Multiparameter Hierarchical Clustering Methods

Gunnar Carlsson and Facundo Mémoli

Stanford University, Mathematics Department.
{gunnar,memoli}@math.stanford.edu

Summary. We propose an extension of hierarchical clustering methods, called *multiparameter hierarchical clustering methods* which are designed to exhibit sensitivity to density while retaining desirable theoretical properties. The input of the method we propose is a triple (X, d, f), where (X, d) is a finite metric space and $f : X \to \mathbb{R}$ is a function defined on the data X, which could be a density estimate or could represent some other type of information. The output of our method is more general than dendrograms in that we track two parameters: the usual scale parameter and a parameter related to the function f. Our construction is motivated by the methods of *persistent topology* [6], the Reeb graph and Cluster Trees [16]. We present both a characterization, and a stability theorem.

Key words: Hierarchical clustering, single linkage, persistent topology.

1 Introduction

Clustering techniques play a very central role in various parts of data analysis. They can give important clues to the structure of datasets, and therefore suggest results and hypotheses in the underlying science. However, despite being one of the most commonly used tools for unsupervised exploratory data analisys, and despite its extensive literature, very little is known about the theoretical foundations of clustering methods. These points have been made prominent by Ben-David and von Luxburg in [1].

The general question of which methods are "best", or most appropriate for a particular problem, or how significant a particular clustering is has not been addressed very frequently. In the context of standard clustering (standard clustering refers to clustering methods that output a single partition of a dataset and hierarchical methods that yield a nested family of partitions), J. Kleinberg proves in [11] a very interesting impossibility result for the problem of even defining a clustering scheme with some rather mild invariance properties.

Inspired by Kleinberg's axiomatic treatment, in [4] we wondered whether in the context of hierarchical clustering (HC from now on) methods, one would

be able to lift the obstruction to existence in his result. Interestingly, we were able to prove that for HC methods, conditions similar to Kleinberg's yield uniqueness instead of non-existence. This HC scheme singled out by our theorem satisfies precise *stability* and *convergence* properties [4]. This unique scheme turned out to be *single linkage* HC. There seems to exist an agreement that amongst hierarchical methods, SL is the one with best theoretical properties, see also the results of Jardine and Sibson in this respect [10].

However, single linkage has frequently been severely criticized for the *chaining effect* it exhibits (see [13] and [17, pp. 296]): SL will disregard the density of samples in a region and may tend to connect two dense clusters when just a few isolated samples produce a chain connecting them. This has had the effect that in practice other clustering methods are typically preferred over SL. Practicioners tend to favour average (AL) or complete (SL) linkage, which are deemed more sensitive to variations of density in datasets. However, since AL and CL are actually *unstable* [10, Section 7.4] in a precise sense, there is a blatant inconsistency between the conclusions of theoretical studies and practical applications of clustering algorithms.

Clustering can be regarded as a statistical problem if we consider the dataset $\mathbb{X} = \{x_1, \ldots, x_n\} \subset X$ as a sample from some unknown probability measure μ_X defined on the Borel sets of a metric space (X, d_X). Consider for the sake of simplicity that X is Euclidean space $\mathbb{R}^d$ and that μ_X is a measure with density ρ. The two main statistical approaches to clustering are the *parametric approach* and the *nonparametric approach*. The former approach is based on the assumption that each group i is represented by a density ρ_i that is a member of some parametric family. The density ρ is then a mixture of the group densities, and the number of components in the mixture together with the parameters values are estimated from the data. The latter approach assumes that groups correspond to *modes* of the density ρ. Searching for modes as a manifestation of the presence of groups can be traced back to D. Wisharts paper [17].

With regards to the chaining effect: it is well understood that one of the shortcomings of SL is its insensitivity to *density*. In this direction, a classical result of Hartigan [8] proves that SL is not *consistent* in the sense that it is unable to recover modes of an underlying density in $\mathbb{R}^d$ for all d. In [17] Wishart proposes *one level mode analysis* as an obvious approach to the amelioration of the chaining effect. The idea is to remove from the observational data all the points that appear to be noise. Define the superlevel set $L_\rho(\sigma)$ of a density ρ at level σ as the subset of the underlying space X for which the density exceeds σ: $L_\rho(\sigma) = \{x | \rho(x) > \sigma\}$. Then, if $\widehat{\rho}$ is some estimate of ρ and σ a given threshold, the idea consists of applying SL clustering to $L_{\widehat{\rho}}(\sigma)$.

