarikonda et al., 2014; Li et al., 2015) and others (Carlsson, 2017; Nielson et al., 2015;

Rucco et al., 2015). Mapper was also applied to a variety of other disciplines, including genomic data analysis (Camara, 2017; Rizvi et al., 2017; Chang et al., 2013; Chan et al., 2013; Bowman et al., 2008), chemistry (Duponchel, 2018a; Lee et al., 2017), the study of aqueous solubility (Pirashvili et al., 2018), remote sensing (Duponchel, 2018b), soil science (Savir et al., 2017), agriculture (Kamruzzaman et al., 2017), sport (Alagappan, 2012) and voting pattern analysis (Lum et al., 2013).

Broadly speaking, the Mapper algorithm provides an approximate representation of the structure of the data, typically given as a point cloud, through a simplicial complex. This complex provides a synthesis of the main topological features of the data in the sense that similar data points are grouped into clusters, and clusters are connected forming loops, flares, etc. An important step in any Mapper implementation is a choice of a clustering procedure that will implement the required notion of similarity of data points. Given that all known clustering procedures display various levels of instability (von Luxburg, 2010), it is to be expected that Mapper will suffer from a similar problem, and indeed, Mapper instability has been well demonstrated (Carrière et al., 2018).

Our main contribution in this paper is a numerical measure of the instability of Mapper as a function of its input parameters. We demonstrate that our notion of instability can be used to select parameter ranges which make the corresponding Mapper output reliable.

To elucidate the problem, it is important to bear in mind that any practical use of Mapper on a dataset X requires a number of choices. In the classical Mapper implementation, we need to choose a real valued function $h\colon X \longrightarrow \mathbb{R}$ (known as a *filter* or a *lens*) and a collection of intervals $\{I_i\}_{i=1}^{t}$ covering $h(X)$, as can be seen in Figure 1. The latter choice involves at least two further parameters, as we need to choose both the length of the intervals and the amount of overlap between successive intervals. We also must choose a clustering method to apply on the bins $h^{-1}(I_i)$ to implement the required notion of similarity.

Because of the choices involved, the creators of Mapper remarked in their foundational paper (Singh et al., 2007) that the method is rather ad hoc, and posed the question of how to create a formal framework that would control the necessary choices and would provide a measure of reliability of a particular Mapper output. In this paper we provide an answer to this problem.

1.1 Contributions and Related Work

Following its many successful applications, several attempts have been made to reduce the number of choices required to create a Mapper output.

Dey, Mémoli, and Wang (2016, 2017) study the structure and stability of a stable signature for what they called *multiscale Mapper*, which uses a hierarchy of covers instead of a single one. However, it is not clear how to translate their findings to the context of the original Mapper.

Jeitziner, Carriére, Rougemont, Oudot, Hess, and Brisken (2017) develop a two-tier version of Mapper applied to clustering gene-expression data in order to identify subgroups. Their version of Mapper is tailored specifically to the type of data for which it was intended and does not require any user choices. Within its intended regime, this version of Mapper is stable. It is not clear at this stage, however, how to extend it to other contexts.

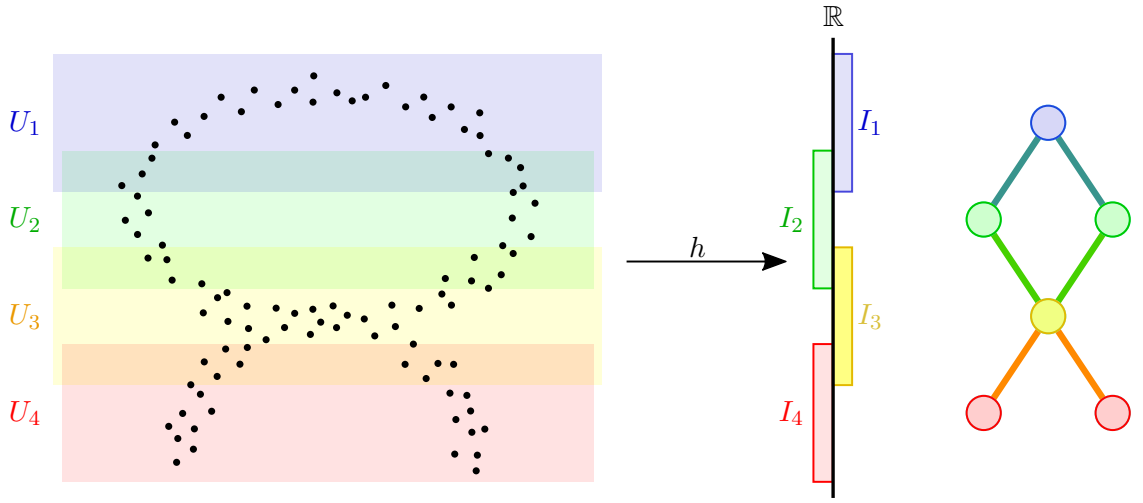


Figure 1: In the picture, X is represented by the black dots on the left, h assigns to each data point $(x, y) \in X$ its y coordinate and the intervals I^i are plotted as rectangles adjacent to the real line. Mapper first clusters each group of data points $h^{-1}(I^i) \subseteq X$ and views each cluster as a node. If two clusters $c_i \subseteq h^{-1}(I^i)$, $c_j \subseteq h^{-1}(I^j)$ $(i \neq j)$ share a point $x \in c_i \cap c_j$, the algorithm connects the nodes of c_i and c_j with an edge. The resulting graph is the output of the classical Mapper algorithm. Note that the resulting graph on the right looks like a simplified version of the point cloud X, exhibiting a hole on top and two flares at the bottom.

Dłotko (2019) sets out a procedure to generate Mapper covers by balls centred around selected points in the data. Once a cover is chosen a sequence of multiscale covers are obtained by expanding the ball sizes.

The work of Carrière, Michel, and Oudot (2018) represents ideas most similar to the present paper. Carrière and Oudot (2016) provide bounds on the stability of Mapper in a deterministic setting on manifolds by comparing it to the Reeb graph. This is achieved through a feature set obtained from an extended persistence diagram of the Mapper graph with respect to the filter function. In particular, the features correspond to loops and flairs in the Mapper graph. Through further statistical analysis (Carrière et al., 2018), bounds are determined on the expectation of the bottleneck distance between the features of the Mapper and Reeb graphs, assuming points are sampled from an underlying manifold. This provides a way to obtain confidence regions for specific features on the persistence diagram that may be used to identify reliable Mapper outputs.

By not restricting to data sampled from a manifold, our approach provides a more general setting than that of (Carrière et al., 2018). Points are only assumed to be sampled from an underlying probability distribution rather than a distribution on a smooth manifold. Furthermore the required covers may be chosen arbitrarily rather than being restricted to arising from an interval cover and a filter function.

By comparing Mapper outputs using a matching distance rather than the bottleneck distance, our approach will account for the size of features in terms of cluster size, not just

their presence. Though our instability measure is less precise, only applying to the whole Mapper output and not individual features, which in our case is only achieved by manually comparing outputs from a sample of outputs in the parameter space. However, these new idea allow us to study the effects of the choice of a clustering algorithm, which can even be picked to be different on different parts of the cover. This possible variability as well as any inherent instability of the chosen clustering procedure have not been investigated so far and we fill that gap here.

Despite the ubiquity of clustering techniques within unsupervised learning, it has proved difficult to establish a good theoretical foundation for this methodology. A lot of effort has been devoted to the study of quality and stability of clustering. Highlights include the famous impossibility theorem of Kleinberg (2003), who proved that there is no clustering procedure satisfying all of his natural axioms. This was taken up by Carlsson and Mémoli (2010), who proposed an axiomatic approach allowing them to provide an existence and uniqueness result for single-linkage clustering. More recently, Strazzeri and Sánchez-García (2018) provided a clustering procedure that satisfies Kleinberg's axioms after an alteration of the consistency axiom.

The work of Ben-David and Ackerman (2009) studied clustering quality measures rather than the clustering functions, which provides a richer setting in which an alternative to Kleinberg's axioms can be consistently stated.

In a similar vein, instability provides a measure of reliability of a particular output for the choice of input parameters. In particular, it will identify regions in the parameter space where the output is very sensitive to the changes of parameter values and so is typically less reliable. Much effort has been invested in studying clustering stability and while the theoretical principles are agreed upon, at present there is no standard implementation to determine its value. For an overview see (von Luxburg, 2010). In particular, methods of data perturbation and resampling have been successful in practice, for instance in the biomedical setting (Bittner et al., 2000; Ben-Hur et al., 2002; Levine and Domany, 2001). Resampling methods such as bagging (Breiman, 1996, 1998) have also long been successfully applied within supervised leaning. A procedure using resampling methods and statistics derived form the Mapper algorithm (Riihimaki et al., 2019) has also been used to obtain very accurate classification results on tree species data.

The most comprehensive theoretical study of clustering stability by Ben-David and von Luxburg (2008) defined a notion of clustering stability and related it to properties of the decision boundaries of the algorithm. This is the starting point of the theoretical part of this work. We extend these notions to account for the considerably more complex Mapper construction.

This paper is organised as follows. In §1.1, we discuss some related work and its connections to the current paper. In §2, we give background on clustering stability required for the remainder of the paper. This allows us in §3 to set out how the ideas of Ben-David and von Luxburg Ben-David and von Luxburg (2008) can be generalised to the Mapper setting. In particular, we introduce Mapper functions in Definition 8, which provide a new way of expressing Mapper outputs. Crucially, this is used to define a similarity metric between Mapper functions, D_M in Definition 9. The Distance D_M captures the structure of the whole Mapper output and leads to the definition of our notion of instability of Mapper (Definition 11) with respect to a large class of clustering procedures.

In the remainder of the paper, we develop theoretical tools to provide bounds on the instability of Mapper and to understand the main contributing factors. To do this, in §4 we introduce another similarity measure D_∂, Definition 19. The pseudo distance D_∂ can be seen as a kind of interleaving distance, and it relates the instability to the Mapper cover, enabling us to obtain useful bounds in §5, Theorems 24 and 30. These theoretical results unravel the main reasons for the instability, which are summarised in Remarks 25 and 31. In §6, we study how to sharpen the bounds on instability obtained in §5 and prove in Theorem 36 that for a large enough sample size and under reasonably constrained conditions these bounds can be arbitrarily small. Implying that the Mapper instability under such conditions is also small. This means that Theorem 36 might be seen as a kind of stability theorem for Mapper and justifies the central experimental observations of §8.

In §7 we present an algorithm allowing us to experimentally obtain values of instability. This leads in section in §8 to experiments demonstrating our theoretically derived reasons for instability and explain how the reasons for instability. Our following more geometrically complex experimental results suggest that regions of relatively high instability correspond to structural changes in the Mapper output. Hence local minima of the instability function with respect to parameter choices are good candidates for parameter selection allowing us to study Mapper through variations of all the parameters. In particular our theoretically derived reasons for instability can also be used to intemperate these observations.

2. Clustering Stability

The question of assessing the quality and stability of clustering procedures has attracted a lot of attention in recent years. In our discussion of Mapper stability, we will build on the foundational work on clustering stability by Ben-David and von Luxburg (2008). Therefore, we begin by introducing our setting in similar terms to theirs.

By a *clustering* of a metric space (U, D) we will mean a partition of U into s disjoint subsets or clusters. Equivalently, we may think of a clustering as a function from U to a finite set of labels. In assessing the performance of a particular clustering procedure, the choice of labels to denote the clusters will typically be unimportant, which motivates the following definition.

Definition 1 *Let (U, D) be a metric space and let F denote the set of all functions $f : U \longrightarrow \{1, 2, \ldots, s\}$. Then a clustering of (U, D) is an element of*

$$\mathcal{F} \coloneqq F/\!\sim,$$

where $f \sim g$ if there is a permutation π of the set $\{1, 2, \ldots, s\}$ such that $f = \pi g$.

To assess the efficiency of a particular clustering procedure we need a clustering quality function, which assigns a notional cost or error to a clustering procedure. The objective of a clustering procedure is then to minimise the cost. Let $M_1(U)$ denote the space of all probability measures on U (with respect to the Borel σ-algebra). For the purposes of this paper, a clustering quality function is a function which assigns a real number (the cost) to a choice of clustering and a choice of a probability measure on U. In other words, a clustering quality function is a map

$$Q \colon \mathcal{F} \times M_1(U) \longrightarrow \mathbb{R}.$$

Clustering is usually presented as an algorithmic procedure, though implicitly most of these constructions have an objective criterion around which the data is to be partitioned. Clustering with respect to a quality function, therefore covers a large variety of clustering processes including most common cluttering procedures. For example, any center-based clustering and spectral clustering fit this description. During the theoretical part of this work it is assumed that clustering algorithms are a method for obtaining a minimum of some objective function, through this function may not be formally stated as part of a given procedure.

Example 1 *To make the previous statement more transparent, consider the K-means clustering. In this case, $Q(g, P)$ measures the expected distance between any point drawn according to the probability distribution P and the cluster centre assigned to that point by the clustering function g. We give the explicit formula for this quality function in Equation 3.*

In the following definition we make the assumption that the clustering quality function has a unique minimizer, and we provide a discussion of the validity of this assumption in what follows.

Definition 2 *Given a probability measure $P \in M_1(U)$, the* optimal *clustering of U is defined as the function $f \in \mathcal{F}$ which minimizes $Q(-, P)$:*

$$f = \operatorname*{argmin}_{g \in \mathcal{F}} Q(g, P).$$

The optimal clustering gives rise to a clustering map

$$C \colon M_1(U) \to \mathcal{F}, \quad P \mapsto \operatorname*{argmin}_{g \in \mathcal{F}} Q(g, P). \tag{1}$$

The clustering f in Definition 2 is only well defined if $Q(\cdot, P)$ has a unique global minimum, which is a commonly made assumption in the literature (Caponnetto et al., 2006; Ben-David et al., 2006, 2007; Ben-David and von Luxburg, 2008; Ribeiro et al., 2016) and will be also be our starting point in this paper. A main reason for this restriction is that in this work we want to understand the relation between the user-selected parameters of the input and the stability of the outcome. In the presence of more local minima of the quality function $Q(\cdot, P)$, clustering instability may be dominated by other phenomena, for example, the symmetry of the data. This case will be discussed in the follow-on work. In fact, as demonstrated by (Ben-David et al., 2007, Theorem 4), k-means is stable if and only if there is a unique global minimiser, so this assumption is quite reasonable. More generally, in (Ben-David et al., 2006, Theorem 15), it is proved that multiple global minimisers with symmetry imply instability. While the presence of underlying symmetries is a main reason for a unique minimizer not to occur, this is very unusual in real data sets. In addition since underlying symmetries are known to cause instability we are interested in situations where this does not occur.

When working on a finite sample of $X = (X_1, \ldots, X_n) \in U^n$, we use another clustering quality function

$$Q_n : \mathcal{F}_n \times U^n \longrightarrow \mathbb{R},$$

which we call the empirical quality function. Unless stated otherwise, we assume that the quality function does not depend on the order of $X_1, \ldots, X_n$.

Example 2 *The empirical K-means quality function for $K = s$ clusters on a finite sample $X = (X_1, \ldots, X_n) \in U^n$ computes the average distance between points in the sample and their corresponding cluster centroid*

$$Q_n(f, X) = \frac{1}{n} \sum_{j=1}^{n} \sum_{k=1}^{K} \mathbb{1}_{f(X_j)=k} D(X_j, c_k), \tag{2}$$

where $D(X_j, c_k)$ denotes the distance between the point $X_j \in U$ and the cluster centre c_k and $\mathbb{1}_{f(X_j)=k}$ is an indicator function,

$$\mathbb{1}_{f(X_j)=k} = \begin{cases} 1, & f(X_j) = k \\ 0, & f(X_j) \neq k. \end{cases}$$

The continuous counterpart of Q_n for K-means clustering is given by:

$$Q(f, P) = \sum_{k=1}^{K} \int_{x \in U} \mathbb{1}_{f(x)=k} D(x, c_k)\, dP(x). \tag{3}$$

Definition 3 *Let $\mathcal{F}_n^X$ denote the space of clusterings of X. Given a point sample $X = (X_1, \ldots, X_n) \in U^n$, define the* optimal *empirical clustering $f^n \in \mathcal{F}_n^X$ of U as*

$$f^n = \operatorname*{argmin}_{g \in \mathcal{F}_n} Q_n(g, X), \tag{4}$$

if $n \geq 1$ and set f^n to be constant for $n = 0$. The optimal empirical clustering gives rise to a clustering map

$$C_n \colon U^n \to \mathcal{F}_n^U, X \mapsto \operatorname*{argmin}_{g \in \mathcal{F}_n} Q_n(g, X)$$

where $\mathcal{F}_n^U$ is the union of all $\mathcal{F}_n^X$ for $X \in U^n$.

Similarly to Definition 2, the clustering f of Definition 3 may not exist. In addition, even if such a global minimum exists, it may not be computable by the clustering algorithm. For example, the empirical clustering quality function for the K-means clustering in Equation 2 need not have a global minimum. However, nearest neighbour clusterings (von Luxburg et al., 2008) or approximation schemes (Ostrovsky et al., 2006) have empirical quality function with a unique global minimum and algorithms to compute them. For the theoretical part of this work, we will assume that $Q_n(-, X)$ has a unique global minimum.

Remark 4 *In practice, the clustering quality function and empirical quality function are related. Intuitively, Q_n is a discretised version of Q, and we will make the additional assumption that Q_n is uniformly consistent with Q in the following sense. The functions $Q_n(f^n, X) \xrightarrow[n \to \infty]{} Q(f, P)$ in probability, uniformly over probability distributions $P \in M_1(U)$. More precisely, $\forall \epsilon > 0, \forall \delta > 0, \exists N \in \mathbb{N}$ such that $\forall n \geq N, \forall P \in M_1(U)$,*

$$P^n \left(|Q_n(f^n, X) - Q(f, P)| > \epsilon \right) \leq \delta.$$

We will need to be able to compare clusterings and for that we now recall the minimal matching distance. This is one of many measures of similarity developed for clusterings, and a good survey on this subject can be found in (Meilă, 2005).

