Multiparameter clustering algorithms

Gunnar Carlsson and Facundo Mémoli[1]

March 18,2009

[1] {gunnar,memoli}@math.stanfod.edu

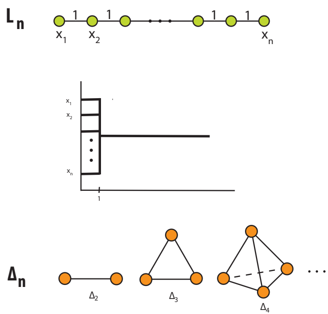
Table of Contents

Introduction

- Not much is known about the theoretical properties of clustering methods, [Kle02, vLBD05, BDvLP06]
- In Hierarchical Clustering (HC) average linkage (AL) and complete linkage (CL) HC methods are preferred over single linkage (SL) HC.
- Reasons for this are
  - SL is insensitive to density: **chaining effect.**
  - CL tends to create clusters that are highly connected and compact: **cliques**.
  - AL performs averaging in order to define the linkage value of two clusters, and this gives some sensitivity to density.

SLHC and the chaining effect

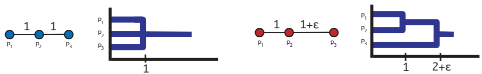
SLHC applied to the two metric spaces below yields the dendrogram in the center.



CLHC applied to a slighly perturbed version of Δ_n produces a dendrogram very similar (in a precise metric sense) to the one in the center of the figure. In contrast, when applied to a slighly perturbed version of L_n, it shows that the set gets connected with much more *effort*.. see next slide

continued...

- AC and CL are **unstable** in a precise sense [CM09]. CLHC applied to the two metric spaces below yields very different dendrograms.



- It can therefore be said that CLHC's sensitivity to density is actually directly related to its instability . This is a rather unsatisfactory situation.

- SL is the **unique** HC scheme that satisfies certain reasonable axioms *à la* Kleinberg [CM08].

- Kleinberg proved in [Kle02] that there exist no **standard clustering** algorithm that simultaneously satisfies 3 natural conditions: *scale invariance, consistency* and *richness*.

- In [CM08] we proved that in the *relaxed* context of HC methods, conditions similar to Kleinberg's yield **uniqueness** instead of non-existence.

Overview of previous work

Assume $\mathbb{X} := \{x_1, \ldots, x_n\} \subset X$ is sample from an unknown Borel probability measure (with compact support) μ_X defined on a metric space (X, d_X). Think of $(X, d_X) = (\mathbb{R}^d, \|\cdot\|)$ and that μ_X admits a density ρ (w.r.t. Lebesgue measure).

- Wishart's **Mode analysis** [Wis]: clusters should reflect *modes* of the underlying density.
- Hartigan followed this line and in [Har75] proposed looking at the *high density clusters* of ρ: for each $\sigma \geqslant 0$ define $L_\rho(\sigma) := \{x | \, \rho(x) > \sigma\}$.
- The high density clusters at level σ are defined to be the *connected components* of $L_\rho(\sigma)$.

Single mode analysis

- Typical procedures consist of fixing the level σ, estimating ρ by $\widehat{\rho}$ and then *computing* $L_{\widehat{\rho}}(\sigma)$ using single linkage (for a fixed threshold).
- Typically, methods consist of four steps:
  1. From the observations $\mathbb{X}$ construct a density estimate $\widehat{\rho}$.
  2. Choose a level σ and find all observations $\mathbb{X}^\sigma := \{x \in \mathbb{X} | \, \widehat{\rho}(x) > \sigma\}$.
  3. Construct a **graph** $\mathcal{G}^{(\sigma,\varepsilon)}(\mathbb{X})$ connecting each observation $x \in \mathbb{X}^\sigma$ to all other observations in $y \in \mathbb{X}^\sigma$ s.t. $\|x - y\| \leqslant \varepsilon$.
  4. Define high density clusters to be the connected components of the graph $\mathcal{G}^{(\sigma,\varepsilon)}(\mathbb{X})$.
- One can actually think of performing SL HC instead of the fixed ε clustering.

Single mode analysis and the Cluster tree

- A problem with single mode analysis is that one choice of the cut level σ may not reveal the whole structure of the data. The dependence of the separation of different clusters on the choice of σ is critical. So one idea would be to do this for all σ at the same time.
- Hartigan observed that the collection of all these high density clusters (for all levels) have a *hierarchical structure*: for any clusters A and B, (1) $A \subset B$, (2) $B \subset A$ or (3) $A \cap B = \emptyset$. This hierarchical structure is known as the **cluster tree** $T(\rho)$ of ρ [Stu03, SN08].
- Stuetzle gives algorithm for estimating the cluster tree $T(\rho)$ based on the observations $\mathbb{X}$.