In [9, Section 11] and [8], Hartigan expanded on Wishart's idea and made it more precise: he defined the *high density clusters* at level σ as the connected components of $L_\rho(\sigma)$. Hartigan also pointed out that the collection of high density clusters has a *hierarchical structure*: for any two clusters A and B

(possibly at different levels) either $A \subset B$ or $B \subset A$ or $A \cap B = \varnothing$. This hierarchical structure is summarized by the **cluster tree** of ρ.

More recent instantiations of the one level mode analysis idea can be found in [7, 5, 2]. Tipycally, methods roughly consist of four steps: (1) for each data point calculate a density estimate $\widehat{\rho}$; (2) choose a density threshold σ and construct $L_{\widehat{\rho}}(\sigma)$; (3) construct a graph interconnecting all observations in $L_{\widehat{\rho}}(\sigma)$ within distance ε of each other; (4) define the clusters to be the connected components of this graph.

As was pointed out in [15], a well known weakness of the one level mode analysis is that the degree of separation between connected components of $L_\rho(\sigma)$, and therefore of $L_{\widehat{\rho}}(\sigma)$, depends critically on the choice of the density threshold σ, which is left to the user. Moreover, there might not be a single value of σ that uncovers all the modes. In [17], citing this difficulty Wishart proposed *hierarchical mode analysis*, which can be regarded as a procedure for computing the cluster tree of a density estimate $\widehat{\rho}$. The work of Wong and Lane [18] provides a method of estimating the cluster tree of a density by a construction based on k-nearest neighbor density estimates.

In [16] Stuetzle gives a precise recursive definition of the cluster tree. Stuetzle's method estimates the cluster tree of the density by computing the cluster tree of the nearest neighbor density estimate and then pruning branches believed to correspond to spurious modes. In [15] the authors present a generalization of Stuetzle's method to other density estimates. It is already expressed in the work of Stuetzle and Nugent that it is desirable to prove that the cluster tree estimates one constructs are *stable* to perturbations in the data. Furthermore, the issue of convergence of the sample based cluster tree has to be resolved, see the discussion in [18]. Similar ideas are also present in the work of Klemelä [12].

The construction implicit in many of the methods we mentioned can be paraphrased as follows. Assume (X, d_X, f) is given where (X, d_X) is a finite metric space and $f : X \to \mathbb{R}$ is a given function (which could be a density estimate). For each σ let $X^\sigma := L_f(\sigma)$. For a given $\varepsilon > 0$ consider the graph $G_{\varepsilon,\sigma} = (X^\sigma, E_{\varepsilon,\sigma})$ with $E_{\varepsilon,\sigma} = \{(x, x') \in X^\sigma \times X^\sigma \,|\, d_X(x, x') \leqslant \varepsilon,\, i \neq j\}$. Then, obtain a one-mode-analysis type of summary by computing the connected components of $G_{\varepsilon,\sigma}$. Clearly, this set up can be used for estimating the cluster tree of f as well by following a recursive procedure such as the one delineated by Stuetzle.

The proposal in this paper hinges on the idea that there is more information contained in the whole collection of graphs $\{G_{\varepsilon,\sigma}\}_{\varepsilon \geqslant 0, \sigma \geqslant 0}$ than just an estimate or a family of estimates (one for each ε) of the cluster tree. Much in the same way as single mode analysis suffers from a particular choice of the density threshold, a procedure that tries to estimate the cluster tree from $\{G_{\varepsilon_0,\sigma}\}_{\sigma \geqslant 0}$ for a *fixed* ε_0 will be affected by having made fixed choice for the spatial (metric dependent) scale ε_0. We claim that it may in fact be more informative to encode all possible choices of scale into an *invariant* richer than just a single cluster tree. The invariant we construct out of the family

$\{G_{\varepsilon,\sigma}\}_{\varepsilon\geqslant 0,\sigma\geqslant 0}$ can be regarded as a generalization of both hierarchical clustering and the cluster tree. In fact, a *slice* of the invariant for a fixed value of ε yields a cluster tree estimate, whereas a slice for a fixed value of σ yields the dendrogram corresponding to applying HC to X^σ, i.e. a single mode analysis snapshot. Our construction therefore takes into account both the linkage parameter ε and σ, a parameter related to the function f (e.g. density). This is to be regarded as *multiparameter clustering*.