Definition 5 *The minimal matching distance is a map $D_m : \mathcal{F}_n \times \mathcal{F}_n \longrightarrow \mathbb{R}$ that, for any two clusterings $f, g \in \mathcal{F}_n$ of a set of points $X = (X_1, \ldots, X_n)$, is defined by*

$$D_m(f, g) = \min_{\pi} \frac{1}{n} \sum_{j=1}^{n} \mathbb{1}_{f(X_j) \neq \pi g(X_j)},$$

where π runs over all permutations of the set $\{1, 2, \ldots, s\}$ and $\mathbb{1}_{f(X_j)\neq\pi g(X_j)}$ is an indicator function.

It is well known that D_m is a metric, and that it can be computed efficiently using a minimal bipartite matching algorithm. Given a distance between clusterings of finite samples, we may define the instability with respect to an empirical quality function and a distance. Here we consider the instability with respect to the minimal matching distance.

Any clustering $g \in \mathcal{F}_n^X$ on a finite point sample $X = (X_1, \ldots, X_n) \in U^n$ can be extended to a clustering $g \in \mathcal{F}$ on all of U in the following fashion. Consider the order $X_1 \leq X_2 \leq \ldots \leq X_n$, and denote by V_i the Voronoi cell of X_i, defined by

$$V_i \coloneqq \{y \in U | D(y, X_i) \leq D(y, X_j) \; \forall \; j > i \text{ and } D(y, X_i) < D(y, X_k) \forall \; k < i \; \}. \tag{5}$$

Note that $\{V_i\}_i$ forms a partition of U. In order to extend the clustering $g \in \mathcal{F}_n^X$ to U, we can simply assign the label $g(X_i)$ to all points of U in the Voronoi cell of the point X_i. Which is, we extend $g \in \mathcal{F}_n^X$ so that it is constant on each Voronoi cell. For technical reason that will become clear in the context of Mapper, we also assume that the Voronoi extension of an empty sample labels all of U as a single cluster.

Given an empirical quality function Q_n, using the clustering function C_n of Definition 3 and the minimal matching distance, we obtain the composition

$$\mathcal{I}(Q_n) \colon U^n \times U^n \xrightarrow{C_n \times C_n} \mathcal{F}_n^U \times \mathcal{F}_n^U \xhookrightarrow{i} \mathcal{F}_{2n} \times \mathcal{F}_{2n} \xrightarrow{D_m} \mathbb{R}, \tag{6}$$

where $\mathcal{F}_n^U$ is the union of all clustering functions $\mathcal{F}_n^X$ on X for every $X \in U^n$. To define the inclusion i, we extend a clustering of n points to a clustering of all of U via Voronoi cells as just explained and focus only on the labels assigned to subsets of $2n$ points.

We would like the function $\mathcal{I}(Q_n)$ to be a random variable with respect to the probability measure on U^n, induced by a probability measure on U. From now on we restrict to quality functions such that $\mathcal{I}(Q_n)$ is a random variable, which we justify in the appendix.

Definition 6 *Let (U, D) be a metric space equipped with an n-point clustering quality function $Q_n \colon \mathcal{F}_n \times U^n \longrightarrow \mathbb{R}$ and a probability measure $P \in M_1(U)$. Then the clustering instability is given by*

$$\mathrm{InStab}_{\mathrm{Clustering}}(Q_n, P) = \mathbb{E}\left(\mathcal{I}(Q_n)\right),$$

where the expectation is taken over probability product measures of P on pairs of n-samples in $U^n \times U^n$.

3. Comparing Mapper Functions

We now pass to the main part of this work. Our first goal is to provide a description of Mapper functions analogous to the representation of clusterings as functions introduced in Definition 1. A key part of our construction is a generalization of the minimal matching distance given in Definition 5 to a form suitable for comparing Mapper outputs. The extension works by taking into account the clustering information contained in the resulting complexes. Our new notion of distance between Mapper functions is then used to define instability of the Mapper procedure and to derive upper bounds for this instability in §5.

Let $(\mathcal{X}, D)$ be a metric space and let $\mathcal{U} = \{U_i\}_{i=1}^t$ be a cover of $\mathcal{X}$, that is $\mathcal{X} = \bigcup_{i=1}^t U_i$. Following standard Mapper terminology, we refer to the sets U_i as *bins*. In the classical Mapper algorithm, these bins are obtained by fixing a real valued function $h : \mathcal{X} \longrightarrow \mathbb{R}$ (known as a *filter function* or a *lens*), fixing a collection of intervals $\{I_i\}_{i=1}^t$ covering $h(\mathcal{X})$, and setting $U_i \coloneqq h^{-1}(I_i)$, as in Figure 1. Here, however, we do not assume, as we do not need to, that the cover $\{U_i\}_{i=1}^t$ of X is of this particular form.

In this paper, we will deal with a discrete and finite sample X drawn from a metric space $\mathcal{X}$. The cover $\{U_i\}_{i=1}^t$ of $\mathcal{X}$ restricts to a cover $\{U_i \cap X\}_{i=1}^t$ of the space X, and we will simply write U_i rather than $U_i \cap X$ to lighten the notation. We now use a clustering procedure to cluster each of the sets U_i, so that we have

$$U_i = V_1^i \cup \cdots \cup V_s^i.$$

A *Mapper output* is a simplicial complex where an n-simplex σ is an $(n+1)$-tuple of clusters

$$\sigma = (V_{j_1}^{i_1}, \ldots, V_{j_{n+1}}^{i_{n+1}})$$

with a nonempty intersection.

To avoid the labels of clusters in U_i being mixed up with those of U_j for $i \neq j$, we cluster each U_i separately, that is, a clustering of U_i is of the form

$$f_i : U_i \longrightarrow \left\{c_1^i, c_2^i, \ldots, c_s^i\right\}, \tag{7}$$

where the c_j^i are cluster labels. Similarly to §2, denote by F^i the collection of all functions of the form of Equation 7 and $\mathcal{F}^i = F^i/\sim$, where

$$f_i \sim g_i \Longleftrightarrow \exists \pi : f_i = \pi g_i,$$

with π denoting some permutation of the set $\left\{c_1^i, c_2^i, \ldots, c_s^i\right\}$. To simplify the notation, we assume that every U_i is partitioned into the same number s of clusters. However, all results hold when choosing a different s for each bin.

Given a probability measure on $\mathcal{X}$, $P \in M_1(\mathcal{X})$, we consider the probability measure induced on U_i by restricting P to U_i and setting

$$P_i = \frac{1}{P(U_i)} \cdot P \in M_1\left(U_i\right), \tag{8}$$

and setting P_i as the zero measure if $P(U_i) = 0$. Denote by $Q^i : \mathcal{F}^i \times M_1(U_i) \longrightarrow \mathbb{R}$ the clustering quality function used in U_i, and denote by

$$Q_n^i : \mathcal{F}^i \times U_i^n \longrightarrow \mathbb{R}$$

its empirical counterpart on size-n samples of U_i. As in Definition 2, the clustering quality function Q^i determines a unique optimal clustering for each set U_i, and taken together, these optimal solutions create an optimal Mapper output and a clustering function $C^i\colon M_1(U_i) \to \mathcal{F}^i$, for every $i = 1, \ldots, t$. In a similar way, Definition 3 and an empirical quality function Q^i_n determines a unique optimal empirical clustering for each U_i from which we obtain an optimal Mapper output and a clustering function $C^i_n\colon U^n_i \to \mathcal{F}^{U_i}_n$.

Remark 7 *As is now apparent, a Mapper output (as well as a Mapper function which we will discuss shortly) depends on the choice of a cover, a quality function as well as the particular sample drawn from the ambient metric space. Moreover, implicit in the choice of a quality function is a choice of a clustering procedure (see Defintion 3). We will refer to these choices collectively as* Mapper parameters. *In practice, these various choices usually come down to a list of real parameters. For example, in the standard Mapper algorithm, the cover U_i is the pullback of an interval cover of $\mathbb{R}$, which is specified through a choice of two parameters, resolution and gain. In this case, resolution is the number and size of intervals in the cover, while gain controls the size of the overlap of these intervals.*

Definition 8 *Let $\mathcal{X}$ be a metric space equipped with a cover U_i. Given a clustering $f_i \in \mathcal{F}^i$ for each member U_i of the cover we define the corresponding* Mapper function *as the function which assigns to each $x \in \mathcal{X}$, the set of clustering labels given to x by the clustering functions f_i, for $i = 1, \ldots, t$. In other words, we have*

$$f(x) = \{f_i(x) \,|\, i = 1, \ldots, t,\ x \in U_i\}\,,$$

for each $x \in \mathcal{X}$. We denote the set of all Mapper functions on $(\mathcal{X}, \{U_i\}_{i=1}^t)$ by $\mathcal{N}$ and $\mathcal{N}_n$ on a finite n-point sample $X \in \mathcal{X}^n$.

Note that for each $x \in \mathcal{X}$, the size of $f(x)$ depends only on the cover, since it is equal to the number of sets U_i that contain x.

Notice as well that a Mapper function contains more information than a Mapper output, which is an abstract simplicial complex constructed on the set of clusters. A Mapper function contains the information about the number of points in each cluster, and also in every nonempty intersection of those clusters. A Mapper output will be equivalent to a Mapper function if we label every simplex $\sigma = (V_0, V_1, \ldots, V_k)$ of the Mapper output by the number of points of X contained in the intersection

$$V_0 \cap \cdots \cap V_k$$

of the clusters that are the vertices of σ.

Let $X = (X_1, \ldots, X_n) \in \mathcal{X}^n$ be a point sample of $\mathcal{X}$. Then, for each $1 \leq i \leq t$, denote

$$X^i = \{X_1, \ldots, X_n\} \cap U_i$$

and let $n_i = n_i(X)$ be the number of elements in X^i. We now introduce a Mapper version of definition (5).

Definition 9 *Given a point sample $X = (X_1, \ldots, X_n) \in \mathcal{X}^n$ of $\mathcal{X}$, we define a distance function $D_M : \mathcal{N}_n \times \mathcal{N}_n \longrightarrow \mathbb{R}$ which, for any two Mapper functions $f, g \in \mathcal{N}_n$, is given by*

$$D_M(f,g) = \min_{\pi} \frac{1}{n} \sum_{j=1}^{n} \mathbb{1}_{f(X_j) \neq \pi g(X_j)},$$

where $\pi = \bigoplus_{i=1}^{t} \pi^i$, each π^i runs over all permutations of $\left\{c_1^i, \ldots, c_s^i\right\}$, and $\mathbb{1}_{f(X_j)\neq \pi g(X_j)}$ is an indicator function.

Remark 10 *For two Mapper functions f, g on $\mathcal{X}$ covered by $\{U_i\}_{i=1}^t$, the matching distance $D_m(f_i, g_i)$ of definition 5 counts the proportion of points of X^i for which f_i and g_i disagree. Since the clustering of each U_i corresponds to the vertices of the Mapper output, considering D_m on each U_i would give no information about the higher dimensional simplicies of the Mapper output. However, Definition 9 takes into account not only the points that fall into different vertices of f and g, but also all the edges and the higher dimensional simplices to which the Mapper functions f and g assign different values.*

A drawback of D_M it that it can see certain intuitively larger changes of vertex labeling as equally distant. Consider the following example. Assume that $\mathcal{X}$ is covered by three sets U_1, U_2, U_3 and that each of these sets is clustered into two clusters labeled c_1^i, c_2^i for $i = 1, 2, 3$. Let $f(x) = \{c_1^1, c_1^2, c_1^3\}$, $g(x) = \{c_1^1, c_1^2, c_2^3\}$, $h(x) = \{c_1^1, c_2^2, c_2^3\}$, and $f(y) = g(y) = h(y)$ for all other points $y \neq x$. Provided the clustering labels remain unchanged, then $D_M(f,g) = D_M(f,h)$, despite the fact that h differs from f on two clusters, and it differs from g on only one cluster. However D_M does has the advantages of taking into account edge information and being simple to work with from both a theoretical and a practical perspective.

Since D_M generalizes D_m, we use D_M to generalize Definition (6) to a notion of instability of Mapper. As before, we assume that the metric space $\mathcal{X}$ is equipped with a cover $\{U_i\}_{i=1}^t$. We choose an empirical quality function $Q_{n_i}^i$ for each U_i. Denote by $\mathcal{N}_n^X$ the set of all Mapper functions on $X \in \mathcal{X}^n$ with cover $\{X^i\}_{i=1}^t$ and $\mathcal{N}_n^{\mathcal{X}}$ the union of $\mathcal{N}_n^X$ for all choices of X. Each $Q_{n_i}^i$ determines a corresponding clustering function $C_{n_i}^i$ which we use to define an instability function $\mathcal{I}(\{Q_{n_i}^i\}_{i=1}^t)$ as the composite

$$\mathcal{I}(\{Q_{n_i}^i\}_{i=1}^t) = \mathcal{X}^n \times \mathcal{X}^n \xrightarrow{\coprod_{i=1}^t C_-^i \times C_-^i} \mathcal{N}_n^{\mathcal{X}} \times \mathcal{N}_n^{\mathcal{X}} \overset{i}{\hookrightarrow} \mathcal{N}_{2n} \times \mathcal{N}_{2n} \xrightarrow{D_M} \mathbb{R}. \tag{9}$$

The function $\mathcal{I}(\{Q_{n_i}^i\}_{i=1}^t)$ will be measurable if and only if each $\mathcal{I}(Q_{n_i}^i)$ is measurable. This is because it follows from the definitions of D_M and D_m that the pre-image of a measurable set under $\mathcal{I}(\{Q_{n_i}^i\}_{i=1}^t)$ is a union of the the pre-images of measurable sets for functions $\mathcal{I}(Q_{n_i}^i)$.

Definition 11 *Fix Mapper parameters on $\mathcal{X}$ by choosing quality functions $\{Q_{n_i}^i\}_{i=1}^t$ defined on a cover $\{U_i\}_{i=1}^t$ of $\mathcal{X}$, and a probability measure $P \in M_1(\mathcal{X})$. These choices are made so that $\mathcal{I} = \mathcal{I}(\{Q_{n_i}^i\}_{i=1}^t)$ is a random variable, as discussed at the end of §2.*

The instability of the Mapper algorithm on size-n samples is defined as

$$\mathrm{InStab}_{\mathrm{Mapper}}(\{Q_{n_i}^i\}_{i=1}^t, n, P) = \mathbb{E}\left(\mathcal{I}(\{Q_{n_i}^i\}_{i=1}^t)\right),$$

where the expectation is taken over the probability product measures of P on pairs of n-samples in $\mathcal{X}^n \times \mathcal{X}^n$.

4. Mapper Boundary Distance

To compare clustering functions on a metric space (U, D), Ben-David and von Luxburg (2008) introduced a distance function that captures the size of the regions of U on which two clustering functions disagree. We now expand upon and generalise this boundary distance to the Mapper setting and use it to provide upper bounds of the Mapper instability in the following section.

Definition 12 *Let $(\mathcal{X}, D)$ be a metric space with cover $\{U_i\}_{i=1}^t$. Then given a Mapper function $f \in \mathcal{N}$, define the boundary of each f_i to be*

$$\partial(f_i) = \partial(f_i^{-1}(c_1^i)) \cup \cdots \cup \partial(f_i^{-1}(c_s^i)) \tag{10}$$

where each $\partial(f_i^{-1}(c_j^i))$ is the usual topological boundary of the subset $f_i^{-1}(c_j^i)$ taken over U_i. Following the established conventions, we will refer to $\partial(f_i)$ as the decision boundary *of f_i.*

Intuitively, $\partial(f_i)$ consists of the points of discontinuity of f_i, that is, the points lying in the boundary of some cluster, and an illustration of $\partial(f_i)$ is provided in Figure 2a. As U_i is a metric space, $\partial(f_i)$ can be described using an equivalent metric condition, which defines the boundary ∂A of any subset $A \subseteq U_i$ by

$$\partial A = \{x \in U_i \mid D(x, A) = D(x, A^c) = 0\}, \tag{11}$$

where $A^c = \mathcal{X} \setminus A$ is the complement of A in the metric space $\mathcal{X}$ and the distance of a point $x \in U_i$ from a set $A \subseteq U_i$ is defined as usual by

$$D(x, A) = \inf \{D(x, y) \mid y \in A\}.$$

Remark 13 *If $t = 1$ and $U_1 = \mathcal{X}$, then we recover the notion of boundary for clustering seen in (Ben-David and von Luxburg, 2008). In the case when any f_i is the constant function on each connected component of U_i, that is there is a single cluster in each component, then*

$$\partial f_i = \emptyset$$

and this is the only way to achieve this. For clustering a connected U_i it is not of particular interest to study data with a single cluster and so this does not cause many problems. Mapper constructions however, would commonly consider U_i with a single cluster, so it will be impotent to incorporate this into our work.

To avoid unnecessary technicalities, two clusterings will be considered different if and only if their values differ outside the intersection of their boundaries. Hence, we work on the set of all clusterings $f_i \in \mathcal{F}^i$ that represent elements in the space of equivalence classes

$$\mathcal{F}_\partial^i = \mathcal{F}^i / \sim,$$

where $f_i \sim g_i$ if and only if

$$\exists \pi : f_i(x) = \pi g_i(x), \forall x \in U_i - \partial(g_i), \qquad \text{and} \tag{12}$$
$$\exists \pi' : g_i(x) = \pi' f_i(x), \forall x \in U_i - \partial(f_i),$$

where $\pi, \pi' \in \Sigma_s$ are permutations of the set of labels.

