# Multi-parameter hierarchical clustering and beyond

**Alexander Rolle**
Institute of Geometry, TU Graz
Graz, Austria
rolle@tugraz.at

## Abstract

We survey recent progress on multi-parameter hierarchical clustering, which has developed in several directions since it was introduced by Carlsson–Mémoli in 2010. These lines of research show that tools originally developed in the setting of multi-parameter persistent homology can be applied more broadly, without linearizing via homology.

## 1 Multi-parameter persistence and hierarchical clustering

### 1.1 A short bibliography

Multi-parameter hierarchical clustering was introduced by Carlsson–Mémoli [CM10], who were motivated by the following situation: say you have a finite metric space $X$, together with a function $f : X \to \mathbb{R}$, which could be a density estimate. For $\epsilon > 0$ and $\sigma \in \mathbb{R}$, let $G_{\epsilon,\sigma}$ be the graph with vertices $\{x \in X : f(x) > \sigma\}$, and with an edge between $x$ and $x'$ if $d_X(x, x') \leq \epsilon$. Taking the connected components of the graphs $G_{\epsilon,\sigma}$, along with the functions induced by relations $\epsilon < \epsilon'$ and $\sigma > \sigma'$, you get a two-parameter hierarchical clustering of $X$. This generalizes the single-linkage hierarchical clustering of $X$, and filtering by $f$ may reduce the chaining effect of single-linkage.

Multi-parameter hierarchical clustering has developed in several directions. One line of research concerns zero-dimensional persistent homology [BBOS20, CKMW20]. Another originates with the HDBSCAN clustering algorithm [CMS13]: the connection with multi-parameter persistence was first observed in [MH18], and studied further in [Jar19, Jar20, RS20, Sco20].

In many applications, the goal is to understand clustering structure across a range of distance scales and density thresholds.

### 1.2 Simplifying hierarchical clusterings

A common theme is that multi-parameter hierarchical clusterings contain rich information about the clustering structure of the data, and enjoy good stability properties, but they are often too complicated to use directly. One concrete manifestation of this was established by Bauer, Botnan, Oppermann, and Steen [BBOS20], who consider the natural subcategory of two-parameter persistence modules that includes the modules $H_0(G_{\epsilon,\sigma})$ arising from the two-parameter hierarchical clusterings $\pi_0(G_{\epsilon,\sigma})$ considered by Carlsson–Mémoli. They show that, as with the category of all two-parameter persistence modules, this subcategory is of wild representation type except in very few cases. So, the algebraic complications of two-parameter persistent homology are present even in this subcategory.

The easiest way to extract information (such as a single clustering of the data) from a multi-parameter hierarchical clustering is just to fix some parameters. Of course, this approach cannot capture clustering structure that occurs at different parameter values across the dataset, and furthermore, fixing parameters often introduces stability problems. So, several methods have been proposed to

Topological Data Analysis and Beyond Workshop at the 34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

extract information from a multi-parameter hierarchical clustering without fixing any of the original parameters. In particular, there are two prominent ways to extract a single clustering from a one-parameter hierarchical clustering by finding clusters that persist for a long time in the hierarchy. The ToMATo algorithm [CGOS13] extracts clusters that persist in the sense of zero-dimensional persistent homology, while the HDBSCAN algorithm [CMS13] extracts clusters that persist in a sense specific to hierarchical clusterings: this algorithm scores clusters according to their persistence and the size of their underlying point sets. An analysis of the stability properties of these two methods appears in [RS20]. By applying these methods after taking a suitable one-parameter slice of a multi-parameter hierarchical clustering, one can extract a single clustering of a dataset from a multi-parameter hierarchical clustering, without fixing any of the original parameters. Instead, one chooses other parameters (the slice, and the internal parameters of the "extraction" method of [CGOS13] or [CMS13]), but there are stability results for these.

## 2 Beyond clustering

### 2.1 Multi-parameter persistence and density estimation

A popular way to frame the clustering problem, which is employed by several of the papers we've mentioned, is *density-based clustering*. If $f$ is a density function on Euclidean space, then one can define a hierarchical clustering of the support of $f$ by taking the connected components of its super-level sets. The problem of density-based clustering is to approximate this hierarchical clustering, given samples drawn from $f$. Of course, this problem is closely related to the problem of density estimation: using samples to approximate $f$ itself.

In [RS20], we define the *kernel density bifiltration* (generalizing the degree–Rips bifiltration [LW15]), which filters a metric probability space according to density estimates computed for all possible choices of a bandwidth (distance scale) parameter $s$. This bifiltration satisfies a stability theorem, which implies that, if a sample $X$ is a good approximation of a density function $f$, then the kernel density bifiltration of $X$ approximates the kernel density bifiltration of the support of $f$.

In [RS20], this is used to establish an approach to density-based clustering. Applying single-linkage to the kernel density bifiltration, we get a three-parameter hierarchical clustering; then we take a one-parameter slice, as in Section 1.2 (see Fig. 1).