In this paper, we produce a variation of the theme in [4]. By first identifying desirable properties of such multi-parameter clustering procedures, we then propose a set of axioms for such methods. We prove a *uniqueness/characterization theorem* (Theorem 1) under these axioms. The procedure singled out by this set of axioms can be regarded as a generalization of both SL HC and the *cluster tree* construction. In addition, in Theorem 2 we establish the precise quantitative (or metric) *stability* of the particular clustering scheme which is characterized by our results.

Our presentation is necessarily concise given the space constraints; more details and elaboration will be presented in a future publication.

2 Notation and Terminology

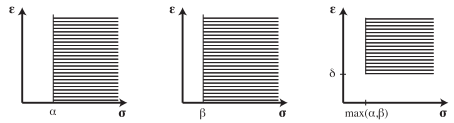
Let $\mathcal{X}$ denote the collection of all finite metric spaces. Let $\mathcal{X}_1$ be the collection of all finite *filtered metric spaces*, that is triples (X, d_X, f_X) where $(X, d_X) \in \mathcal{X}$ and $f_X : X \to \mathbb{R}$. Given $(X, d_X, f_X) \in \mathcal{X}_1$, for each $\sigma \in \mathbb{R}$ let $X_\sigma = f_X^{-1}((-\infty, \sigma])$. For a finite set X and a symmetric function $W : X \times X \to \mathbb{R}^+$ let $\mathcal{L}(W)$ denote the maximal metric on X less than or equal to W, i.e. $\mathcal{L}(W)(x, x') = \min\{\sum_{i=0}^m W(x_i, x_{i+1}) |\, x = x_0, \ldots, x_m = x', m \in \mathbb{N}\}$ for $x, x' \in X$. For a finite metric space (X, d_X), $\mathrm{sep}(X, d_X)$ denotes the minimal distance between any two different points in X. When referring to a metric space (X, d_X) or to a filtered metric space (X, d_X, f_X) we may drop the metric and filter and refer to it by just X. For a topological space S, $\mathcal{B}(S)$ denotes the collection of Borel sets of S. Given a set Z, for a function $h : Z \to \mathbb{R}$, we use the notation $\|h\|_{L_\infty(Z)} = \sup_{z\in Z} |h(z)|$.

3 Two Parameter Hierarchical Clustering: a Characterization Theorem

Definition 1 (Persistent Structures). *Given a finite set X, a* persistent structure *on X is a map $Q_X : X \times X \to \mathcal{B}(\mathbb{R}^+ \times \mathbb{R})$ s.t.*

1. *If $(\varepsilon, \sigma) \in Q_X(x, x')$, then $(\varepsilon + t, \sigma + s) \in Q_X(x, x')$ for all $t, s \geqslant 0$.*
2. *If $(\varepsilon_1, \sigma_1) \in Q_X(x, x')$ and $(\varepsilon_2, \sigma_2) \in Q_X(x', x'')$, then $\big(\max(\varepsilon_1, \varepsilon_2), \max(\sigma_1, \sigma_2)\big) \in Q_X(x, x'')$.*
3. *For all $x, x' \in X$, $\partial Q_X(x, x') \subset Q_X(x, x')$ (technical condition).*

Example 1. Let $\Delta = \{p, q\}$ and Q_Δ be given by the Figure 1, where $\alpha, \beta, \delta \geqslant 0$.



$Q_\Delta(p,p)$ $Q_\Delta(q,q)$ $Q_\Delta(p,q)$

Fig. 1. A simple persistence structure on $\{p,q\}$: Q_Δ.

Remark 1. Persistent structures are useful constructs for expressing nested associations of points. They can be regarded as a certain generalization of the concept of ultrametrics and therefore of dendrograms (nested families of partitions). In fact, one can see that a persistence structure Q on X gives rise to a family of ultrametrics on X.

We use the language of *categories* and *functors*, see [4] for an exposition relevant to clustering and [14] for a comprehensive account. Below, $\underline{Sets}$ denotes the category whose objects are sets and whose morphisms are set maps.