Definition 14 *Let f be a Mapper function constructed using clustering functions f_i of the sets U_i. Then the decision boundary $\partial(f)$ of the Mapper function f can be defined using the decision boundaries of the functions f_i by*

$$\partial(f) = \bigcup_{i=1}^{t} \partial(f_i).$$

We denote by $\mathcal{N}_\partial$ the set of Mapper functions $f \in \mathcal{N}$ such that each f_i is an element of $\mathcal{F}_\partial^i$.

For any $\gamma > 0$, we may define the γ-tube of f_i to be

$$\mathbb{T}_\gamma(f_i) = \{x \in U_i \mid D\left(x, \partial(f_i)\right) \leq \gamma\},$$

as is the case in (Ben-David and von Luxburg, 2008) and Figure 2 illustrates the construction $\mathbb{T}_\gamma(f_i)$. However as discussed in Remark 13, in the case of Mapper covers, it becomes more relevant to consider U_i for which a clustering has no boundary or where the bounder can intuitively be seen as fully or partially lying outside U_i. Therefore, following the suggestion in (Ben-David and von Luxburg, 2008) to avoid reference to the boundary, we define for any $\gamma > 0$, the γ-tube of f_i to be the closed set

$$T_\gamma(f_i) = \overline{\{x \in U_i \,|\, \exists y \in U_i \colon d(x,y) \leq \gamma \text{ and } f_i(x) \neq f_i(y)\}}.$$

The next proposition expresses the relationship between $T_\gamma(f_i)$, $\mathbb{T}_\gamma(f_i)$ and $\partial(f_i)$.

Proposition 15 *For all $\gamma > 0$,*

$$\partial(f_i) \subseteq \mathbb{T}_\gamma(f_i) \subseteq T_\gamma(f_i).$$

When U_i is a complete and convex metric space, then $\mathbb{T}_\gamma(f_i) = T_\gamma(f_i)$.

Proof Let f_i be a clustering of U_i, $x \in \mathbb{T}_\gamma(f_i)$ and $\gamma > 0$. Then there is a $b \in \partial(f_i)$ such that $D(x,b) \leq \gamma$. By definition of the metric condition on the boundary in Equation 11, for any $\epsilon > 0$ the open ball $B(b,\epsilon)$ in U_i contains a points y such that $f_i(y) \neq f_i(x)$. By the triangle inequality we have that $D(x,y) \leq D(x,b) + D(b,y) = \gamma + \epsilon$. As a closed set, it contain its limit points and ϵ can be arbitrarily small, so $x \in T_\gamma(f_i)$ and hence $\mathbb{T}_\gamma(f_i) \subseteq T_\gamma(f_i)$. In addition we have that $\partial(f_i) \subseteq \mathbb{T}_\gamma(f_i)$ by definition of $\mathbb{T}_\gamma(f_i)$. Assuming now that U_i is convex and complete, then for any points $x, y \in U_i$ such that $d(x,y) \leq \gamma$ there is metric line segment between them containing a boundary point b. As this boundary point lies on the metric line segment $D(x,b) \leq \gamma$. Any point in $T_\gamma(f_i)$ is a limit point of such a pair of points, therefore $T_\gamma(f_i) \subseteq \mathbb{T}_\gamma(f_i)$ in this case. ■

Definition 16 *We define the γ-tube around a Mapper function $f \in \mathcal{N}$ to be*

$$T_\gamma(f) = \bigcup_{i=1}^{t} T_\gamma(f_i). \tag{13}$$

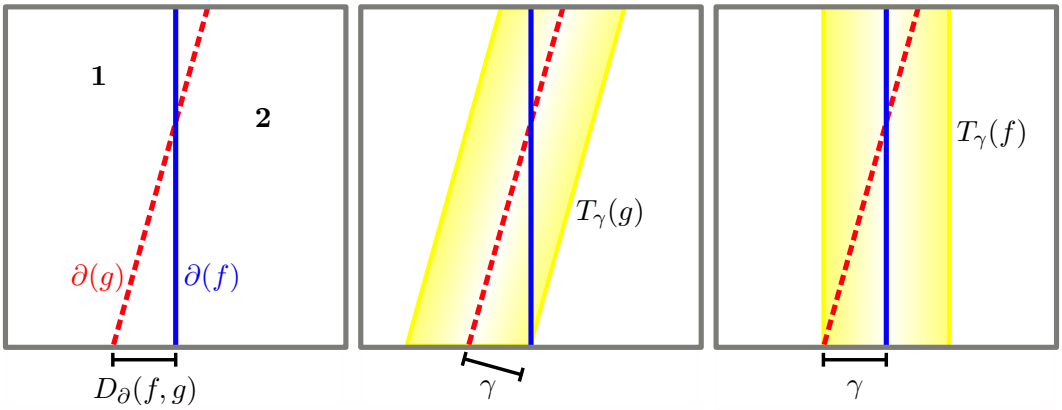


Figure 2: Assume that $t = 1$, so that a Mapper function f coincides with the clustering function f_1. In (a), everything left from the solid vertical blue line is labelled by f as cluster 1 and everything right from that line is labelled by f as cluster 2. Hence, $\partial(f)$ coincides with the solid line. Analogously, the left and right side of the dashed tilted line correspond to clusters 1 and 2 of g, respectively. Hence, $\partial(g)$ is precisely the dashed line. $D_\partial(f,g)$ is also illustrated in (a). In particular, $\forall\gamma > D_\partial(f,g)$, f and g agree both outside the γ-tube of g (*i.e.*, $f \triangleleft T_\gamma(g)$; see (b)) and outside the γ-tube of f (*i.e.*, $g \triangleleft T_\gamma(f)$; see (c)).

Figure 2 illustrates the construction $T_\gamma(f_i)$. If two clusterings $f_i, g_i \in \mathcal{F}_\partial^i$ agree outside the γ-tube of f_i, we will write $g_i \triangleleft T_\gamma(f_i)$. Thus the condition $g_i \triangleleft T_\gamma(f_i)$ holds if and only if for all x, y in the complement $U_i - T_\gamma(f_i)$ of the γ-tube of f_i we have that

$$f_i(x) = f_i(y) \Leftrightarrow g_i(x) = g_i(y).$$

Definition 17 *Given Mapper functions $f, g \in \mathcal{N}_\partial$, we say that g is contained in the γ-tube of f, written $g \triangleleft T_\gamma(f)$, if for all x, y in the complement of $T_\gamma(f)$*

$$f(x) = f(y) \Leftrightarrow g(x) = g(y).$$

It is clear that this statement is equivalent to saying that,

$$g \triangleleft T_\gamma(f) \Longleftrightarrow g_i \triangleleft T_\gamma(f_i) \text{ for all } i = 1, \ldots, t.$$

The mass $P\left(T_\gamma(f)\right)$ of the γ-tube $T_\gamma(f)$ with respect to the probability measure P depends on the overlap between the bins U_i. We have the following natural estimate.

Proposition 18 *The mass $P\left(T_\gamma(f)\right)$ of the tube $T_\gamma(f)$ is bounded by the mass of the γ tubes $T_\gamma(f_i)$ as follows:*

$$\max_{i=1,\ldots,t} P\left(T_\gamma(f_i)\right) \leq P\left(T_\gamma(f)\right) \leq \sum_{i=1}^{t} P\left(T_\gamma(f_i)\right).$$

The inequality on the left becomes an equality when all the elements U_i are contained in a single U_j. The inequality on the right becomes an equality when the bins U_i are all disjoint.

Proof Since $T_\gamma(f) = \bigcup_{i=1}^{t} T_\gamma(f_i)$, we have $P\left(T_\gamma(f_i)\right) \leq P\left(T_\gamma(f)\right)$ for all $i = 1, \ldots, t$ and hence, $\max_{1\leq i\leq m} P\left(T_\gamma(f_i)\right) \leq P\left(T_\gamma(f)\right)$, proving the inequality on the left. The other inequality follows in a similar way.

Turning to the second part of the Proposition, if there is some $1 \leq i_0 \leq t$ such that $T_\gamma(f_j) \subseteq T_\gamma(f_{i_0})$ for all $1 \leq j \leq t$, then $T_\gamma(f) = T_\gamma(f_{i_0})$. Hence $P\left(T_\gamma(f)\right) = P\left(T_\gamma(f_{i_0})\right) = \max_{1\leq i\leq t} P\left(T_\gamma(f_i)\right)$, realizing the lower bound.

Similarly, if $T_\gamma(f_i) \cap T_\gamma(f_j) = \emptyset$ for all $i \neq j$, then $P\left(T_\gamma(f)\right) = \sum_{i=1}^{t} P\left(T_\gamma(f_i)\right)$, realizing the upper bound. ■

Definition 19 *Let f and g be Mapper functions in $\mathcal{N}_\partial$. The boundary distance D_∂ is defined by*

$$D_\partial(f, g) = \inf\left\{\gamma > 0 \,|\, f \triangleleft T_\gamma(g) \text{ and } g \triangleleft T_\gamma(f)\right\}.$$

The metric D_∂ is therefore an interleaving distance between the γ-tubes of the functions f and g.

Example 3 *Once the distance parameter $\epsilon > 0$ and sample X are fixed, the Voronoi extension (see Equation 5) of neighbourhood clustering provides a unique clustering $f_\epsilon(X)$. Any quality function with a unique minimum at this clustering function will be a quality function corresponding to ϵ-neighbourhood clustering. Therefore as clustering is equivalent to Mapper with a single bin, an empirical quality function would be*

$$Q_n(f, X) = D_\partial(f, f_\epsilon(X)).$$

Similarly given the probability measure P from which X is sampled, define ϵ-neighbourhood clusterings on P as the function $f_\epsilon(P)$ whose clusters are the equivalence classes of connected components of P with distance less than ϵ form each over. This extends to the whole space by assigning labels outside the support to be the same label as that of their nearest labelled point (the boundary assignment in F_∂ making no difference). Then set:

$$Q(f, P) = D_\partial(f, f_\epsilon(P)).$$

For a large enough sample size the point sample X will be a dense enough representative of the support of P in probability. So, these functions will be uniformly consistent in the sense of Remark 4.

Remark 20 *If some U_i is unbounded then D_∂ may be infinite. In practice the support of P will always be bounded, hence if U_i is unbounded, then we may restrict $\mathcal{X}$ and U_i to some bounded subset containing the support of P. Therefore unless stated otherwise, we assume from now on that each U_i is bounded. In this case we note (and leave it to the reader to check) that the condition in Equation 12 makes D_∂ a metric. Without this restriction, D_∂ is only a pseudo-metric, as is also the case for clusterings.*

Furthermore as observed in Remark 29, the boundary $\partial(f)$ may be empty, however as $T_\gamma(f)$ does not depend on the boundary and $D_\partial(f,g)$ is well defined for any pair of Mapper function $f, g \in \mathcal{N}_\partial$.

To get more information regarding the boundary metric, we need to examine in a bit more detail the relationship between the space $\mathcal{N}$ of all Mapper functions and the spaces $\mathcal{F}_i$ of clusterings of the individual sets U_i in the cover of $\mathcal{X}$. We have the following.

Lemma 21 *There exists a bijection*

$$\phi : \mathcal{N} \longrightarrow \prod_{i=1}^{t} \mathcal{F}^i.$$

Proof Let ϕ be a map

$$\begin{array}{rccc} \varphi: & \mathcal{N} & \longrightarrow & \prod_{i=1}^t \mathcal{F}^i \\ & f & \longmapsto & (f_1, \ldots, f_t)\,, \end{array}$$

where each $f_i \colon U_i \longrightarrow \{c_1^i, \ldots, c_s^i\}$ is a function defined as follows: for every $x \in U_i$, $f_i(x)$ is the only value in the singleton set $f(x) \cap \{c_1^i, \ldots, c_s^i\}$.

The inverse map to φ is given by the construction of a Mapper function f from clustering functions $f_1, \ldots, f_t$ as described in Definition 8. ∎

It follows that we can view $\mathcal{N}_\partial$ as the product $\prod_{i=1}^t \mathcal{F}_\partial^i$ and so the space $\mathcal{N}_\partial$ is naturally a product metric space in the following way. If f and g are represented as $f = (f_1, \ldots, f_t)$ and $g = (g_1, \ldots, g_t)$, then by the bijection of Lemma 21, it is straightforward to check that

$$D_\partial(f,g) = \max_{i=1,\ldots,t} D_\partial(f_i, g_i), \tag{14}$$

where $D_\partial(f_i, g_i)$ is the Mapper distance D_∂ restricted to clustering functions on a single U_i. From now on, we will think of $\mathcal{N}_\partial$ as the product metric spaces $\left(\mathcal{F}_\partial^i, D_\partial\right)$. The metric D_∂ has several nice properties exhibited in the next Proposition, which provides a crucial step in the proof of Theorem 24 that provides an upper bound for the instability of Mapper.

Proposition 22 *Denote by $\{U_i\}_{i=1}^t$ a cover of the metric space $\mathcal{X}$. Then, with the notation above, the following properties hold:*

1. *Let $f, g \in \mathcal{N}_\partial$ and $\gamma > 0$. Then, $g \triangleleft T_\gamma(f)$ implies that $\partial(g) \subseteq T_\gamma(f)$.*
2. *Let $f, g \in \mathcal{N}_\partial$ and $\gamma > 0$ be such that $D_\partial(f,g) \leq \gamma$. Then for any choice of clustering labels, there exists a permutation π such that for all $x \in \mathcal{X}$,*

   $$f(x) \neq \pi(g(x)) \Longrightarrow x \in T_\gamma(g),$$

   where $\pi = \bigoplus_{i=1}^t \pi^i$, and π^i denotes a permutation of the set $\left\{c_1^i, \ldots, c_s^i\right\}$.
3. *If $\mathcal{X}$ is a subset of $\mathbb{R}^a$, the metric on $\mathcal{X}$ is induced by a norm on $\mathbb{R}^a$ and each $U_i \subseteq \mathcal{X}$ is compact, then $(\mathcal{N}_\partial, D_\partial)$ is relatively compact.*

Proof Let $f, g \in \mathcal{N}_\partial$ and $\gamma > 0$ be such that $g \triangleleft T_\gamma(f)$. By definition, this means that for each i, $g_i \triangleleft T_\gamma(f_i)$. By definition of the clustering boundary using the metric condition in Equation 11, for every $x \in \partial(g_i) - \partial U_i$ and every $\epsilon > 0$, the open ball $B(x, \epsilon)$ in U_i contains two points y and z such that $g_i(y) \neq g_i(z)$. Hence by definition of $g_i \triangleleft T_\gamma(f_i)$, for every $\epsilon > 0$, we have that $B(x, \epsilon) \cap T_\gamma(f)$ is nonempty. Since $T_\gamma(f_i)$ is a closed set, we have that $x \in T_\gamma(f_i)$, which implies that $\partial(g_i) \subseteq T_\gamma(f_i)$ for all i. Therefore, using Definition 14 and equality in Equation 13, we have that

$$\partial(g) = \bigcup_{i=1}^{t} \partial(g_i) \subseteq \bigcup_{i=1}^{t} T_\gamma(f_i) = T_\gamma(f),$$

which proves (1).

Let $f, g \in \mathcal{N}_\partial$ and $\gamma > 0$ be such that $D_\partial(f, g) \leq \gamma$. This means that $f_i \triangleleft T_\gamma(g_i)$ for all i. Suppose that for some $i = 1, \ldots, t$ and $j, a, b = 1, \ldots, s$ we set

$$A = \left(f_i^{-1}(c_j^i) - T_\gamma(g_i)\right) \cap \left(g_i^{-1}(c_a^i) - T_\gamma(g_i)\right) \text{ and } B = \left(f_i^{-1}(c_j^i) - T_\gamma(g_i)\right) \cap \left(g_i^{-1}(c_b^i) - T_\gamma(g_i)\right).$$

Since f_i takes the same value on A, B and $f_i \triangleleft T_\gamma(g_i)$, so g_i takes the same value on A and B. Therefore $a = b$ or $A = B = \phi$. Hence for every i and any choice of cluster labels on $f_i, g_i \in F_i$, we may construct a permutation π^i of the set $\left\{c_1^i, c_2^i \ldots, c_s^i\right\}$ such that $f_i(x) = \pi^i(g_i)(x)$ for $x \in U_i - T_\gamma(g_i)$. Setting $\pi = \bigoplus_{i=1}^t \pi^i$, the following holds for all $x \in \mathcal{X}$,

$$f(x) \neq \pi(g(x)) \Longrightarrow \exists i : f_i(x) \neq \pi^i(g_i(x)) \Longrightarrow \exists i : x \in T_\gamma(g_i) \Longrightarrow x \in T_\gamma(g),$$

where the last implication follows from Equation 13, proving (2).

It is stated in part 6 of (Ben-David and von Luxburg, 2008, Proposition 1) that provided U is a compact subset of $\mathbb{R}^a$, clusterings $\mathcal{F}_\partial$ on U are relatively compact in D_∂. Therefore, as we assume each $\mathcal{F}_\partial^i$ is relatively compact. Since $\mathcal{N}_\partial$ is endowed with a product metric D_∂, it follows that is $\mathcal{N}_\partial$ is relatively compact too, which proves (3). ■

Remark 23 *In Proposition 22, point (3), we assumed each bin U_i to be compact. Consider the classical Mapper algorithm, where a real-valued function $h \colon \mathcal{X} \longrightarrow \mathbb{R}$ and a collection of intervals $\{I_i\}_{i=1}^t$ covering $h(\mathcal{X})$ are used to define each bin U_i as $h^{-1}(I_i)$. Basic topology shows that a sufficient condition for each U_i to be compact consists of each interval I_i being of the form $[a_i, b_i]$ for some $a_i, b_i \in \mathbb{R}$ and the function h being a proper map, that is a function such that inverse images of compact subsets are compact. Furthermore, it is enough to assume $\mathcal{X}$ to be compact and h to be continuous to guarantee h to be a proper map. Notice also that if all bins are compact, so is $\mathcal{X}$, as a finite union of compact sets.*

5. Mapper Stability as a Function of Mapper Parameters

In this section, in Theorems 24 and 30 we prove two results that provide estimates of the instability of Mapper. Moreover, as we shall see, these results provide practical insights into how the stability of the Mapper algorithm can be affected by the specific choice of the Mapper parameters, including the filter function, the cover, the clustering algorithm, the metric and the sample size.