Cluster tree

- Constructions such as single mode analysis (with SLHC) and the cluster tree both can be seen as *slices* of the information contained in the whole collection of graphs

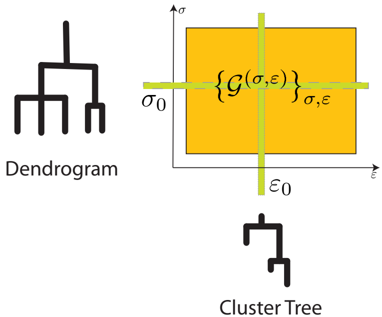
$$\left\{\mathcal{G}^{(\sigma,\varepsilon)}\right\}_{\sigma,\varepsilon}.$$

- Indeed, if you fix $\varepsilon = \varepsilon_0$, then $\left\{\mathcal{G}^{(\sigma,\varepsilon_0)}\right\}_{\sigma}$ contains all the information necessary for constructing an estimate of the cluster tree.

- If you fix $\sigma = \sigma_0$, $\left\{\mathcal{G}^{(\sigma_0,\varepsilon)}\right\}_{\varepsilon}$ contains all the information for single mode analysis (a dendrogram representation thereof).

Our proposal: Multiparameter Clustering

We claim that it is more powerful and general to obtain information directly from the whole two-parameter family $\left\{\mathcal{G}^{(\sigma,\varepsilon)}\right\}_{\sigma,\varepsilon}$.



- By looking at this two parameter family of graphs, we construct **invariants** that summarize the **multiscale** (ε) and **multilevel** (σ) information contained in the density estimate.

Formulation

Definition 1. *A* filtered metric space *is a triple* (X, d_X, f_X) *where* X *is a finite set,* d_X *is a metric on* X *and* $f_X : X \to \mathbb{R}$. *The function* f_X *is called the* filter. *For* $\sigma \in \mathbb{R}$ *let* $X_\sigma := \{x \in X,\ f_X(x) \leqslant \sigma\}$.

Definition 2 (Persistent Structures)**.** *Given a finite set* X*, a* persistent structure *on* X *is a map* $Q_X : X \times X \to \mathcal{B}(\mathbb{R}^2)$ *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. *Technical condition ("semi-closedness").*

Categories, functors..

We formulate our results in the language of **Category Theory**. Categories are useful mathematical constructs that encode the nature of certain objects of interest together with a set of admissible/interesting/useful maps between them. This formalism is extremely useful for studying classes of mathematical objects which share a common structure, such as sets, groups, vector spaces, or topological spaces.

Categories

Definition 3. *A category $\underline{C}$ consists of*

- *A collection of objects $ob(\underline{C})$ (e.g. sets, groups, vector spaces, etc.)*
- *For each pair of objects $X, Y \in ob(\underline{C})$, a set $Mor_{\underline{C}}(X, Y)$, the morphisms from X to Y (e.g. maps of sets from X to Y, homomorphisms of groups from X to Y, linear transformations from X to Y, etc. respectively).*
- *For each object $X \in \underline{C}$, a distinguished element $id_X \in Mor_{\underline{C}}(X, X)$*
- *Composition operations: $\circ : Mor_{\underline{C}}(X, Y) \times Mor_{\underline{C}}(Y, Z) \to Mor_{\underline{C}}(X, Z)$, corresponding to composition of set maps, group homomorphisms, linear transformations, etc.*

The composition is assumed to be associative in the obvious sense, and for any $f \in Mor_{\underline{C}}(X, Y)$, it is assumed that $id_Y \circ f = f$ and $f \circ id_X = f$.

Simple examples

Example 1. *Here we consider four extremely simple categories, that we are going to refer to as $\underline{0}, \underline{1}, \underline{2}$ and $\underline{3}$ respectively.*

- *The category $\underline{0}$ has $ob(\underline{0}) = \varnothing$ and all the conditions in the definition above are trivially satisfied.*
- *Consider the category $\underline{1}$ with exactly one object A and one morphism: $Mor_{\underline{1}}(A, A) = f$. It follows that f must be the identity morphism id_A. This is represented graphically as follows:*

$$\begin{tikzcd} A \arrow[loop right, "id_A"] \end{tikzcd}$$

- *The category $\underline{2}$ has exactly two objects A and B and three morphisms: the identities from A to A and from B to B and exactly one morphism in $Mor_{\underline{2}}(A, B)$:*

$$\begin{tikzcd} A \arrow[loop left, "id_A"] \arrow[rr, "f"] & & B \arrow[loop right, "id_B"] \end{tikzcd}$$

continued..