Consider the category $\underline{\mathcal{Q}}$ whose objects are pairs (X, Q_X) where X is a finite set and Q_X is a persistent structure on X. Let $\mathcal{Q}$ denote the objects in $\underline{\mathcal{Q}}$. A map $\phi : X \to Y$ is called *persistence preserving* if for all $x, x' \in X$, $Q_X(x,x') \subseteq Q_Y(\phi(x),\phi(x'))$. We declare that $Mor_{\underline{\mathcal{Q}}}((X,Q_X),(Y,Q_Y))$ consists of all persistence preserving maps between X and Y. We define $\underline{\mathcal{M}}^{gen}$ to be the category that has all finite filtered metric spaces as objects, and as morphisms all those maps that are distance non-increasing and filter non-increasing. That is, $\phi \in Mor_{\underline{\mathcal{M}}^{gen}}(X,Y)$ if and only if for all $x, x' \in X$, $d_X(x,x') \geqslant d_Y(\phi(x),\phi(x'))$ and $f_X(x) \geqslant f_Y(\phi(x'))$.

In this context, a **clustering functor** will be a functor $\mathcal{C} : \underline{\mathcal{M}}^{gen} \to \underline{\mathcal{Q}}$. Consider the equivalence relation on X_σ given by $x \sim_{(\varepsilon,\sigma)} x'$ if and only if there exists $x_0, \ldots, x_m$ in X s.t. $x_0 = x$, $x_m = x'$, $\max_i d_X(x_i, x_{i+1}) \leqslant \varepsilon$ and $\max_i f_X(x_i) \leqslant \sigma$. For $x \in X_\sigma$ let $[x]_{(\varepsilon,\sigma)}$ denote the equivalence class to which x belongs.

Example 2. Consider the functor $\mathcal{C}^* : \underline{\mathcal{M}}^{gen} \to \underline{\mathcal{Q}}$ that when applied to (X, d_X, f_X) produces the object (persistent structure) (X, Q^*_X) where $Q^*_X(x,x') := \left\{(\varepsilon,\sigma) \in \mathbb{R}^2 \middle| \, x \sim_{(\varepsilon,\sigma)} x'\right\}$. That $\mathcal{C}^*$ is a functor follows easily from the definitions. The following observations are in order:

- The sets $Q^*_X(x,x')$ are obviously unbounded. They are of the form $\bigcup_{i=1}^K [\varepsilon^{(i)}, \infty) \times [\sigma_1^{(i)}, \infty)$. Note that for $x \in X$, $Q^*_X(x,x) = \{(\varepsilon,\sigma) \in \mathbb{R}^2 | \, \varepsilon \geqslant 0, \, \sigma \geqslant f_X(x)\}$.
- Let $B = [x]_{(\varepsilon,\sigma)} \neq [x']_{(\varepsilon,\sigma)} = B'$. Then, clearly, $\min_{x \in B, x' \in B'} d_X(x,x') > \varepsilon$.
- If (ε,σ) are s.t. $\mathrm{sep}(X_\sigma, d_X) > \varepsilon$, then $(\varepsilon,\sigma) \notin Q^*_X(x,x')$ for all x, x' in X_σ with $x \neq x'$. Indeed, otherwise let $x, x', x_0, \ldots, x_n \in X$ be s.t. $x_0 = x$, $x_n = x'$, $d_X(x_i, x_{i+1}) \leqslant \varepsilon$ and $f_X(x_i) \leqslant \sigma$. Since $x_i \in X_\sigma$ for $i \in \{0, \ldots, n\}$, and $x \neq x'$, there are at least two different consecutive points in $\{x_0, x_1, \ldots, x_n\}$ whose distance is not greater than ε, a contradiction.

- For $t \geqslant 0$ let $\sigma_X^t : X \times X \to \mathbb{R}$ be defined by $(x, x') \mapsto \inf \left\{ \sigma \,|\, x \sim_{(t,\sigma)} x' \right\}$. This gives rise to a tree and can be likened to the cluster tree construction of Stuetzle.

We have the following characterization/uniqueness theorem.