Throughout this section and the remainder of the paper, we assume that $\mathcal{X}$ is a metric space equipped with a probability measure $P \in M_1(\mathcal{X})$ and that $\mathcal{X}$ is given a cover $\mathcal{X} = \bigcup_{i=1}^{t} U_i$ such that each U_i is bounded. As before, we assume given quality functions Q^i on each U_i, together with the empirical quality functions Q^i_n. Furthermore, we will use the following additional notation.

- Denote by f the unique optimal Mapper function of $\mathcal{X}$, given by

$$f = \prod_{i=1}^{t} C^i(P_i).$$

  where C^i is an optimal clustering function defined in Equation 1

- Denote by f^n the unique optimal empirical Mapper function, that is the function

$$f^n = \prod_{i=1}^{t} C^i_{n_i}(X)$$

  obtained from size-n samples $X \in \mathcal{X}^n$ using the empirical clustering functions $C^i_{n_i}$.

Following Proposition 22 (2), we will also assume that all clusterings on U_i have connected clusters, however this automatically is the case for all common clustering procedures. We begin by generalizing the estimates obtained in (Ben-David and von Luxburg, 2008, Proposition 2).

Theorem 24 *Using the above assumption and notation, the instability of the Mapper algorithm satisfies*

$$\mathrm{InStab}_{\mathrm{Mapper}}(\{Q^i_{n_i}\}_{i=1}^{t}, n, P) \leq 2\Big(P(T_\gamma(f)) + P\left(D_\partial(f^n, f) > \gamma\right)\Big),$$

where $\gamma \geq 0$ *and*

- $P(T_\gamma(f))$ *denotes the mass of the* γ*-tube of* f*,*
- $P\left(D_\partial(f^n, f) > \gamma\right)$ *denotes the probability that the optimal empirical Mapper function* f^n *satisfies* $D_\partial(f^n, f) > \gamma$ *when each* f^n_i *is extended in* U_i *using Voronoi cells as given in Equation 5*

Proof

Define the following three collections of size-n samples $X' = (X_1, \ldots, X_n) \in \mathcal{X}^n$:

- Let $M_{\leq\gamma}$ be the set of $X' \in \mathcal{X}^n$ for which $D_\partial(f^n, f) \leq \gamma$.
- Let $M_{>\gamma}$ be the set of $X' \in \mathcal{X}^n$ for which $D_\partial(f^n, f) > \gamma$.

In particular following Remark 23, we have that $\mathcal{X}^n = M_{\leq\gamma} \cup M_{>\gamma}$ as even when a sample is empty f_i^n is assumed to extend to the constant clustering and D_∂ is well defined all of $\mathcal{N}_\partial$. Without loss of generality, let us assume that the permutation π for which the minimum value of D_M is attained (see Definition 9) is the identity. By Definition 11,

$$\mathrm{InStab}_{\mathrm{Mapper}}(\left\{Q_{n_i}^i\right\}_{i=1}^t, n, P) = \mathbb{E}\left(\mathcal{I}(\{Q_{n_i}^i\}_{i=1}^t)\right).$$

To simplify notation, we will write InStab for the left hand side of the above equation. Let $f^n = \prod_{i=1}^t C_{n_i(X')}^i(X')$ and $g^n = \prod_{i=1}^t C_{n_i(X'')}^i(X'')$ denote the optimal empirical Mapper functions for samples $X', X'' \in \mathcal{X}^n$, respectively. Recall that, using the Voronoi cell construction in Equation 5, Mapper functions f^n and g^n can be extended to the $2n$-point sample $X = (X', X'') \in \mathcal{X}^{2n}$. Then by Equation 9, taking D_M over all point in $X = (X_1, \ldots, X_{2n})$ and using the triangle inequality,

$$\begin{aligned}\mathrm{InStab} &= \int_{X\in\mathcal{X}^{2n}} D_M(f^n(X), g^n(X))dP^{2n}(X)\\ &\leq \int_{X\in\mathcal{X}^{2n}} \left(D_M(f^n(X), f(X)) + D_M(f(X), g^n(X))\right) dP^{2n}(X)\end{aligned}$$

where we note that each of the two terms now depends only on the variables either f^n or g^n, respectively. Therefore, we can now write

$$\begin{aligned}\mathrm{InStab} &= 2\int_{X\in\mathcal{X}^{2n}} D_M(f^n(X), f(X))dP^{2n}(X)\\ &= 2\left(\int_{\substack{X'\in M_{\leq\gamma},\\ X''\in\mathcal{X}^n}} D_M(f^n(X), f(X))dP^{2n}(X) + \int_{\substack{X'\in M_{>\gamma},\\ X''\in\mathcal{X}^n}} D_M(f^n(X), f(X))dP^{2n}(X)\right).\end{aligned}$$

Since $D_M(f^n(X), f(X)) \in [0,1]$ and using Defintion 9 for D_M, we obtain

$$\mathrm{InStab} \leq 2\left(\frac{1}{2n}\int_{\substack{X'\in M_{\leq\gamma},\\ X''\in\mathcal{X}^n}} \sum_{i=1}^{2n} \mathbb{1}_{f^n(X_i)\neq f(X_i)} dP^{2n}(X) + P\left(X'\in M_{>\gamma}\right)\right).$$

If f^n is obtained from a sample in $M_{\leq\gamma}$, then Proposition 22 (3) gives that for all $x \in \mathcal{X}$,

$$f^n(x) \neq f(x) \Longrightarrow x \in T_\gamma(f).$$

On the other hand, by definition, we have $P\left(M_{>\gamma}\right) = P\left(D_\partial(f^n, f) > \gamma\right)$. Therefore, we conclude

$$\begin{aligned}\mathrm{InStab} &\leq 2\left(\frac{1}{2n}\int_{\substack{X'\in M_{\leq\gamma},\\ X''\in\mathcal{X}^n}} \sum_{i=1}^{2n} \mathbb{1}_{X_i\in T_\gamma(f)} dP^{2n}(X) + P\left(D_\partial(f^n, f) > \gamma\right)\right)\\ &= 2\left(\frac{1}{2n}\cdot 2n\int_{x\in M_{\leq\gamma}} \mathbb{1}_{x\in T_\gamma(f)} dP + P\left(D_\partial(f^n, f) > \gamma\right)\right)\\ &= 2\left(P(T_\gamma(f)) + P\left(D_\partial(f^n, f) > \gamma\right)\right).\end{aligned}$$

■

Remark 25 (Reasons for instability - Part I) *Theorem 24 can be used to identify the effect of particular parameter choices on the instability of the Mapper output as follows. The bound becomes large if $P(T_\gamma(f))$ is large, when the mass is concentrated around the decision boundary $\partial(f)$ of the optimal clustering f. This may happen when any of the following conditions hold (but note that these conditions are not sufficient for Mapper instability).*

(a) The decision boundaries $\partial(f_i)$ lie in a highly dense area.

(b) The decision boundaries $\partial(f_i)$ are 'long' in the sense of a suitably defined path distance along $\partial(f_i)$.

(c) There is low overlap between bins.

Moreover, Proposition 18 suggests $P(T_\gamma(f))$ can also be large if this holds:

(d) The decision boundaries of different members of the cover are relatively far apart.

Indeed, small changes to decision boundaries that are far apart necessarily increase the distance between the Mapper functions. This is not always true for decision boundaries that are close since they are more likely to mismatch on the same points, see Figure 3 for an illustration.

Even if $P(T_\gamma(f))$ is small, $P\left(D_\partial(f_n, f) > \gamma\right)$ can still increase the bound, which happens if:

(e) The decision boundaries $\partial(f_i^n)$ vary a lot with the choice of the sample.

While points (a), (b) and (e) above are generalizations of those stated in (Ben-David and von Luxburg, 2008, §3), the phenomena described in (c) and (d) are unique to Mapper, since they involve interactions within the cover.

We now explore in more detail the instability of the Mapper output that result from parameter choices. High instability suggests that the Mapper output varies significantly with small variations of the input data. In particular, it is not surprising that Mapper instability increases if the decision boundaries $\partial(f_i)$ vary a lot with slight changes in the input sample. However, it is hard to identify explicitly the situations that make the term $P\left(D_\partial(f^n, f) > \gamma\right)$ large. To deal with this, in Theorem 30 we provide an upper bound for $P\left(D_\partial(f^n, f) > \gamma\right)$ in terms that more clearly depend on the choice of Mapper parameters. While in general this leads to a less sharp bound, we gain a greater insight into how these variables affect the instability of Mapper.

Remark 26 *To state Theorem 30, we make the following assumptions on the quality functions $Q_n^i \colon \mathcal{F}_n \times U_i^n \longrightarrow \mathbb{R}$ and $Q^i \colon \mathcal{F}^i \times M_1(U_i) \longrightarrow \mathbb{R}$.*

1. *The functions Q_n^i and Q^i have a unique global minimizer $f_i \in \mathcal{F}_n$, as we are assuming throughout the paper.*

2. *The functions Q_n^i are continuous, with respect to the topology on $\mathcal{F}_n \times U_i^n$ given by the metric D_∂.*

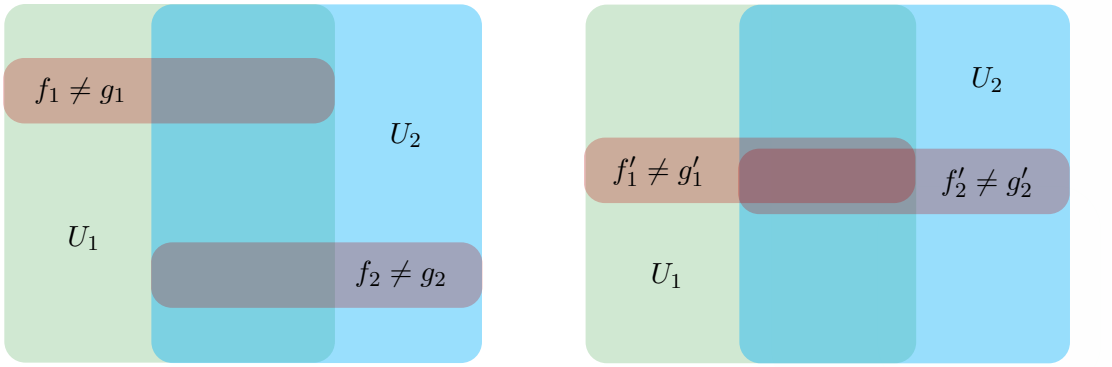


Figure 3: The images show two overlapping bins U_1 and U_2 (green and blue respectively) as well as the regions (red) where clustering functions assigned to each bin do not agree. On the left, the region in U_1 where f_1, g_1 do not agree, does not intersect the region where f_2, g_2 disagree in U_2. On the right, the mismatch regions between f_1', g_1' in U_1 and f_2', g_2' in U_2 have the same size as their counterparts on the left diagram. However, the ones on the right have a large intersection. Therefore assuming that point samples in the regions are similar, $D_M(f,g) > D_M(f',g')$ while $D_m(f_1,g_1)$ and $D_m(f_2,g_2)$ have similar values to $D_m(f_1',g_1')$ and $D_m(f_2',g_2')$, respectively.

3. *The functions Q_n^i are uniformly consistent with the functions Q^i in the sense that for every i and every $\gamma > 0$, $Q_n^i(f_i^n, X) \xrightarrow[n\to\infty]{} Q^i(f_i, P_i)$ in probability, uniformly over all probability distributions $P_i \in M_1(U_i)$. That is $\forall \epsilon > 0, \forall \delta > 0, \exists N \in \mathbb{N}$ such that $\forall n \geq N, \forall P_i \in M_1(U_i)$,*

$$P_i\left(|Q_{n_i}^i(f_i^n, X) - Q^i(f_i, P_i)| > \epsilon\right) \leq \delta. \tag{15}$$

Note that N does not depend on P_i. For future reference, we denote by $N^i(\epsilon, \delta)$ the minimum of the set of the numbers N for which the condition of Equation 15 is satisfied.

Ben-David showed that uniform consistency holds for the algorithm constructing the global minimum of the k-means objective function (Ben-David, 2007). Similar results occur with the normalized cut used in spectral clustering (von Luxburg et al., 2008). For more on consistency of clustering algorithms, see (von Luxburg et al., 2008).

The next proposition shows that for a large enough sample size, Equation 15 guarantees that the minimal quality function and empirical minimal quality functions will be close in the boundary metric.

Proposition 27 *In addition to the assumptions above, let us also assume that*

- *each U_i is compact; and*

- *for every* $0 < \eta < 1$ *and* $\zeta > 0$ *there is* $N \in \mathbb{N}$ *such that for all* $n \geq N$,

$$P_i(|Q^i(f_i^n, P_i) - Q^i_{n_i}(f_i^n, X)| \leq \zeta) \geq \eta,$$

for each $i = 1, \ldots, t$. *Note that is property differs form Equation 15 as it depends on the minimal empirical function* f_i^n *within both quality functions.*

Then for each $\epsilon > 0$,

$$P(D_\partial(f_i^n, f_i) \leq \epsilon) \xrightarrow[n\to\infty]{} 1.$$

The second assumption in the proposition is similar to the uniform consistency assumption of Equation 15, in that it states that for a large enough n the functions Q^i and $Q^i_{n_i}$ give similar values at a specified point, in particular that the functions take similar values at f_i^n.

Proof If U_i has zero mass that the second condition on the proposition implies that f_i^n is the unique minimise of Q^i, hence $f_i^n = f_i$ and $D_\partial(f_i^n, f_i) = 0$ satisfying the conclusion of the proposition. So assume form now on that U_i has nonzero mass. By (Ben-David and von Luxburg, 2008, Proposition 3) (whose proof is the same under our conditions) for all $\epsilon \geq 0$ there is an $\xi \geq 0$, such that for each $g \in \mathcal{F}^i_\partial$,

$$|Q^i(g, P_i) - Q^i(f_i, P_i)| \leq \xi \Longrightarrow D_\partial(f_i^n, f_i) \leq \epsilon.$$

Hence by the triangle inequality

$$\begin{aligned} P_i(D_\partial(f_i^n, f_i) \leq \epsilon) &\geq P_i\left(|Q^i(f_i^n, P_i) - Q^i(f_i, P_i)| \leq \xi\right) \\ &\geq P_i\left(|Q^i_{n_i}(f_i^n, X) - Q^i(f_i, P_i)| + |Q^i(f_i^n, P_i) - Q^i_{n_i}(f_i^n, X)| \leq \xi\right) \\ &\geq P_i\left(|Q^i_{n_i}(f_i^n, X) - Q^i(f_i, P_i)| \leq \frac{\xi}{2} \text{ and } |Q^i(f_i^n, P_i) - Q^i_{n_i}(f_i^n, X)| \leq \frac{\xi}{2}\right). \end{aligned} \tag{16}$$

On the other hand from Equation 15, since for any $0 \leq \delta < 1$ there is an $N \in \mathbb{N}$ such that for $n \geq N$,

$$P_i\left(|Q^i_{n_i}(f_i^n, X) - Q^i(f_i, P_i)| > \zeta\right) \leq 1 - \delta$$

which implies that

$$P_i\left(|Q^i_{n_i}(f_i^n, X) - Q^i(f_i, P_i)| \leq \zeta\right) \geq \delta.$$

Therefore picking $\zeta = \frac{\xi}{2}$, combining with the second assumption in the proposition and Equation 16, since U_i have nonzero mass we obtain that

$$P_i(D_\partial(f_i^n, f_i) \leq \epsilon) \xrightarrow[n\to\infty]{} 1.$$

The statement of the proposition now follows. ■

To state Theorem 30, we now introduce the term $\iota(n)$, which describes in probabilistic terms the dependence of the behaviour of the Mapper function on the properties of the clusterings for each U_i.

Definition 28 *If $P\left(D_\partial(f_i^n, f_i) \leq \gamma\right) \neq 0$ for all $i = 1, \ldots, t$, denote by $\iota(n)$ the real number $\iota(n) \geq 0$ such that*

$$P\left(D_\partial(f^n, f) \leq \gamma\right) = \iota(n) \prod_{i=1}^{t} P\left(D_\partial(f_i^n, f_i) \leq \gamma\right). \tag{17}$$

If $P\left(D_\partial(f_i^n, f_i) \leq \gamma\right) = 0$ for some $i = 1, \ldots, t$, then define $\iota(n) = 1$.

The relationship between $\iota(n)$ and n is not necessarily monotone. To see this, recall that, as stated in Equation 14, for any $g, h \in \mathcal{N}_\partial$, we have that

$$D_\partial(g, h) = \max_{i=1,\ldots,t} D_\partial(g_i, h_i).$$

For example, if for some $i = 1, \ldots, t$, $P\left(D_\partial(f_i^n, f_i) \leq \gamma\right)$ decreases at a slower rate than the others with respect to n, then the value of $\iota(n)$ will rise. Under the assumptions of Proposition 27, we have that $P(D_\partial(f_i^n, f_i) \leq \epsilon) \xrightarrow[n\to\infty]{} 1$, implying that

$$\iota(n) \xrightarrow[n\to\infty]{} 1, \tag{18}$$

so the behaviour of $\iota(n)$ for a large enough n is determined.