Example. • *Finally, the category* $\underline{3}$ *has exactly three objects* A, B *and* C *and six morphisms: the identities from* A *to* A, *from* B *to* B *and* C *to* C, *and three more morphisms,* $Mor_{\underline{3}}(A,B) = f$, $Mor_{\underline{3}}(B,C) = g$ *and* $Mor_{\underline{3}}(A,C) = h$*:*

$$\begin{tikzcd}
A \arrow[loop left, "id_A"] \arrow[rr, "f"] \arrow[dr, "h"'] & & B \arrow[loop right, "id_B"] \arrow[dl, "g"] \\
& C \arrow[loop below, "id_C"] &
\end{tikzcd}$$

Now, note that in order to satisfy composition one must have $h = g \circ f$.

Example: A category of Sets

Example 2. *A category of sets Consider the category* $\underline{Sets}$ *whose objects are sets and whose morphisms are maps between two sets:* $Mor_{\underline{Sets}}(A, B)$ *comprises all maps from the set* A *to the set* B. *The identity map* $id_A : A \to A$ *is the obvious* $a \mapsto a$ *and composition is defined as usual by* $(g \circ f)(a) = g(f(a))$.

Example: a category of persistent structures

Example 3. *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 compatible*** *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 compatible maps between X and Y.

Another example: a category of filtered metric spaces

Example 4. *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

$$f_X(x) \geqslant f_Y(\phi(x')).$$

Functors

Definition 4. *Let $\underline{C}$ and $\underline{D}$ be categories. Then a* functor *from $\underline{C}$ to $\underline{D}$ consists of*

- *A map of sets $F : ob(\underline{C}) \to ob(\underline{D})$*
- *For every pair of objects $X, Y \in \underline{C}$ a map of sets $\Phi(X,Y) : Mor_{\underline{C}}(X,Y) \to Mor_{\underline{D}}(FX, FY)$ so that*
  1. *$\Phi(X,X)(id_X) = id_{F(X)}$ for all $X \in ob(\underline{C})$*
  2. *$\Phi(X,Z)(g \circ f) = \Phi(Y,Z)(g) \circ \Phi(X,Y)(f)$ for all $f \in Mor_{\underline{C}}(X,Y)$ and $g \in Mor_{\underline{C}}(Y,Z)$*

Remark 1. *In the interest of clarity, we often refer to the pair (F, Φ) as a single letter F.*

Clustering Functor

Definition 5. *A* clustering functor *will is a functor*

$$\mathcal{C} : \underline{\mathcal{M}}^{gen} \to \underline{\mathcal{Q}}.$$

$$\begin{array}{ccc} (X, d_X, f_X) & \xrightarrow{\mathcal{C}} & (X, Q_X) \\ \phi \big\downarrow & & \big\downarrow \mathcal{C}(\phi) \\ (Y, d_Y, f_Y) & \xrightarrow{\mathcal{C}} & (Y, Q_Y) \end{array}$$

We now construct the main example.

Main example

Definition 6. *For each* $\varepsilon \geqslant 0$ *and* $\sigma \in \mathbb{R}$ *we define the equivalence relation on* X_σ *given by* $x \sim_{(\varepsilon,\sigma)} x'$ *if and only if there exists* $x = x_0, \ldots, x_m = x'$ *s.t.*

- $\max_i d_X(x_i, x_{i+1}) \leqslant \varepsilon$, *and*
- $\max_i f_X(x_i) \leqslant \sigma$.

For each $x \in X_\sigma$, *let* $[x]_{(\varepsilon,\sigma)}$ *denote the equivalence class to which* x *belongs.*

Example 5. *Consider the functor* $\mathcal{C}^*$ *that when applied to* (X, d_X, f_X) *produces the object* (X, Q_X^*) *where*

$$Q_X^*(x, x') := \{(\varepsilon, \sigma) \in \mathbb{R}^2 |\, x \sim_{(\varepsilon,\sigma)} x'\}.$$

A remark: SLHC

Remark 2. *The sets* $Q^*_X(x,x')$ *are unbounded. They are of the form*

$$\bigcup_{i=1}^{K}[\varepsilon^{(i)},\infty)\times[\sigma^{(i)},\infty).$$

Remark 3. *Assume that* f_X = *constant. Then, the clustering functor* $\mathcal{C}^*$ *generates the same hierarchical information as SLHC.*

It turns out that $\mathcal{C}^*$ is the **unique** clustering functor that satisfies certain properties.