Theorem 1. *Let $\mathcal{C} : \underline{\mathcal{M}}^{gen} \to \underline{\mathcal{Q}}$ be a functor which satisfies the following conditions.*

(I) Let $\alpha : \underline{\mathcal{M}}^{gen} \to \underline{Sets}$ and $\beta : \underline{\mathcal{Q}} \to \underline{Sets}$ be the forgetful functors $(X, d_X, f_X) \to X$ and $(X, Q_X) \to X$, which forget the metric and filter, and persistence structure, respectively, and only "remember" the underlying sets X. Then we assume that $\beta \circ \Psi = \alpha$. This means that the underlying set of the persistent structure associated to a metric space is just the underlying set of the metric space.

(II) For $\delta \geqslant 0$ and $\alpha, \beta \in \mathbb{R}$ let $\Delta(\delta, \alpha, \beta) = (\{p, q\}, \left(\begin{smallmatrix} 0 & \delta \\ \delta & 0 \end{smallmatrix}\right), \{\alpha, \beta\})$ denote the two point filtered metric space with underlying set $\{p, q\}$, where $dist(p, q) = \delta$ and $f_\Delta(p) = \alpha$ and $f_\Delta(q) = \beta$. Then $\mathcal{C}(\Delta(\delta, \alpha, \beta))$ is the persistent structure $(\{p, q\}, Q_\Delta)$ whose underlying set is $\{p, q\}$ and where Q_Δ is given by the construction shown in Figure 1.

(III) Given $(\varepsilon, \sigma) \in \mathbb{R}^+ \times \mathbb{R}$ and finite filtered metric space (X, d_X, f_X), then $sep(X_\sigma) > \varepsilon$ implies that $(\varepsilon, \sigma) \notin Q_X(x, x')$ for any $x, x' \in X_\sigma$, $x \neq x'$.

Then $\mathcal{C}$ is equal to the functor $\mathcal{C}^*$.

Proof. We sketch the proof. Let (X, d_X, f_X) be a finite filtered metric space. Write $(X, Q_X) = \mathcal{C}(X, d_X, f_X)$. Also, write $(X, Q_X^*) = \mathcal{C}^*(X, d_X, f_X)$.
(1) Let $x, x' \in X$ and $(\varepsilon, \sigma) \in \mathbb{R}^+ \times \mathbb{R}$ be s.t. $(\varepsilon, \sigma) \in Q_X(x, x')$. We will prove that $(\varepsilon, \sigma) \in Q_X^*(x, x')$ as well. Consider the filtered metric space (X', d', f') where $X' = X \backslash \sim_{(\varepsilon,\sigma)}$. Let $\phi : X \to X'$ be given by $x \mapsto [x]_{(\varepsilon,\sigma)}$. For $\alpha, \beta \in X'$ let $W(\alpha, \beta) := \min_{x \in \phi^{-1}(\alpha),\, x' \in \phi^{-1}(\beta)} d_X(x, x')$. Note that by the discussion in Example 2, $\min_{\alpha \neq \beta} W(\alpha, \beta) > \varepsilon$ for $\alpha, \beta \in X'$. Define d' to be the maximal metric pointwisely less than or equal W, i.e. $d' = \mathcal{L}(W)$. Finally , let $f' : X' \to \mathbb{R}$ be given by $\alpha \mapsto \min_{x \in \phi^{-1}(\alpha)} f_X(x)$. Note that by construction, $X'_\sigma = X'$ and $\mathrm{sep}(X', d') > \varepsilon$.