Remark 29 *It is however not clear what range of values $\iota(n)$ may take. Intuitively given a large enough point sample $X \in \mathcal{X}^n$, if $D_\partial(f_i^n, f_i) \leq \gamma$ for some $i = 1, \ldots, t$, this would indicate that the sample well represented the underlying probability distribution P_i on U_i. So the subset of the point sample contained in another bin U_j intersecting U_i would be more likely to well represent P_j. This in turn should result in a lower value of $P(D_\partial(f_j^n, f_j) \leq \gamma)$. More precisely for each $i = 1, \ldots, t$, we would expect that*

$$P(D_\partial(f_j^n, f_j) \leq \gamma \mid D_\partial(f_i^n, f_i) \leq \gamma) \geq P(D_\partial(f_j^n, f_j) \leq \gamma).$$

In this case, since the event $D_\partial(f^n, f) \leq \gamma$ is the intersection of events $D_\partial(f_i^n, f_i) \leq \gamma$, using conditional probability we obtain that

$$P\left(D_\partial(f^n, f) \leq \gamma\right) \geq \prod_{i=1}^{t} P\left(D_\partial(f_i^n, f_i) \leq \gamma\right).$$

In particular by Defintion 28, this implies that

$$\iota(n) \geq 1.$$

It then follows from Equation 18, that $\iota(n)$ is minimised as n grows. To make these points more precise we would require more information, especially regarding the properties of the clustering functions.

To find an upper bound on the term $P(D_\partial(f^n, f) > \gamma)$ of Theorem 24, we use properties of the cluster quality function $Q^i(-, P_i)$ in a neighbourhood of the global minimum f_i. Assuming that each U_i is compact, by (Ben-David and von Luxburg, 2008, Proposition 3)

(whose proof is the same under our conditions) for every $\gamma > 0$ and every $i = 1, \ldots, t$, there exists $\epsilon > 0$ such that for all $g \in \mathcal{N}$, written as $g = \prod_{i=1}^{t} g_i$, the condition that

$$|Q^i(g_i, P_i) - Q^i(f_i, P_i)| \leq \epsilon$$

for all $i = 1, \ldots, t$ implies that

$$D_\partial(g, f) \leq \gamma.$$

Let us denote by $S_{P_i}^{Q^i}(\gamma)$ the supremum of the set of all such ϵ. See Figure 4 for an illustration of what $S_{P_i}^{Q^i}(\gamma)$ represents.

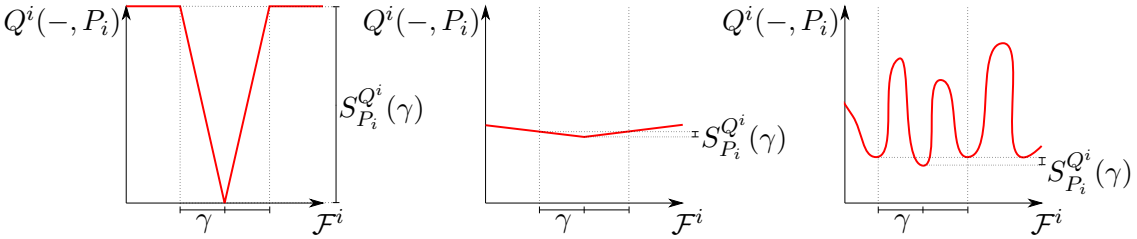


Figure 4: $S_{P_i}^{Q^i}(\gamma)$ measures how distinctly unique the global minimum of $Q^i(-, P_i)$ is, by looking at a ball of radius γ around the minimum of $Q^i(-, P_i)$. In the illustration, we identify $(\mathcal{F}^i, D_\partial)$ with a subset of the reals $(\mathbb{R}, d_{\text{Euclidean}})$. The quality function on the left has a very distinct global minimum, and hence a large $S_{P_i}^{Q^i}(\gamma)$. The other two functions exhibit different ways in which points can have values close to the global minimum and hence have a small $S_{P_i}^{Q^i}(\gamma)$.

We can now express an upper bound on instability which involves, among others, the mass of the bins that form the cover of $\mathcal{X}$.

Theorem 30 *Fix a sample size n, given the assumptions and notations presented at the beginning of the section, Remark 26 and that each U_i is compact. Then, for all $\gamma > 0, \delta > 0$, the instability of the Mapper algorithm satisfies*

$$\text{InStab}_{\text{Mapper}}(\{Q_{n_i}^i\}_{i=1}^t, n, P) \leq 2P(T_\gamma(f)) + 2\phi, \tag{19}$$

where $\phi \in [0, 1]$ has the form

$$\phi = 1 - \iota(n)\,(1 - \delta)^t \prod_{i=1}^{t} P(U_i) \cdot P\left(n_i \geq N^i(S_{P_i}^{Q^i}(\gamma), \delta)\right)$$

and the function N^i is defined in part (3) of Remark 26.

Proof Fix $\gamma > 0$ and some $\delta > 0$. We first find a lower bound for $P(D_\partial(f^n, f) \leq \gamma)$. This yields the upper bound ϕ for the term $P\left(D_\partial(f_n, f) > \gamma\right)$ of Theorem 24, from which we will conclude that

$$\text{InStab}_{\text{Mapper}}(\{Q_{n_i}^i\}_{i=1}^t, n, P) \leq 2P(T_\gamma(f)) + 2\phi.$$

We denote by $n_i \geq N^i(S_{P_i}^{Q^i}(\gamma), \delta)$ the event consisting of picking $X \in \mathcal{X}^n$ according to P^n such that $N^i(S_{P_i}^{Q^i}(\gamma), \delta)$ is greater than n_i. Since for any events A and B, $P(A) \geq P(A \cap B)$, for each $i = 1, \ldots, t$:

$$P\left(D_\partial(f_i^n, f_i) \leq \gamma\right) \geq P\left(\{D_\partial(f_i^n, f_i) \leq \gamma\} \cap \left\{n_i \geq N^i(S_{P_i}^{Q^i}(\gamma), \delta)\right\}\right). \tag{20}$$

By conditional probability, $P\left(A \cap B\right) = P\left(A \mid B\right) P(B)$ for any events A and B. In particular, the expression on the right hand side of Equation 20 is equal to

$$P\left(D_\partial(f_i^n, f_i) \leq \gamma \;\middle|\; n_i \geq N^i(S_{P_i}^{Q^i}(\gamma), \delta)\right) \cdot P\left(n_i \geq N^i(S_{P_i}^{Q^i}(\gamma), \delta)\right). \tag{21}$$

We now find a lower bound for the left multiplicand in Equation 21. By definition of $S_{P_i}^{Q^i}(\gamma)$,

$$|Q^i(f_i^n, P_i) - Q^i(f_i, P_i)| \leq S_{P_i}^{Q^i}(\gamma) \Longrightarrow D_\partial(f_i^n, f_i) \leq \gamma.$$

Hence, $P_i\left(D_\partial(f_i^n, f_i) \leq \gamma \;\middle|\; n_i \geq N^i(S_{P_i}^{Q^i}(\gamma), \delta)\right)$ is bounded from below by

$$P_i\left(|Q_{n_i}^i(f_i^n, X) - Q^i(f_i, P_i)| \leq S_{P_i}^{Q^i}(\gamma) \;\middle|\; n_i \geq N^i(S_{P_i}^{Q^i}(\gamma), \delta)\right), \tag{22}$$

Additionally, by definition of $N^i(S_{P_i}^{Q^i}(\gamma), \delta)$, if $n_i \geq N^i(S_{P_i}^{Q^i}(\gamma), \delta)$ then

$$P_i\left(|Q^i(f_i^n, P_i) - Q^i(f_i, P_i)| \leq S_{P_i}^{Q^i}(\gamma)\right) \geq 1 - \delta,$$

and therefore, the expression in Equation 22 is bounded below by $1 - \delta$. Hence by Equation 8 the left factor in Equation 21 is bounded below by

$$(1 - \delta) \cdot P(U_i).$$

This provides the following lower bound for the full expression in Equation 21:

$$(1 - \delta) \cdot P(U_i) \cdot P\left(n_i \geq N^i(S_{P_i}^{Q^i}(\gamma), \delta)\right),$$

and hence, using Equation 17, the following lower bound for $P\left(D_\partial(f_i^n, f_i) \leq \gamma\right)$:

$$\iota(n)\,(1 - \delta)^t \prod_{i=1}^{t} P(U_i) \cdot P\left(n_i \geq N^i(S_{P_i}^{Q^i}(\gamma), \delta)\right). \tag{23}$$

If we define ϕ so that $1 - \phi$ is the expression in Equation 23, then,

$$P\left(D_\partial(f^n, f) > \gamma\right) \leq \phi,$$

which combined with Theorem 24, provides us with Equation 19.

Finally we show that $\phi \in [0, 1]$. First note that $\phi \geq P\left(D_\partial(f_n, f) > \gamma\right) \geq 0$. On the other hand, the terms $\iota(n), (1 - \delta), P(U_i)$ and $P\left(n_i \geq N^i(S_{P_i}^{Q^i}(\gamma), \delta)\right)$ are non-negative. Therefore Equation 23 is non-negative. Subtracting this expression from 1 yields a result no larger than 1 and this is ϕ by definition. ■

We now discuss the consequences of Theorem 30 on the instability of Mapper.

Remark 31 (Reasons for instability - Part II) *Theorem 24 revealed that a large mass $P(T_\gamma(f))$ around the minimizer f of a Mapper quality function corresponded to an unstable Mapper output. Assuming the mass $P(T_\gamma(f))$ to be small, a closer look at the term ϕ introduced in Theorem 30 reveals the following additional reasons for the instability of a Mapper algorithm (but note that these conditions are not sufficient for Mapper instability).*

(A) A small sample size n makes ϕ large. The term $P\left(n_i \geq N^i(S^{Q^i}_{P_i}(\gamma), \delta)\right)$ decreases as n increases. While $\iota(n)$ need not decrease monotonically with n, we know by Equation 18 it does tend to 1 when n tends to infinity, under the assumptions of Proposition 27. So the term $\iota(n)$ can be ignored for a large sample, possibly even minimal using the reasoning of Remark 29. Therefore for a large enough sample size, ϕ becomes small.

(B) A small value of $P(U_i)$ for some $i = 1, \ldots, t$, makes ϕ large, i.e., *close to* 1.

(C) If a clustering quality function Q^i has many points with values very near the global minimum then ϕ is close to 1. *Indeed, if there are many local minima of $Q^i(-, P_i)$, $g_i \in \mathcal{F}^i$ such that $|Q^i(g_i, P_i) - Q^i(f_i, P_i)|$ is small, for the given global minimizer f_i of $Q^i(-, P_i)$, then $S^{Q^i}_{P_i}(\gamma)$ is small (see Figure 4 for an illustration), making $N^i(S^{Q^i}_{P_i}(\gamma), \delta)$ large, and in consequence, making ϕ large too.*

The above points add to the reasons for instability presented in Remark 25. The conditions given in (A) and (B) can be seen as the global and local versions, respectively, of a similar phenomenon, since a small $P(U_i)$ means that the proportion of sampled points from $\mathcal{X}$ that fall into a U_i is likely to be small.

The chosen clustering method and metric play a crucial role in the causes for instability of Remark 25 and in cause (C) in Remark 31. Among all the parameters selected for the classical Mapper algorithm given by a filter function and interval cover of $\mathbb{R}$, the weight $P(U_i)$ in (B) depends only on the cover $\{I_i\}_{i=1}^t$ of $\mathbb{R}$ and the filter function $h\colon \mathcal{X} \longrightarrow \mathbb{R}$. Hence, by choosing suitable $\{I_i\}_{i=1}^t$ and h, we would be able to control the value of $P(U_i)$, providing we have sufficient information of the distribution P.

Finally, notice that if (C) applies, this may produce not only instability but also inaccuracy. This would arise in situations when the global minimum is not distinct enough, which leads to a possible error in finding the minimizer. This is often a sign of a mismatch between the model and the data (Shamir and Tishby, 2010).

6. On the Sharpness of Bounds on Instability

In the previous section, we proved two theorems describing upper bounds on the instability of Mapper in terms of the behaviour of the Mapper parameters necessary to produce an output from some given data. In this Section, we discuss the efficiency of these estimates. To get a feel for the problem, let us first address the obvious question of the possible range of values for instability and its upper bounds. Let

$$Bound_{D_\partial} = 2P(T_\gamma(f)) + 2P\left(D_\partial(f^n, f) > \gamma\right),$$

denote the bound from Theorem 24. Let us also denote by

$$Bound_\phi = 2P(T_\gamma(f)) + 2\phi,$$

the bound from Theorem 30. Fix all parameters except $\gamma > 0$ and $\delta \in (0,1)$. Since $P(T_\gamma(f))$, $P\left(D_\partial(f_n, f) > \gamma\right), \phi \in [0,1]$, we have that

$$Bound_{D_\partial}, Bound_\phi \in [0,4].$$

In contrast, the instability is by definition an expectation over the image of D_M and $D_M(f,g) \in [0,1]$ for any $f, g \in \mathcal{N}_n$ (see Definition 9), so

$$\mathrm{InStab}_{\mathrm{Mapper}}(\left\{Q^i_{n_i}\right\}_{i=1}^t, n, P) \in [0,1].$$

This shows that the choice of specific values of the parameters γ and δ is crucial if we want to be able to control the value of instability, it particularly important to be able to obtain

$$\inf_{\gamma>0} Bound_{D_\partial}, \qquad \text{and} \qquad \inf_{\gamma>0,\delta\in(0,1)} Bound_\phi.$$

In the remainder of the section, we discuss choices of parameters γ and δ for which tight bounds are attained.

Remark 32 *We can make the following simple observation about varying γ and δ.*

1. *As γ increases, $P\left(T_\gamma(f)\right)$ increases and $P\left(D_\partial(f^n, f) > \gamma\right)$ decreases.*

2. *Analogously, as γ increases, each $S^{Q^i}_{P_i}(\gamma)$ increases, forcing $N^i(S^{Q^i}_{P_i}(\gamma), \delta)$ to increase, with the overall effect of making ϕ smaller. However, increasing γ also increases $P\left(T_\gamma(f)\right)$.*

3. *Similarly, as δ grows, each $N^i(S^{Q^i}_{P_i}(\gamma), \delta)$ decreases, which diminishes the value of ϕ. However, when δ grows, the value of $(1-\delta)$ gets smaller, which increases the value of ϕ.*

From Remark 32, we see that in general there is no straightforward way to identify optimal values of γ and δ. However, the following Corollary of Theorems 24 and 30 shows that to obtain useful boundaries we need to consider small values of γ.

Corollary 33 *If $\mathcal{X}$ is bounded then there exists some $\Gamma > 0$ such that for $\gamma \geq \Gamma$, we have*

$$1 \leq Bound_{D_\partial}(\gamma) \leq Bound_\phi(\gamma, \delta),$$

for all $\delta \in (0,1)$.

Proof If $\mathcal{X}$ is bounded, then there is some $\gamma > 0$ such that $P\left(T_\gamma(f)\right) \geq \frac{1}{2}$, hence the corollary follows from Theorem 24 and 30. ∎

Since $\mathrm{InStab}_{\mathrm{Mapper}}(\left\{Q^i_{n_i}\right\}_{i=1}^t, n, P) \leq 1$, an upper bound above 1 gives no information. A consequence of Corollary 33 is that large values of γ produce such large bounds. In the next theorem we show that under reasonable conditions selecting suitable $\gamma > 0$, $\phi > 0$ with a large enough n, make $Bound_{D_\partial}$ arbitrarily close to 0 and therefore to the instability.

Definition 34 *We call the pair (P, Q) consisting of a probability measure on metric space $(\mathcal{X}, D)$ and a clustering quality function Q on point samples $X \in \mathcal{X}^n$, a* proper pair *if all decision boundaries of the clustering function in the image of C (the associated clustering function of Q, see (2)) are of zero mass with respect to P.*

In most applications a proper pair would be expected. For example following Proposition 15 if U_i are convex and complete the we may take the boundary definition

$$T_\gamma(f_i) = \{x \in U_i \mid D\left(x, \partial(f_i)\right) \leq \gamma\},$$

In this case if U_i are also $\mathbb{R}^a$ and if the probability measure is obtained form a continuous probability distribution and the boundaries of f_i are possibly empty finite unions of Jordan arcs.

Remark 35 *In particular a proper pair implies that for $\gamma > 0$ and optimal clustering function f, tube $T_\gamma(f)$ may be of arbitrarily small mass. The clustering boundary ∂f is by definition a finite union of boundaries, hence nowhere dense. Therefore if a nonzero lower bound existed on $P(T_\gamma(f))$, then for any sequence γ_j such that $\gamma_j \to 0$ as $j \to \infty$, we would have that $\partial f = \bigcap_{j\in\mathbb{N}} T_{\gamma_j}(f)$ is of nonzero mass.*

Theorem 36 *Given the assumptions of Remark 26, Proposition 27, that each U_i is a bounded, convex , compact, complete subset of $\mathbb{R}^a$ and that each (P_i, Q^i) is a proper pair on $(\mathcal{X}, D)$. Then for each $1 > \epsilon > 0$, there is a $\gamma > 0$ and $N \in \mathbb{N}$, such that for $n \geq N$ the instability of the Mapper algorithm satisfies*

$$0 \leq \mathrm{InStab}_{\mathrm{Mapper}}(\left\{Q^i_{n_i}\right\}_{i=1}^t, n, P) \leq Bound_{D_\partial}(\gamma) \leq \epsilon. \tag{24}$$

Weakening boundedness of U_i to bounded support of P and on subset of $\mathbb{R}^a$ removing the compactness assumption, we obtain that

$$\mathrm{InStab}_{\mathrm{Mapper}}(\left\{Q^i_{n_i}\right\}_{i=1}^t, n, P) \xrightarrow[n\to\infty]{} 0.$$

Proof Pick $1 > \epsilon > 0$ and recall that

$$Bound_{D_\partial}(\gamma) = 2P(T_\gamma(f)) + 2P\left(D_\partial(f^n, f) > \gamma\right).$$

Following Remark 35 and Proposition 15, since U_i is convex complete and (P_i, Q^i) is a proper probability measure, $P_i(T_\gamma(f_i))$ becomes arbitrarily small as γ goes to zero. By Equation 13, we have $T_\gamma(f) = \bigcup_{i=1}^t T_\gamma(f_i)$, so we may choose $\gamma > 0$ so that

$$2P(T_\gamma(f)) \leq \frac{\epsilon}{3}.$$

By Proposition 27, we have $P(D_\partial(f_i^n, f_i) \leq \gamma) \xrightarrow[n\to\infty]{} 1$ and, in addition by Equation 14, we also have $D_\partial(f, g) = \max_{i=1,\ldots,t} D_\partial(f_i, g_i)$. Therefore, we may choose $N \in \mathbb{N}$, with $N \geq N'$ such that for all $n \geq N$

$$2P\left(D_\partial(f^n, f) > \gamma\right) \leq \frac{\epsilon}{3},$$

which proves Equation 24. By construction, $0 \leq \mathrm{InStab}_{\mathrm{Mapper}} \leq Bound_{D_\partial}(\gamma)$, so

$$\mathrm{InStab}_{\mathrm{Mapper}}(\left\{Q^i_{n_i}\right\}_{i=1}^t, n, P) \xrightarrow[n\to\infty]{} 0.$$

As pointed out in Remark 20, if the support of P is bounded and U_i is unbounded, then we may restrict $\mathcal{X}$ and U_i to some bounded subset containing the support of P. If U_i is a subset of $\mathbb{R}^a$ then we may take its closure, so this bounded subset may also be assumed to be closed, hence compact. These alterations of the cover do not change the value of the instability while allowing $Bound_{D_\partial}(\gamma)$ to be well defined. ■

Considering the Mapper output over the space of possible Mapper parameters, we would expect most choices of parameters to satisfy the conditioned of Theorem 36. Setting aside conditions on the underlying probability distribution, most other conditions can be satisfied by choosing a reasonable Mapper setup, such as the classical Mapper algorithm and a sensible clustering procedure. The exception to this is the assumption that the quality functions Q^i has a unique global minimizer. However as discussed below Definition 2, this is most likely caused by a symmetry of P_i in U_i, which we might interpret as a transition in the structure of the Mapper output at a particular choice of parameters. Therefore following Theorem 36 and as observed experimentally in Table 3 and Figure 5, for a large enough sample size, we would expect the values of instability over the parameter space to form regions of low instability separated by ridges of instability. In this sense, as Theorem 36 justifies the existence of regions of stability, it could be considered a stability theorem for Mapper.