A characterization 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 persistent structure, respectively, and only "remember" the underlying sets X. Then we assume that $\beta \circ \Psi = \alpha$.*

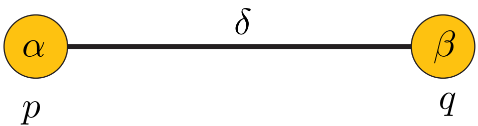
(II) *For $\delta \geqslant 0$ and $\alpha, \beta \in \mathbb{R}$ let $\Delta(\delta, \alpha, \beta) = \left(\{p, q\}, \left(\begin{smallmatrix} 0 & \delta \\ \delta & 0 \end{smallmatrix}\right), \{\alpha, \beta\}\right)$ denote the two point filtered metric space with underlying set $\{p, q\}$, where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

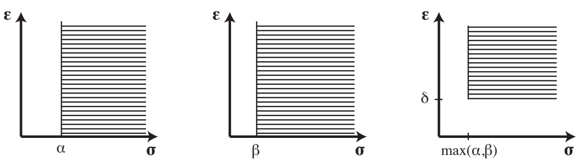
(III) *Given $(\varepsilon, \sigma) \in \mathbb{R}^+ \times \mathbb{R}$ and the filtered metric space (X, d_X, f_X), then $sep(X_\sigma) > \varepsilon$ implies that $(\varepsilon, \sigma) \notin Q_X(x, x')$ for any $x, x' \in X_\sigma$, $x \neq x'$.*

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

continued

$$\Delta(\delta,\alpha,\beta) = \left(\{p,q\}, \left(\begin{smallmatrix} 0 & \delta \\ \delta & 0 \end{smallmatrix}\right), \{\alpha,\beta\}\right)$$

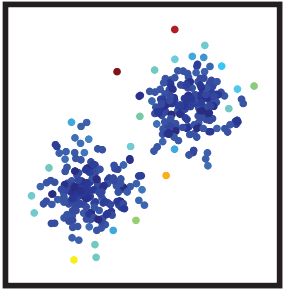


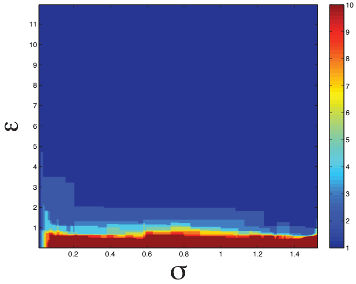


$Q_\Delta(p,p)$ $Q_\Delta(q,q)$ $Q_\Delta(p,q)$

How to use this?

One simple invariant we can look at is the number of connected components of $\mathcal{G}^{(\sigma,\varepsilon)}$ for each value of $\varepsilon \geqslant 0$ and $\sigma \in \mathbb{R}$ and plot this as an image. Large regions with constant number of components suggest relevant features in the data. Let $N(\varepsilon, \sigma)$ denote the number of connected components of $\mathcal{G}^{(\sigma,\varepsilon)}$. Below, we show an example for the sum of two gaussians. The filter was chosen to be the distance to the k-th nearest neighbor ($k = 2$).





A stability theorem

One can define a distance $\mathbf{D}$ on filtered metric spaces and then also a suitable distance $d_{\mathcal{Q}}$ on collection on persistent structures.

Theorem 2. *One has*

$$d_{\mathcal{Q}}((X, Q^*_X), (Y, Q^*_Y)) \leqslant \mathbf{D}(X, Y)$$

This theorem in particular implies the metric stability of SLHC, see [CM08, CM09].

Discussion

- Both SLHC on X^σ and the cluster tree (with fixed connectivity parameter ε) suffer from making an arbitrary choice of parameters.
- Following our extension of Kleinberg's result, and the acceptance that one would like a stable HC method that is sensitive to density, we propose a structure that captures the variations of both a scale and a density parameter.
- We have solid theoretical results for our proposal. The ideas can be applied to reason about clustering methods in general.
- Can be connected to **persistent topology**, [Car09].

Preprints available `http://comptop.stanford.edu/`

Bibliography

References

[BDvLP06] 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.

[Car09] Gunnar Carlsson. Topoloy and data. *Bull. Amer. Math. Soc.*, 46, 2009.

[CM08] Gunnar Carlsson and Facundo Mémoli. Persistent clustering and a theorem of J. Kleinberg. Technical report, 2008.

[CM09] Gunnar Carlsson and Facundo Mémoli. Characterization, stability and convergence of hierarchical clustering algorithms. Technical report, 2009.

[Har75] John A. Hartigan. *Clustering algorithms*. John Wiley & Sons, New York-London-Sydney, 1975. Wiley Series in Probability and Mathematical Statistics.

[Kle02] Jon M. Kleinberg. An impossibility theorem for clustering. In Suzanna Becker, Sebastian Thrun, and Klaus Obermayer, editors, *NIPS*, pages 446–453. MIT Press, 2002.

[SN08] W. Stuetzle and R. Nugent. A generalized single linkage method for estimating the cluster tree of a density, 2008.

[Stu03] 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.

[vLBD05] U. von Luxburg and S. Ben-David. Towards a statistical theory of clustering. presented at the pascal workshop on clustering, london. Technical report, Presented at the PASCAL workshop on clustering, London, 2005.

[Wis] D. Wishart. Mode analysis: a generalization of nearest neighbor which reduces chaining effects.