Now, also by construction it holds that $\phi \in Mor_{\underline{\mathcal{M}}^{gen}}(X, X')$. By functoriality we then have $Q_X \subseteq Q_{X'} \circ (\phi, \phi)$, and in particular, we have that $(\varepsilon, \sigma) \in Q_{X'}(\phi(x), \phi(x'))$. Note that we must have $\phi(x) = \phi(x')$ for otherwise, condition (III) together with $\mathrm{sep}(X', d_{X'}) > \varepsilon$ give a contradiction. This means that $[x]_{(\varepsilon,\sigma)} = [x']_{(\varepsilon,\sigma)}$, hence, by definition of $\mathcal{C}^*$, $(\varepsilon, \sigma) \in Q_X^*(x, x')$.
(2) Let $x, x' \in X$ and $(\varepsilon, \sigma) \in \mathbb{R}^+ \times \mathbb{R}$ be s.t. $(\varepsilon, \sigma) \in Q_X^*(x, x')$. Let $x = x_0, x_1, \ldots, x_t = x'$ be points in X_σ s.t. $\max_i d_X(x_i, x_{i+1}) \leqslant \varepsilon$. Fix $i \in \{0, 1, \ldots, t-1\}$. Consider the two point filtered metric space $\Delta(\varepsilon, \sigma, \sigma)$ and the map $\psi : \Delta \to X$ given by $\psi(p) = x_i$ and $\psi(q) = x_{i+1}$. Note that by construction $\psi \in Mor_{\underline{\mathcal{M}}^{gen}}(\Delta, X)$. Then, $Q_\Delta \subseteq Q_X \circ (\psi, \psi)$, and in particular (check Figure 1), $(\varepsilon, \sigma) \in Q_X(x_i, x_{i+1})$. Since i was arbitrary, by applying property 2. in the Definition 1 repeatedly, we obtain that $(\varepsilon, \sigma) \in Q_X(x, x')$. This concludes the proof.

Example 3. As a simple practical tool for the analysis of data one could use the following construction: for a given triple (X, d, f) let $K_X : \mathbb{R}^+ \times \mathbb{R} \to \mathbb{N}$ be given by $(\varepsilon, \sigma) \mapsto \#\big(X_\sigma \backslash \sim_{(\varepsilon,\sigma)}\big)$, i.e. the number of equivalence classes of X_σ under $\sim_{(\varepsilon,\sigma)}$.

4 Metric Stability of $\mathcal{C}^*$

For sets A and B, a subset $R \subset A \times B$ is a *correspondence* (between A and B) if and and only if (1) $\forall\ a \in A$, there exists $b \in B$ s.t. $(a, b) \in R$; and (2) $\forall\ b \in B$, there exists $a \in X$ s.t. $(a, b) \in R$. Let $\mathcal{R}(A, B)$ denote the set of all possible correspondences between sets A and B.

Consider compact metric spaces (X, d_X) and (Y, d_Y). Let $\Gamma_{X,Y} : X \times Y \times X \times Y \to \mathbb{R}^+$ be given by $(x, y, x', y') \mapsto |d_X(x, x') - d_Y(y, y')|$. Then, the **Gromov-Hausdorff distance** [3] between X and Y is given by $d_{\mathcal{GH}}(X, Y) := \inf_{R\in\mathcal{R}(X,Y)} \|\Gamma_{X,Y}\|_{L^\infty(R\times R)}$. The Gromov-Hausdorff distance is a metric on the collection of all isometry classes of compact metric spaces [3]. We modify the expression of the Gromov-Hausdorff distance in order to define a metric for filtered metric spaces. We deem two spaces X, Y in $\mathcal{X}_1$ *isomorphic* whenever there exists an isometry $\Psi : (X, d_X) \to (Y, d_Y)$ such that $f(x) = g \circ \Psi(x)$ for all $x \in X$.

Definition 2. *Let* $\mathbf{D} : \mathcal{X}_1 \times \mathcal{X}_1 \to \mathbb{R}^+$ *be given by*

$$\mathbf{D}(X, Y) := \min_{R\in\mathcal{R}(X,Y)} \max\left(\|\Gamma_{X,Y}\|_{L^\infty(R\times R)}, \|f_X - f_Y\|_{L^\infty(R)}\right),\ X, Y \in \mathcal{X}_1.$$

Proposition 1. *The function* $\mathbf{D}$ *defined above is a metric on (the set of isomorphism classes of)* $\mathcal{X}_1$.

We say that two persistent structures (X, Q_X) and (Y, Q_Y) are *isomorphic* and write $(X, Q_X) \simeq (Y, Q_Y)$, if and only if there exist a bijection $\Phi : X \to Y$ s.t. $Q_Y = Q_X \circ (\Phi, \Phi)$. We define a metric on the collection $\mathcal{Q}$ of all persistent structures by

$$d_{\mathcal{Q}}(X, Y) := \min_{R\in\mathcal{R}(X,Y)} \max_{(x,y),(x',y')\in R} d_{\mathcal{H}}^{(\mathbb{R}^2,L^\infty)}\big(Q_X(x, x'), Q_Y(y, y')\big) \qquad (1)$$

In (1) above, $d_{\mathcal{H}}^{(\mathbb{R}^2,L^\infty)}$ stands for the *Hausdorff distance* ([3]) on subsets of the plane under the L^∞ metric.