In the case of clustering, Theorem 36 recovers partially the stability theorem (Ben-David et al., 2006, Theroem 10), hence may be seen as generalisation to the Mapper setting. In particular this suggests that the assumptions of Theorem 36, when considering the limit of Mapper instability may be further weakened.

Remark 37 *Equation 24 will not hold if $Bound_{D_\partial}(\gamma)$ is replace with $Bound_\phi(\gamma, \delta)$. Recall that*

$$Bound_\phi(\gamma,\delta) = 2\left(P(T_\gamma(f)) + 1 - \iota(n)\,(1-\delta)^t \prod_{i=1}^t P(U_i)\cdot P\left(n_i \geq N^i(S^{Q^i}_{P_i}(\gamma),\delta)\right)\right).$$

As shown in the proof of Theorem 36, the term $P(T_\gamma(f))$ can be made arbitrarily small for large n. By equation 18, we have $\iota(n) \xrightarrow[n\to\infty]{} 1$. Also by construction $P\left(n_i \geq N^i(S^{Q^i}_{P_i}(\gamma),\delta)\right)$ may be arbitrarily close to 1 if n is large enough. Therefore under the conditions of Theorem 36, it follows that

$$\inf_{\gamma>0,\ \delta\in(0,1)} Bound_\phi(\gamma,\delta) \xrightarrow[n\to\infty]{} 2\left(1 - \prod_{i=1}^t P(U_i)\right)$$

and $\prod_{i=1}^t P(U_i)$ is fixed by the choice of cover.

7. Computing Mapper Instability

In this section, we present a procedure for experimentally estimating the Mapper instability given in Definition 11. It is important to note that there is no standard procedure to determine clustering instability, and a discussion of the subject can be found in (von Luxburg, 2010). Our approach is to generalise to the Mapper setting a method for computing clustering instability detailed in (Ben-Hur et al., 2002), which is based on sub-sampling of the data. A similar approach applicable to points sampled from a manifold is taken by (Carrière et al., 2018), where the clustering procedure is chosen to be the neighbourhood clustering. On the other hand, in that work the authors are able to compute confidence measures for specific Mapper features, while here we need to consider the whole Mapper output. To interpret the stability of Mapper features from our perspective we make observations of variations over sampled outputs form the parameter space.

To begin with, we assume that all necessary Mapper parameters, as explained in Remark 7, have been selected and that we have a sample of n points taken independently and identically distributed (i.i.d) from an underlying probability distribution. Then we may computationally estimate the Mapper instability based on the method of k-fold cross validation as follows.

1. Split the data into $k \geq 2$ sub-samples. That is, choose $m, k \in \mathbb{N}$ such that $n = km$ and remove for each $1 \leq i \leq k$ the m points $m(i-1)+1$ to mi, leaving k sub-samples of $(k-1)m$ points.

2. Compute the Mapper distance between the Mapper functions of each pair of sub-samples, on the $(k-2)m$ points of their intersection.

3. Average the distances between Mapper functions restricted to the sub-samples by summing the distances and dividing by $\frac{(k-1)k}{2}$.

The outcome of this procedure is an approximation of the instability of the Mapper function.

Choosing a small k leads to inaccurate results since there are too few samples and the intersection between the samples is small. However too large a choice of k may result in samples that are too similar which in turn decreases the speed of computation as many more distances need to be calculated. Hence the best results are achieved with values of k and m in the middle of their range, such that m is not too large. Greater accuracy can still be gained by averaging the results of the procedure applied to several randomly shuffled copies of the dataset. We now explain the details of the procedure for computing the Mapper distance between the Mapper functions on two sub-samples.

Given a dataset X, we describe in Algorithm 1 a procedure to compute n times the Mapper distance $D_M(f,g)$ between two Mapper functions $f, g \in \mathcal{N}^X$ on a cover $\{U_i\}_{i=1}^t$ of X. We denote by

$$\{c_i^1, \ldots, c_i^{k_i}\} \text{ and } \{s_i^1, \ldots, s_i^{k_i}\}$$

the clusters of f and g in each U_i respectively, where k_i is the maximum number of clusters of either f or g in each U_i. If k_i is larger than the number of clusters, then the additional clusters are assumed to be empty. Additionally, with $l = \sum_{i=1}^t k_i$ let,

$$(c_{\zeta_1}^{\eta_1}, \ldots, c_{\zeta_l}^{\eta_l}) \tag{25}$$

be a size-ordered list of clusters of f, that is $|c_{\zeta_1}^{\eta_1}| \geq |c_{\zeta_2}^{\eta_2}| \geq \cdots \geq |c_{\zeta_l}^{\eta_l}|$.

Algorithm 1 is a recursive backtracking procedure, which is initialized with an upper bound, and a possible choice here is the total number of points in the sample. However, we will indicate shortly how to significantly improve this choice which will greatly shorten the computation time.

The mismatch between two clusters c_i^a and s_i^b is the symmetric difference $c_i^a \triangle s_i^b$ of the sets, consisting of the points that are elements of one of the sets but not the other. Algorithm 1 takes in order each cluster of $(c_{\zeta_1}^{\eta_1}, \ldots, c_{\zeta_l}^{\eta_l})$, and looks for the first cluster of g that has not yet been matched. We compute any additional mismatch that arises from any new matching. If the total mismatch exceeds the upper bound the algorithm backtracks and looks for a better matching. We replace the upper bound if a better one is obtained. Ordering the clusters in Equation 25 is therefore a good idea because obtaining a large mismatch is only possible if at least one of the clutters is large. If we obtain a large mismatch quickly, this reduces the execution time of the algorithm by reducing the number of possibilities that need to be checked.

Algorithm 1 Obtains n times Mapper distance $D_M(f, g)$ of Mapper functions f and g

Input

$(c_{\zeta_1}^{\eta_1}, \ldots, c_{\zeta_l}^{\eta_l})$	Size ordered list of cluster from f
Bound	Upper bound on the Mapper distance
$p \leftarrow 1$	Cluster position p in $(c_{\zeta_1}^{\eta_1}, \ldots, c_{\zeta_l}^{\eta_l})$
Mismatch $\leftarrow \emptyset$	Set of mismatched points
U_i-*matches* $\leftarrow \{s_i^1, \ldots, s_i^{k_i}\}$	Clusters of g_i not yet matched with f_i clusters

Output

$nD_M(f, g)$	n times Mapper distance between f and g

procedure DISTANCE
  for each member S of U_{ζ_p}-*matches* **do**
    NewMismatch $\leftarrow$ *Mismatch* $\cup$ $(c_{\zeta_p}^{\eta_p} \triangle S)$
    if $|$*NewMismatch*$|$ $<$ *Bound* **then**
      if $p = l$ **then**
        Bound $\leftarrow |$*NewMismatch*$|$
      else
        Matches $\leftarrow U_1$-*matches*,...,U_p-*matches* $- S$,...,U_l-*matches*
        Bound $\leftarrow$ DISTANCE($(c_{\zeta_1}^{\eta_1}, \ldots, c_{\zeta_l}^{\eta_l})$, *Bound*, $p+1$, *NewMismatch*, *Matches*)
  return *Bound*

A drawback of Algorithm 1 is that despite executing significantly faster than a procedure that considers all cluster permutations, computation time can still be slow. The main reason for this is that if the initial upper bound is large, improved bounds may only be obtained in small increments, requiring most permutations to be checked.

A very good upper estimate for the Mapper distance can be obtained by finding the permutations corresponding to the minimal matching distances $D_m(f_i, g_i)$ within each clustering of U_i. Then finding the size of the set of mismatched points across the Mapper functions corresponds to the permutation obtained by combining the optimal permutations in each U_i.

In practice, this upper bound can be obtained by performing Algorithm 1 restricted to each clustering on U_i and returning the minimal $Mismatch$ in addition to the corresponding *Bound*. An upper bound is then given by the size of the union of the mismatches from each U_i. Alternatively the optimal permutation within each U_i could be obtained using the Hungarian algorithm.

8. Experimental Tests for the Instability of Mapper

In this section we demonstrate how the procedure detailed in the previous section might be used to determine good Mapper outputs over varying parameter selections. Mapper is a standard tool from topological data analysis and there are several available implementations (Müllner and Babu, 2013; Müllner et al., 2010; Hendrik and Nathaniel, 2017). Our results were obtained using the Kepler Mapper (Hendrik and Nathaniel, 2017). The code for the experiment on resolution and gain used to produce Figure 5 can be obtained at (Burfitt, 2019).

We begging by presenting numerical experiments to investigate and demonstrate the causes of instability given in Remarks 25 and 31. In particular we focus on causes of instability unique to the Mapper algorithm.

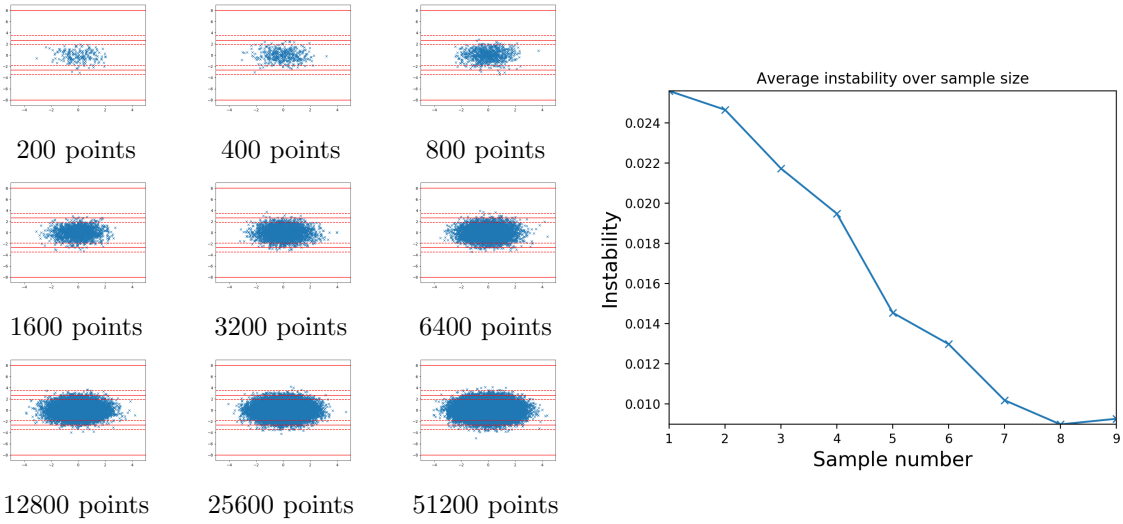
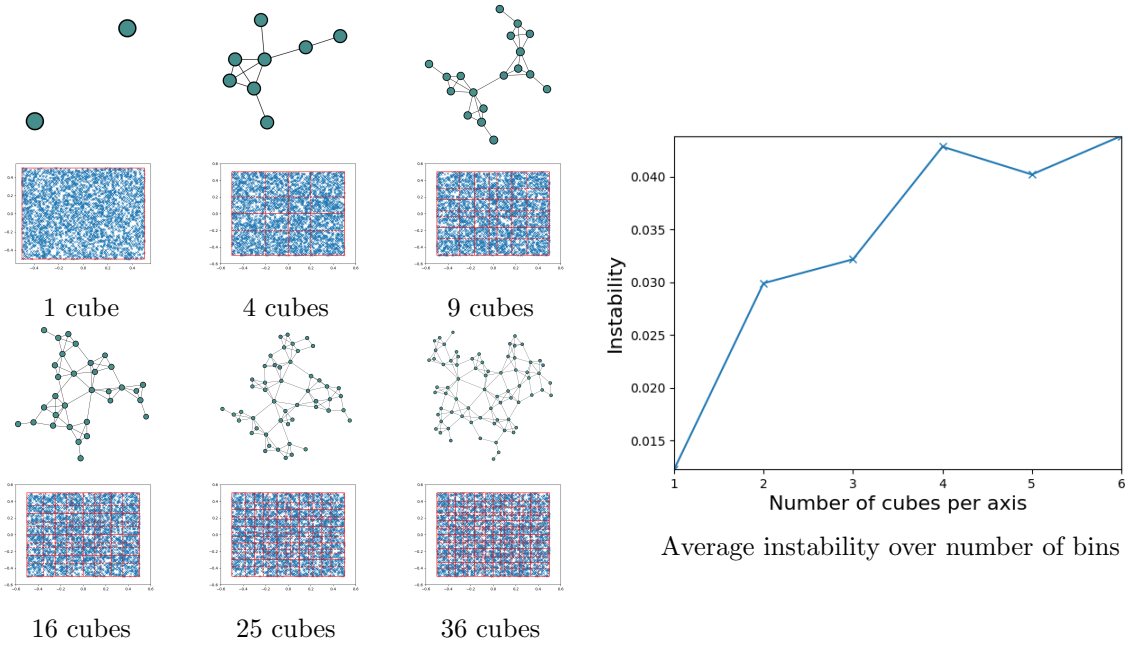


Table 1: On the left are nine samples from a bivariant Gaussian distribution centred at the origin. The first sample has 200 points doubled each time up to 51200 points. The dashed red lines denote the boundaries of the overlapping bins and the solid lines show the centre of the overlap. On the right we give a plot of the corresponding Mapper instability for each of the datasets on the left. The clustering procedure used was K-means with $K = 2$ cluster on 15 percent overlap between bins, the instabilities were averaged over 30 different samples and each instability was computed using 40 sub-samples. See §7 for details of the procedure.

Table 1 demonstrates a relationship between increasing numbers of points and lower values of instability as discussed in part (A) of Remark 31 and Theorem 36. While it is intuitively clear that larger samples should lower the instability, experiments of this kind allow one to quantify the sample size necessary to ensure that is is not, by itself, a source of instability.



Average instability over number of bins

Table 2: On the left 6 uniform samples of 3600 points from a unit square centred at the origin, alongside Mapper graphs obtained by increasing the number of bins per axis from 1 to 6. The dashed red lines denote the boundaries of the overlapping bins and the solid lines the centre of the overlap. On the right is a plot of the corresponding Mapper instabilities for each dataset to the left. The clustering procedure used was K-means with 2 clusters, there was 40 percent overlap between bins, the instabilities were averaged over 30 different samples and the instability for each sample was computed using 40 sub-samples. See §7 for details of the procedure.

Table 2 demonstrates the relationship between increasing numbers of bins and higher values of instability. In each case, we draw the same number of points from a uniform distribution in a unit square. As the sample size is constant, by increasing the number of bins, the number of points in each bin decreases, hence $P(U_i)$ decreases as explained by part (B) of Remark 31.

Observe also that the instability values in Table 1 are lower than all but the first value appearing in Table 2. This is a consequence of parts (b), (d) (e) of Remark 25 and part (C) of Remark 31, clustering decision boundaries in Table 2 become longer as the number of boxes increase, since the boxes are square depending on the sample the boundary many possible lines intersecting the square, this also increases the chance of boundaries lying far

apart when they are of opposite orientations in neighbouring boxes and the rectangle point samples in Table 2 do not give a clear place for a k-means decision boundary increasing the number of distinct clustering boundaries with a low quality value. On the other hand, in Table 1, the decision boundaries converge to a vertical line in all the boxes.