There is a related approach to density estimation. The resulting estimate can be written as a variable-bandwidth kernel density estimate (a *balloon estimate*, in the language of [TS92]), and it follows from the results of [RS20] that this estimate has appealing stability properties. In this application, there is no longer an invariant from algebraic topology, such as $H_0$ or $\pi_0$, in the picture. However, both the method and the proofs of its guarantees are fundamentally guided by the theory of multi-parameter persistence.

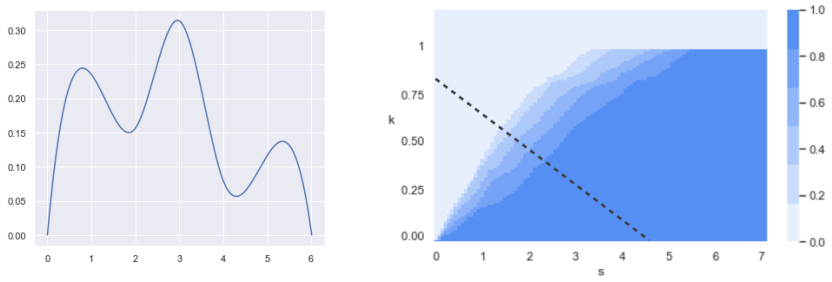


Figure 1: On the left, a probability density function $f$ supported on the real line. On the right, a heat map representing the kernel density bifiltration of the support of $f$. The intensity of color at a point $(s, k)$ in the heat map indicates the proportion of points in the support of $f$ whose density estimate w.r.t. the bandwidth parameter $s$ is greater than $k$. The dashed line indicates a possible choice of one-parameter slice. *Figure:* [RS20].

### 2.2 Further directions

The theory of multi-parameter persistent homology has generated powerful tools for organizing and comparing data "all at once". Some of these tools make sense outside of the linearized setting in which they were first introduced: for example, it is well known that, if one defines a *multi-persistent object* in any category $\mathcal{C}$ as a functor $\mathbb{R}^n \to \mathcal{C}$ (where $\mathbb{R}^n$ is given its usual partial order), then one immediately obtains an interleaving distance on these multi-persistent objects.

These tools have now been applied in various ways to the problem of clustering, and beyond. In many applications, the goal is to avoid fixing distance scales or density thresholds. As such parameters appear in many geometrically motivated algorithms, it is natural to ask: what other methods of exploratory data analysis have multi-parameter analogues? Where else can the tools of multi-parameter persistence be applied?

## Acknowledgments and Disclosure of Funding

Supported by Austrian Science Fund (FWF) grant number P 29984-N35. I would like to thank the referees for their comments.

## References

[BBOS20] Ulrich Bauer, Magnus B. Botnan, Steffen Oppermann, and Johan Steen. Cotorsion torsion triples and the representation theory of filtered hierarchical clustering. *Advances in Mathematics*, 369:107171, 2020.

[CGOS13] Frédéric Chazal, Leonidas J. Guibas, Steve Y. Oudot, and Primoz Skraba. Persistence-based clustering in riemannian manifolds. *J. ACM*, 60(6), November 2013.

[CKMW20] Chen Cai, Woojin Kim, Facundo Mémoli, and Yusu Wang. Elder-rule-staircodes for augmented metric spaces. In *36th International Symposium on Computational Geometry (SoCG 2020)*, 2020.

[CM10] Gunnar Carlsson and Facundo Mémoli. Multiparameter hierarchical clustering methods. In Hermann Locarek-Junge and Claus Weihs, editors, *Classification as a Tool for Research*, pages 63–70, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg.

[CMS13] Ricardo J. G. B. Campello, Davoud Moulavi, and Jörg Sander. Density-based clustering based on hierarchical density estimates. In *Advances in Knowledge Discovery and Data Mining*, volume 7819 of *Lecture Notes in Computer Science*, pages 160–172. Springer, 2013.

[Jar19] John F. Jardine. Stable components and layers. *Canad. Math. Bull.*, page 1–15, 2019.

[Jar20] John F. Jardine. Persistent homotopy theory. arXiv:2002.10013, 2020.

[LW15] Michael Lesnick and Matthew Wright. Interactive visualization of 2-D persistence modules. arXiv:1512.00180v1, 2015.

[MH18] Leland McInnes and John Healy. Accelerated hierarchical density based clustering. In *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*, volume 00, pages 33–42, Nov. 2018.

[RS20] Alexander Rolle and Luis Scoccola. Stable and consistent density-based clustering. arXiv:2005.09048, 2020.

[Sco20] Luis Scoccola. Locally persistent categories and metric properties of interleaving distances. *Electronic Thesis and Dissertation Repository*. `https://ir.lib.uwo.ca/etd/7119`, 2020.

[TS92] George R. Terrell and David W. Scott. Variable kernel density estimation. *The Annals of Statistics*, 20(3):1236–1265, 1992.