Proposition 2. $d_{\mathcal{Q}}$ *defines a metric on (the isomorphism classes of)* $\mathcal{Q}$.

Now one has *stability* on the functor $\mathcal{C}^*$, i.e. the application $(X, d_X, f_X) \mapsto (X, Q_X^*)$ is stable in an appropiate sense.

Theorem 2. *For two filtered spaces* (X, d_X, f_X) *and* (Y, d_Y, f_Y) *in* $\mathcal{X}_1$ *consider the associated persistent structures* (X, Q_X^*) *and* (Y, Q_Y^*) *defined in Example 2. Then, one has* $d_{\mathcal{Q}}((X, Q_X^*), (Y, Q_Y^*)) \leqslant \mathbf{D}(X, Y)$.

5 Acknowledgements

Our research has been supported by DARPA grant number HR0011-05-1-0007 (GC and FM), NSF DMS-0406992 (GC) and ONR grant number N00014-09-1-0783 (FM).

References

1. Shai Ben-David, Ulrike von Luxburg, and Dávid Pál. A sober look at clustering stability. In Gábor Lugosi and Hans-Ulrich Simon, editors, *COLT*, volume 4005 of *Lecture Notes in Computer Science*, pages 5–19. Springer, 2006.
2. Gérard Biau, Benoît Cadre, and Bruno Pelletier. A graph-based estimator of the number of clusters. *ESAIM Probab. Stat.*, 11:272–280, 2007.
3. D. Burago, Y. Burago, and S. Ivanov. *A Course in Metric Geometry*, volume 33 of *AMS Graduate Studies in Math.* American Mathematical Society, 2001.
4. G. Carlsson and F. Mémoli. Persistent Clustering and a Theorem of J. Kleinberg. *ArXiv e-prints*, August 2008.
5. Antonio Cuevas, Manuel Febrero, and Ricardo Fraiman. Cluster analysis: a further approach based on density estimation. *Comput. Statist. Data Anal.*, 36(4):441–459, 2001.
6. H. Edelsbrunner, D. Letscher, and A. Zomorodian. Topological persistence and simplification. In *Proc. 41st Ann. IEEE Sympos. Found Comput. Sci.*, pages 454–463, 2000.
7. Martin Ester, Hans Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. pages 226–231. AAAI Press, 1996.
8. J. A. Hartigan. Consistency of single linkage for high-density clusters. *J. Amer. Statist. Assoc.*, 76(374):388–394, 1981.
9. John A. Hartigan. *Clustering algorithms*. John Wiley & Sons, New York-London-Sydney, 1975. Wiley Series in Probability and Mathematical Statistics.
10. Nicholas Jardine and Robin Sibson. *Mathematical taxonomy*. John Wiley & Sons Ltd., London, 1971. Wiley Series in Probability and Mathematical Statistics.
11. Jon M. Kleinberg. An impossibility theorem for clustering. In Suzanna Becker, Sebastian Thrun, and Klaus Obermayer, editors, *NIPS*, pages 446–453. MIT Press, 2002.
12. Jussi Klemelä. Visualization of multivariate density estimates with level set trees. *J. Comput. Graph. Statist.*, 13(3):599–620, 2004.
13. G. N. Lance and W. T. Williams. A general theory of classificatory sorting strategies 1. Hierarchical systems. *Computer Journal*, 9(4):373–380, February 1967.
14. Saunders Mac Lane. *Categories for the working mathematician*, volume 5 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, second edition, 1998.
15. W. Stuetzle and R. Nugent. A generalized single linkage method for estimating the cluster tree of a density, 2008.
16. Werner Stuetzle. Estimating the cluster type of a density by analyzing the minimal spanning tree of a sample. *J. Classification*, 20(1):25–47, 2003.
17. D. Wishart. Mode analysis: a generalization of nearest neighbor which reduces chaining effects. In *Numerical Taxonomy*, pages 282–311. Academic Press, 1969.
18. M. Anthony Wong and Tom Lane. A kth nearest neighbour clustering procedure. *J. Roy. Statist. Soc. Ser. B*, 45(3):362–368, 1983.