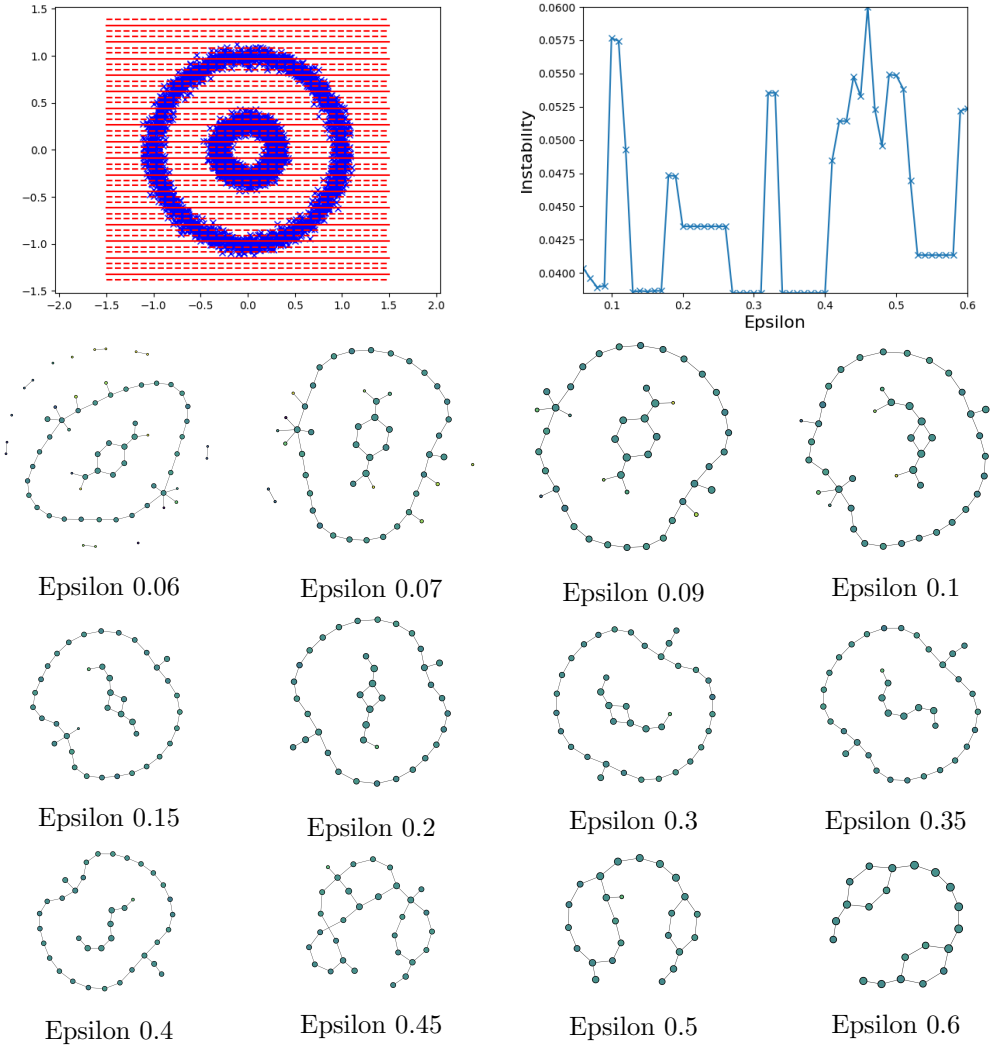


Table 3: The panel in the top left shows a dataset of 5000 points sampled with noise from two concentric circles. The dashed red lines denote the boundaries of the overlapping bins and the solid lines are the centres of the overlap. Displayed below are the Mapper graphs corresponding to the increasing values of ϵ, the parameter guiding the ϵ-neigbourhood clustering used to construct the Mapper outputs. The plot in the top right shows the instability as a function of ϵ. The bin overlap was 35 percent, with 17 bins and the instabilities were averaged over 30 different sub-samples, with 10 sub-samples in each case. See §7 for details of the procedure.

Table 3 considers a dataset with two noisy concentric circles. We produce a family of Mapper graphs using the ϵ-neighbourhood clustering with varying values of epsilon. The construction of quality functions for ϵ-neighbourhood cluster in the context of our theory is given in Example 3. The specific clustering procedure used was DBSCAN from the sklearn python package with the minimum sample size set to 0 so that it coincides with ϵ-neighbourhood clustering. For the values of epsilon of 0.06 and 0.09, the instability decreases due to the disappearance of noise represented by spurious small connected components in the Mapper graph. The major part of the structure of the Mapper graph remains the same, revealing both the inner and outer circle. Above the epsilon values of 0.1, there is a spike in the instability value corresponding to a loss of detail in the inner circle within the Mapper graph. A similar spike occurs around the 0.32 value of epsilon, corresponding to the loss of the inner circle from the Mapper graph. The final large increase in instability occurs around the 0.45 value of epsilon, and it corresponds to the gradual merging of the two circles in the Mapper graph, eventually stabilising with two connected cycles between 0.53 and 0.58 ending with a last rise in instability corresponding to the shrinking of one of the cycles.

We now pass to experiments that explore the dependence of the Mapper graph on the values of resolution and gain. Figure 5 presents a contour plot of the instability of Mapper on another dataset consisting of noisy concentric circles created by varying the percentage overlap between bins (gain) and the number of bins (resolution).

Similarly to the discussion on Table 3, it is possible to identify a number of global features within the plot with structural changes in the Mapper graph.

Running between bin numbers of 7 and 13, there is a diagonal of high peaks in instability. Restricting to odd number of bins, this range of peaks appears to correspond to the emergence of the inner circle within the Mapper graph. All graphs below the first distinct diagonal show the inner circle as a cluster without a cycle. Mapper graphs for odd bin numbers above the diagonal contain the structure of the inner circle.

Along the horizontal value of 14 bins, there is a relative rise in instability. This appears to correspond to the fact that if we use an even number of bins the correct structure of the inner circle is revealed.

The region determined by bin numbers from 8 to 12 and percentage overlaps from 25 to 50 is a negatively sloped diagonal of relatively high instability. This appears to correspond to the emergence in the Mapper graph of a new relatively large cluster attached to the structure of the outer circle forming a flare corresponding to either a number of points at the top or at the bottom of the outer circle.

Running between bin numbers 14 and 20 is another diagonal range in peaks of instability. These peaks seem to appear when restricted to even numbers of bins and correspond to the emergence of a better defined structure of the inner circle within the Mapper graphs.

Finally, the high instability in the top left hand corner of the contour plot appears to capture the moment when the part of the Mapper graph corresponding to outer circle breaks up.

The spikes and ridges in instability that occur around changes in the structure of the Mapper graph in Table 3 and Figure 5 are inaction to Theorem 36, explained by part (e) of Remark 25 and part (b) of Remark 31. This is because at the boundary values of epsilon between structural changes in the graph, the clustering function in some bins changes dramatically with the choices of sample.

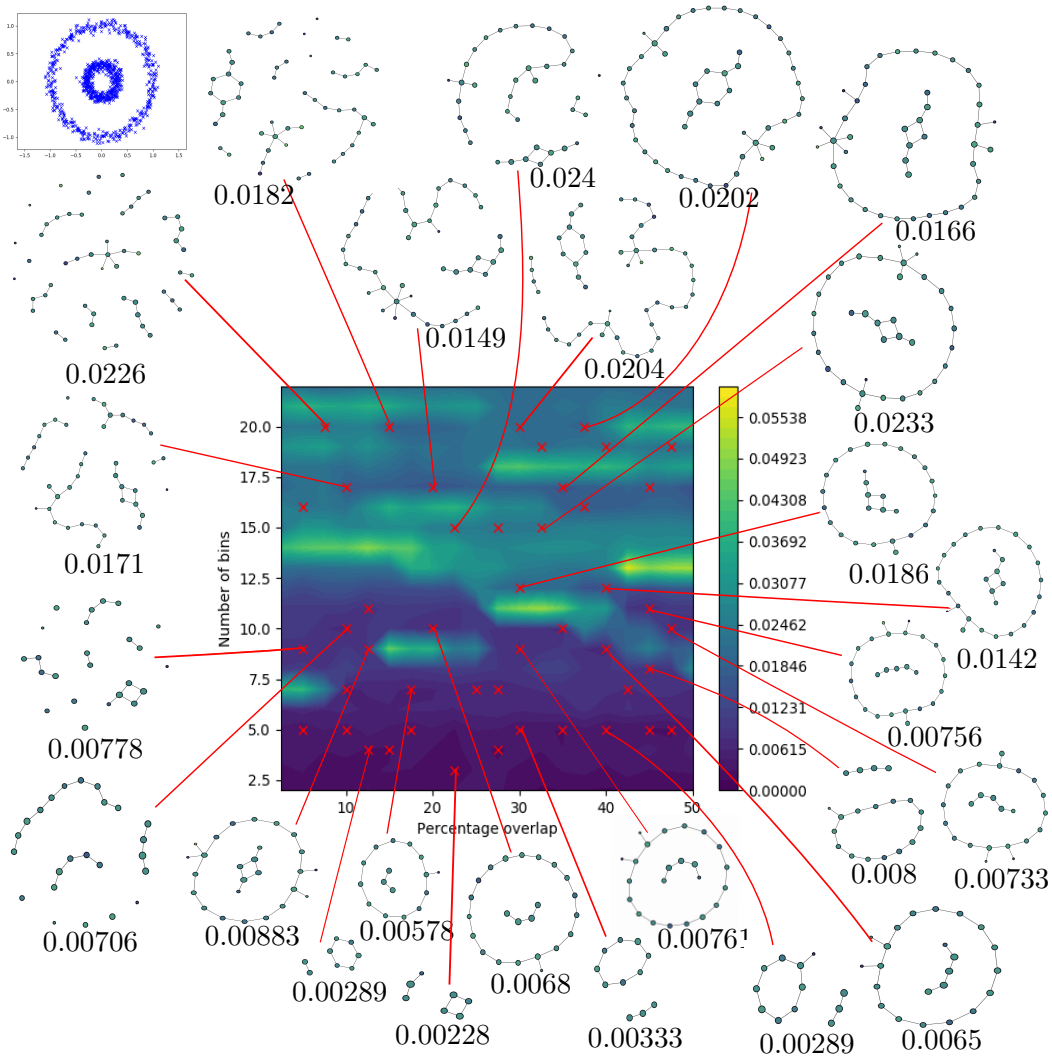


Figure 5: We consider 1000 points sampled with noise from two concentric circles. The centre of the figure shows a contour plot of instability values varying over the percentage overlap (gain) of the bins and the number of bins (resolution). The red crosses correspond to the local minima. The numbers next to the vertical bar on the right are values of instability. Surrounding the plot are the Mapper graphs corresponding the various local minima. Below each Mapper graph is the corresponding instability value. The bin overlap was between 2.5% and 50% at 2.5% interval steps. The number of bins varied between 2 to 22. The instabilities were averaged over 10 runs where we selected 10 random sub-samples in each case. See §7 for details of the procedure.

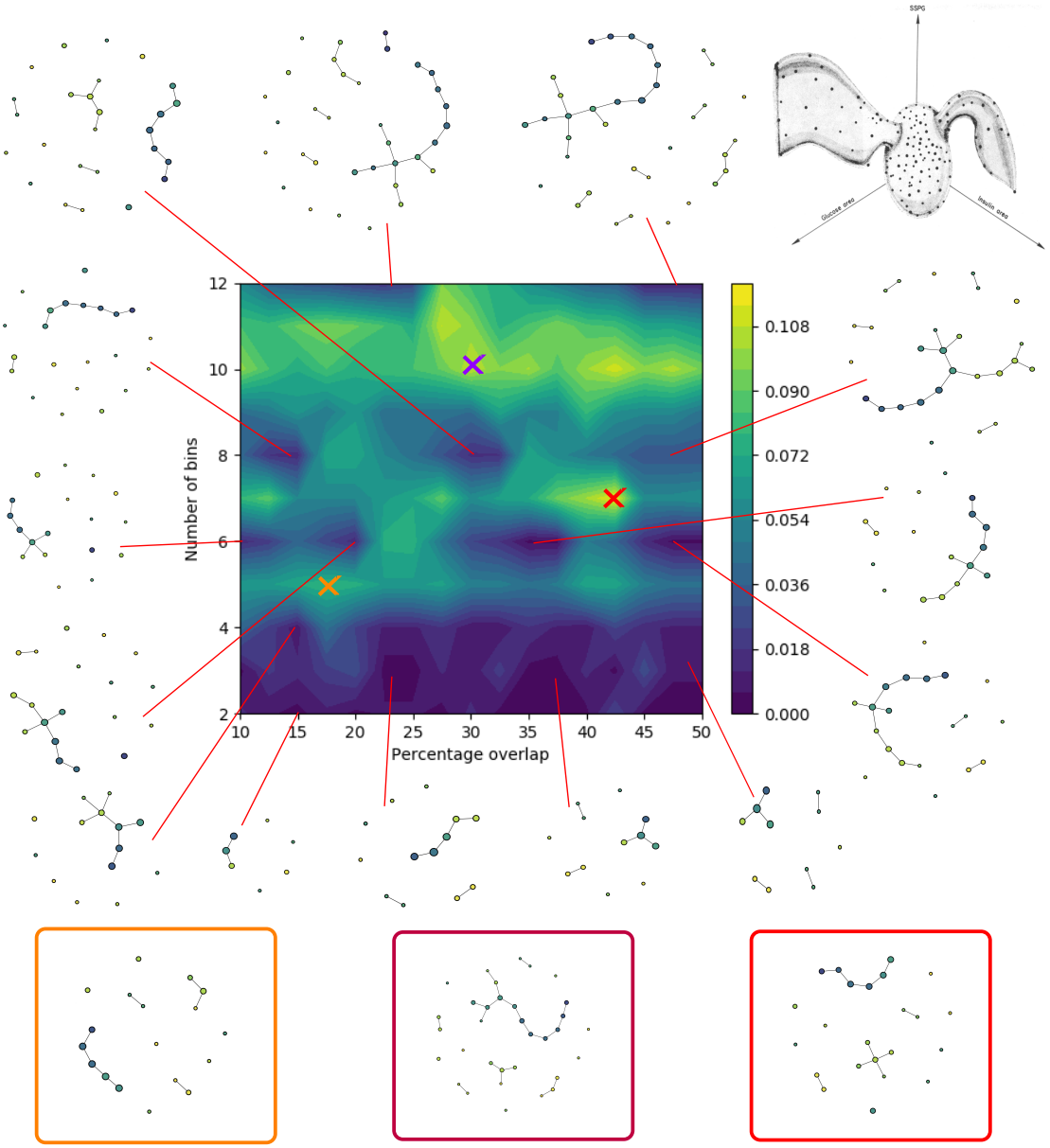


Table 4: An example of the diabetes study by Reaven and Miller (1979). The contour plot in the centre shows instability as a function of the number of the bins (resolution) and the percentage of their overlap (gain). We give Mapper graphs corresponding to the various local minima. The three Mapper graphs at the bottom correspond to the local maxima (crosses) differ significantly from the target configuation. We used the epsilon neighbourhood clustering $\epsilon = 0.3$. The bin overlap was between 10% and 50% at 2.5% interval steps; the number of bins varied between 2 and 12. The instabilities were averaged over 10 runs with 5 random sub-samples selected in each case. See §7 for more details.

High instability in the top left hand corner of the contour plot in Figure 5 and to a lesser extent most of the left hand side of the plot, appears to correspond to Mapper graphs with a fragmented outer circle. This feature can be explained by (c) of Remark 25, since the low percentage overlap between the bins is causing fragments of the outer circle to partially join together in an inconsistent fashion over varying subsamples.

Table 4 considers the diabetes data set first analysed in (Reaven and Miller, 1979) (available form the locfit R-package) and later studied using Mapper (Singh et al., 2007), in which the data is observed to form three distinct subgroups corresponding to distinct flares in the Mapper analysis.

The Mapper images in Table 4 at local minima usually give a good description of the underlying data particularly when the percentage overlap is high. In particular Mapper instability detects well the transitions between structure in this direction.

In contrast the ridges of high Mapper instability are most visible horizontally along particular bin values. On these ridges, the Mapper graphs corresponding to higher values of instability more often have missing flares and greater amounts of erroneous structure within the outliers when compared to the more stable Mapper graphs in a similar region of the parameter space.

We conclude that to infer the reliability of the Mapper graph, the Mapper instability should be considered over the whole parameter space. While it is intuitively clear that a more complicated Mapper output often gives a more unstable result we show that jumps in instability appear to correspond well with the structural changes in the Mapper output. A jump in complexity accompanied by a relatively low jump in instability, suggests that the additional structure is indeed present in the data, providing a method to determine the reliability of features present within relatively stable regions in the parameter space.

9. Conclusion

In this paper we have demonstrated that changes in the choice of particular parameters to create Mapper outputs can lead to very unstable results. To help alleviate this shortcoming, we have created a framework that can be used to select regions in the parameter space which are likely to create reliable Mapper outputs. We have introduced Mapper instability to provide a numerical measure of reliability of a particular Mapper output, especially when considered over a range of parameters. In particular our construction makes very few assumption on the specifics of the chosen Mapper construction, which makes it applicable to any Mapper-type algorithm.

We provide theoretical results to describe and explain the behaviour of the Mapper instability and in our discussion we make very few assumptions about the specifics of the structure of the data or the particular cover used to create Mapper outputs and show that in most circumstances the instability converges to zero as the sample size is increased. We construct explicit bounds which lead to practical criteria for Mapper instability. We provide a number of experimental results to further support the practical use our findings.

An important outcome of our discussion is that we are now able to verify when a change in the Mapper output is indeed supported by the structure of the data. Specifically, while more complicated Mapper outputs often suffer from a greater instability, we show that when

the increase in instability is accompanied by low instability, the resulting structure is indeed present in the data.

Acknowledgements: This research was supported by the EPSRC grant EP/N014189/1.

Appendix A: Justification of Measurability for Neighbourhood Clustering

In this Appendix we justify the assumption that $\mathcal{I}(Q_n)$ of Equation 6 is a random variable. In other words, $\mathcal{I}(Q_n)$ needs to be a measurable function with respect to the Borel σ-algebra on $\mathbb{R}$ and the product probability measure on $U^n \times U^n$. The measurability of $\mathcal{I}(Q_n)$ can be guaranteed provided the empirical quality function satisfies the condition of the following lemma.

Lemma 38 *Let $i\colon \mathcal{F}_n^U \times \mathcal{F}_n^U \to \mathcal{F}_{2n} \times \mathcal{F}_{2n}$ be the inclusion map given by the Voronoi cells in Equation 5. Then for each pair (f, g) of clustering functions on $2n$ points, if the pre-image of*

$$U^n \times U^n \to \mathcal{F}_{2n} \times \mathcal{F}_{2n}, \quad (X, X') \mapsto i(C_n(X), C_n(X')) \tag{26}$$

at (f, g) across $\mathcal{F}_{2n} \times \mathcal{F}_{2n}$ is measurable, then $\mathcal{I}(Q_n)$ is a random variable.

Proof Given that there are only finitely many clustering functions on $2n$ points, the map

$$D_m : \mathcal{F}_{2n} \times \mathcal{F}_{2n} \longrightarrow \mathbb{R}$$

determined by the matching metric is measurable. In consequence, by Equation 6, the map $\mathcal{I}(Q_n)$ is a random variable when the assumption of the Lemma holds. ■

The condition of the Lemma 38 is easily verified for common quality functions. For example in the case of nearest neighbour clusterings, given $\epsilon > 0$ and clusterings f, g on $2n$ points, we can describe the preimage in Equation 26 by a set of simple conditions. More precisely, the preimage is given by the set of points $((X_1, \ldots, X_n), (X'_1, \ldots, X'_n)) \in U^n \times U^n$ that satisfy the following. First, we define an ϵ-path in a metric space to be a sequence of points $(X_1, \ldots, X_k)$ such that $D(X_i, X_{i+1}) \leq \epsilon$ for $i = 1, \ldots, k-1$.

1. For every two points X_{i_α} and X_{i_β} chosen from $(X_1, \ldots, X_n)$, we have that $f(X_{i_\alpha}) = f(X_{i_\beta})$ if and only there is an ϵ-path consisting of points from the list $(X_1, \ldots, X_n)$ connecting X_{i_α} and X_{i_β}.

2. The function g satisfies an analogous condition on the sequence of points $(X'_1, \ldots, X'_n)$.

3. For every $i = 1, \ldots, n$, let j be the smallest index so that the element X_j from the list $(X_1, \ldots, X_n)$ minimises the distance $D(X'_i, X_k)$, for $k = 1, \ldots, n$. Then if $f(X_j) = C$ then also $f(X'_i) = C$.

4. An analogous condition holds for the clustering g.

Lemma 39 *For each $(f, g) \in \mathcal{F}_{2n} \times \mathcal{F}_{2n}$, the subsets of $(\mathbb{R}^a)^n \times (\mathbb{R}^a)^n$ described above are measurable.*

Proof Given $(f, g) \in \mathcal{F}_{2n} \times \mathcal{F}_{2n}$, consider in turn the restrictions imposed by each of the conditions (1), (2), (3) and (4) given above the lemma.

For (1), since all points sharing a label are connected by ϵ-paths and any two such points are connected by a path, we may consider adding these points inductively in the following way. When $n = 1$ there is a single point which can take any value in $\mathbb{R}^a$. In particular, $\mathbb{R}^a$ is a measurable set. Now assume inductively that for some $k = 1, \dots, n$, the possible values of the points $X_1, \dots, X_k$ under condition (1) form a measurable set $S_k \subseteq (\mathbb{R}^a)^k$. The corresponding set S_{k+1} on points $X_1, \dots, X_k, X_{k+1}$ is a subspace of $S_k \times \mathbb{R}^a$, under the condition that the final point X_{k+1} is at most a distance of ϵ from any of the points of $X_1, \dots, X_k$ with the same label and at least a distance greater than ϵ form any with a different label. More precisely X_{k+1} satisfies that, for each $j = 1, \dots, k$,

$$D(X_j, X_{k+1}) \leq \epsilon \text{ if } f(X_j) = f(X_{k+1}) \text{ and } D(X_j, X_{k+1}) \leq \epsilon \text{ if } f(X_j) \neq f(X_{k+1}).$$

Note that the possible values of X_{k+1} are nonempty. If X_{k+1} shares a label with one of $X_1, \dots, X_k$, then it may for example take the same value and if not the union of the epsilon neighbourhoods of points $X_1, \dots, X_k$ cannot cover all of $\mathbb{R}^a$. So the possible values of X_{k+1} are the nonempty intersection of a closed set determined by the first set of strict bounds and an open set determined by the second set of non-strict bounds. Since S_k is measurable, the above inequalities on X_{k+1} extend it to a measurable set S_{k+1}. Hence the possible values of $X_1, \dots, X_n$ under condition (1) lie in a measurable set $A = S_n$. Analogously we see that the set B of the possible values of $X'_1, \dots, X'_n$ under condition (2) is measurable.

For each $i = 1, \dots, n$, consider the subsets

$$X^i_\alpha = \{X_{\alpha_1}, \dots X_{\alpha_k}\} \subseteq \{X_1, \dots, X_n\},$$

such that $f(X_{\alpha_j}) = f(X'_i)$ for each $j = 1, \dots, k$. The Voronoi cells of $X_1, \dots, X_n$ are defined in equation 5. For each X_{α_j} its corresponding cell is obtained by a finite set of inequalities. Each inequality is strict if it arises from a pair of points $X^i_{\alpha_j}$ and X_p such that $p < \alpha_j$ and non-strict if $p > \alpha_j$. Condition (3) is equivalent to requiring X'_i is contained in the Voronoi cell of one of the elements of X^i_α. We may split the conditions on the Voronoi cells of X_α onto those with a strict inequality and those with an non-strict inequality. Using a similar inductive augment used when considering condition (1) in the previous part of the proof, we may now describe the possible values of $X'_1, \dots, X'_n$ under condition (3) as the intersection of an open and closed set, built from the strict and non-strict inequalities respectively to obtain a measurable set C. Similarly (4) gives us a measurable subset D of $(\mathbb{R}^a)^n$.

Putting this all together, the subset of points in $(\mathbb{R}^a)^n \times (\mathbb{R}^a)^n$ we wish to describe, is the intersection of the sets $A \times (\mathbb{R}^a)^n$, $(\mathbb{R}^a)^n \times B$, $C \times (\mathbb{R}^a)^n$ and $(\mathbb{R}^a)^n \times D$. Since each of A, B, C and D are measurable sets, the intersection is a measurable set. ■

References

M. Alagappan. From 5 to 13: Redefining the positions in basketball. *MIT Sloan Sports Analytics Conference*, 2012.

S. Ben-David. A framework for statistical clustering with constant time approximation algorithms for k-median and k-means clustering. *Machine Learning*, 66(2):243–257, Mar 2007. ISSN 1573-0565. doi: 10.1007/s10994-006-0587-3.

S. Ben-David and M. Ackerman. Measures of clustering quality: A working set of axioms for clustering. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 121–128. Curran Associates, Inc., 2009.

S. Ben-David and U. von Luxburg. Relating clustering stability to properties of cluster boundaries. In *COLT 2008*, pages 379–390, Madison, WI, USA, July 2008. Max-Planck-Gesellschaft, Omnipress.

S. Ben-David, U. von Luxburg, and D. Pál. A sober look at clustering stability. In G. Lugosi and H. U. Simon, editors, *Learning Theory: 19th Annual Conference on Learning Theory, COLT 2006, Pittsburgh, PA, USA, June 22-25, 2006. Proceedings*, pages 5–19. Springer Berlin Heidelberg, 2006. ISBN 978-3-540-35296-9. doi: 10.1007/11776420_4.

S. Ben-David, D. Pál, and H. U. Simon. Stability of k-means clustering. In N. H. Bshouty and C. Gentile, editors, *Learning Theory: 20th Annual Conference on Learning Theory, COLT 2007, San Diego, CA, USA; June 13-15, 2007. Proceedings*, pages 20–34. Springer Berlin Heidelberg, 2007. ISBN 978-3-540-72927-3. doi: 10.1007/978-3-540-72927-3_4.

A. Ben-Hur, A. Elisseeff, and I Guyon. A stability based method for discovering structure in clustered data. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, pages 6–17, 2002.

M. Bittner et al. Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature*, 406:536 EP –, 2000.

G. R. Bowman, X. Huang, Y. Yao, J. Sun, G. Carlsson, L. J. Guibas, and V. S. Pande. Structural insight into rna hairpin folding intermediates. *JACS Communications, pp*, pages 9676–9678, 2008.

L. Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, Aug 1996. ISSN 1573-0565. doi: 10.1023/A:1018054314350.

L. Breiman. Arcing classifier (with discussion and a rejoinder by the author). *Ann. Statist.*, 26(3):801–849, June 1998. doi: 10.1214/aos/1024691079.

M. Burfitt. Mapper instability. https://github.com/Quiff789/Mapper-instability, June 2019.

P. G. Camara. Topological methods for genomics: Present and future direction. *Current Opinion in Systems Biology*, 1:95–101, 2017.

A. Caponnetto, A. Rakhlin, and P. Kaelbling. Stability properties of empirical risk minimization over donsker classes. volume 6, pages 2565–2583, Dec 2006.

G. Carlsson. Topology and data. *Bull. Amer. Math. Soc. (N.S.)*, 46(2):255–308, 2009. ISSN 0273-0979. doi: 10.1090/S0273-0979-09-01249-X.

G Carlsson. The shape of biomedical data. *Current Opinion in Systems Biology*, 1, 2017.

G. Carlsson and F. Mémoli. Characterization, stability and convergence of hierarchical clustering methods. *Journal of Machine Learning Research*, 11:1425–1470, 04 2010.

M. Carrière and S. Oudot. Structure and Stability of the 1-Dimensional Mapper. In Sándor Fekete and Anna Lubiw, editors, *32nd International Symposium on Computational Geometry (SoCG 2016)*, volume 51 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 25:1–25:16, Dagstuhl, Germany, 2016. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik. ISBN 978-3-95977-009-5. doi: 10.4230/LIPIcs.SoCG.2016.25.

M. Carrière, B. Michel, and S. Oudot. Statistical analysis and parameter selection for mapper. *Journal of Machine Learning Research*, 19(12):1–39, 2018.

L. D. Cecco et al. Head and neck cancer subtypes with biological and clinical relevance: Meta-analysis of gene-expression data. *Oncotarget*, 6:9627–9642, 2015.

J. M. Chan, G. Carlsson, and R. Rabadana. Topology of viral evolution. *Proceedings of the National Academy of Science*, 110, 2013.

J. Chang, M. M. Nicolau, T. R. Cox, D. Wetterskog, J. W. Martens, H. E. Barker, and J. T. Erler. Loxll2 induces aberrant acinar morphogenesis via erbb2 signaling. *Breast Cancer Research*, 15, 2013.

K. T. Dey, F. Mémoli, and Y. Wang. Multiscale mapper: Topological summarization via codomain covers. In *SODA*, pages 997–1013. SIAM, 2016.

K. T. Dey, F. Mémoli, and Y. Wang. Topological analysis of nerves, reeb spaces, mappers, and multiscale mappers. In *Symposium on Computational Geometry*, volume 77 of *LIPIcs*, pages 36:1–36:16. Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik, 2017.

P. Dłotko. Ball mapper: a shape summary for topological data analysis. *arXiv e-prints*, January 2019.

L. Duponchel. Exploring hyperspectral imaging data sets with topological data analysis. *Analytica Chimica Acta*, 1000:123–131, 2018a.

L. Duponchel. When remote sensing meets topological data analysis. *Journal of Spectral Imaging*, 2018b.

J. Hendrik and S. Nathaniel. Keplermapper. http://doi.org/10.5281/zenodo.1054444, nov 2017.

T. S. C. Hinks et al. Innate and adaptive t cells in asthmatic patients: Relationship to severity and disease mechanisms. *Journal of Allergy and Clinical Immunology*, 136(2): 323–333, 2015.

T. S. C. Hinks et al. Multidimensional endotyping in patients with severe asthma reveals inflammatory heterogeneity in matrix metalloproteinases and chitinase 3-like protein 1. *J.Allergy Clin Immunol*, 138(1), 2016.

R. Jeitziner, M. Carriére, J. Rougemont, S. Oudot, K. Hess, and C. Brisken. Two-Tier Mapper: a user-independent clustering method for global gene expression analysis based on topology. *arXiv e-prints*, December 2017.

M. Kamruzzaman, A. Kalyanaraman, B. Krishnamoorthy, and P. Schnable. Toward a scalable exploratory framework for complex high-dimensional phenomics data. *arXiv e-prints*, 2017.

J. M. Kleinberg. An impossibility theorem for clustering. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information Processing Systems 15*, pages 463–470. MIT Press, 2003.

Y. Lee, S. D. Barthel, P. Dłotko, S. M. Moosavi, K. Hess, and B. Smit. Quantifying similarity of pore-geometry in nanoporous materials. *Nature Communications*, 8:15396, May 2017. doi: 10.1038/ncomms15396.

E. Levine and E. Domany. Resampling method for unsupervised estimation of cluster validity. *Neural Computation*, 13(11):2573–2593, 2001. doi: 10.1162/089976601753196030.

L. Li, W. Cheng, B. S. Glicksberg, O. Gottesman, R. Tamler, R. Chen, E. P. Bottinger, and J. T. Dudley. Identification of type 2 diabetes sub-groups through topological analysis of patient similarity. *Science Translational Medicine*, 7(311), 2015.

P. Y. Lum et al. Extracting insights from the shape of complex data using topology. *Scientific Reports*, 3(1236), 2013.

M. Meilă. Comparing clusterings: An axiomatic view. In *Proceedings of the 22Nd International Conference on Machine Learning*, ICML '05, pages 577–584, New York, NY, USA, 2005. ACM. ISBN 1-59593-180-5. doi: 10.1145/1102351.1102424.

N. Monica, A. J. Levine, and G. Carlsson. Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival. *Proceedings of the National Academy of Sciences of the United States of America*, 108(17):7265–7270, 2011.

D. Müllner and A. Babu. Python mapper: An open-source toolchain for data exploration, analysis and visualization. http://danifold.net/mapper, 2013.

D. Müllner, P. Pearson, and G. Singh. TDA mapper. https://github.com/paultpearson/TDAmapper, May 2010.

J. L. Nielson et al. Topological data analysis for discovery in preclinical spinal cord injury and traumatic brain injury. *Nature Communications*, 6:8581+, 2015.

R. Ostrovsky, Y. Rabani, L. Schulman, and C. Swamy. The effectiveness of lloyd-type methods for the k-means problem. In *2006 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS'06)*. IEEE, 2006. doi: 10.1109/focs.2006.75.

M. Pirashvili, L. Steinberg, F. Belchí Guillamon, M. Niranjan, J. G. Frey, and J. Brodzki. Improved understanding of aqueous solubility modeling through topological data analysis. *Journal of Cheminformatics*, 10(1):54, Nov 2018. ISSN 1758-2946. doi: 10.1186/s13321-018-0308-5.

G. M. Reaven and R. G. Miller. An attempt to define the nature of chemical diabetes using a multidimensional analysis. *Diabetologia*, 16(1):17–24, Jan 1979. ISSN 1432-0428. doi: 10.1007/BF00423145.

M. T. Ribeiro, S. Singh, and C. Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 1135–1144. ACM, 2016. ISBN 978-1-4503-4232-2. doi: 10.1145/2939672.2939778.

H. Riihimaki, W. Chacholski, J. Theorell, J. Hillert, and R. Ramanujam. A topological data analysis based classification method for multiple measurement. *arXiv e-prints*, Mar 2019. doi: 10.1101/569210.

A. H. Rizvi, P. G. Camara, E. K. Kandror, T. J. Roberts, I. Schieren, T. Maniatis, and R. Rabadan. Single-cell topological rna-seq analysis reveals insights into cellular differentiation and development. *Nature Biotechnology*, 35, 2017.

D. Romano et al. Topological methods reveal high and low functioning neuro-phenotypes within fragile x syndrome. *Human Brain Mapping*, 35:4904–4915, 2014.

M. Rucco, E. Merelli, D. Herman, D. Ramanan, T. Petrossian, L. Falsetti, C. Nitti, and A. Salvi. Using topological data analysis for diagnosis pulmonary embolism. *Journal of Theoretical and Applied Computer Science*, 9:41–55, 2015.

G. Sarikonda et al. Cd8 t-cell reactivity to islet antigens is unique to type 1 while cd4 t-cell reactivity exists in both type 1 and type 2 diabetes. *Journal of Autoimmunity*, 50(Supplement C):77–82, 2014.

A. Savir, G. Toth, and L. Duponchel. Topological data analysis (tda) applied to reveal pedoge- netic principles of european topsoil system. *Science of the Total Environment*, 586(2):1091–1100, 2017.

J. P. R. Schofield et al. Stratification of asthma phenotypes by airway proteomic signatures. *Journal of Allergy and Clinical Immunology*, 2019. ISSN 0091-6749. doi: https://doi.org/10.1016/j.jaci.2019.03.013.

O. Shamir and N. Tishby. Stability and model selection in k-means clustering. *Machine Learning*, 80(2):213–243, Sep 2010. ISSN 1573-0565. doi: 10.1007/s10994-010-5177-8.

G. Singh, F. Mémoli, and G. Carlsson. Topological methods for the analysis of high dimensional data sets and 3D object recognition. In *SPBG*. The Eurographics Association, 2007.

F. Strazzeri and R. J. Sánchez-García. Morse theory and an impossibility theorem for graph clustering. *CoRR*, abs/1806.06142, 2018.

B. Y. Torres, J. H. O. Oliveira, A. T. Tate, R. Poonam, K. Cumnock, and D. S. Schneider. Tracking resilience to infections by mapping disease space. *PLOS Biology*, 14(6):e1002436, 2016.

U. von Luxburg. Clustering stability: An overview. *Found. Trends Mach. Learn.*, 2(3): 235–274, March 2010. ISSN 1935-8237. doi: 10.1561/2200000008.

U. von Luxburg, S. Bubeck, S. Jegelka, and M. Kaufmann. Consistent minimization of clustering objective functions. In *Advances in Neural Information Processing Systems 20*, pages 961–968, Red Hook, NY, USA, September 2008. Max-Planck-Gesellschaft, Curran.