Journal of Machine Learning Research 26 (2025) 1-26 Submitted 7/24; Revised 1/25; Published 1/25

An Axiomatic Definition of Hierarchical Clustering

Ery Arias-Castro EARIASCASTRO@UCSD.EDU
Department of Mathematics and Halıcıoğlu Data Science Institute
University of California, San Diego
La Jolla, CA 92093, USA

Elizabeth Coda ECODA@UCSD.EDU
Department of Mathematics
University of California, San Diego
La Jolla, CA 92093, USA

Editor: Daniel Hsu

Abstract

In this paper, we take an axiomatic approach to defining a population hierarchical clustering for piecewise constant densities, and in a similar manner to Lebesgue integration, extend this definition to more general densities. When the density satisfies some mild conditions, e.g., when it has connected support, is continuous, and vanishes only at infinity, or when the connected components of the density satisfy these conditions, our axiomatic definition results in Hartigan's definition of cluster tree.

Keywords: clustering, hierarchical clustering, Hartigan cluster tree

1. Introduction

Clustering, informally understood as the grouping of data, is a central task in statistics and computer science with broad applications. Modern clustering algorithms originated in the work of numerical taxonomists, who developed methods to identify hierarchical structures in the classification of plant and animal species. Since then clustering has been used in disciplines such as medicine, astronomy, anthropology, economics, etc., with aims such as exploratory analysis, data summarization, the identification of salient structures in data, and information organization.

The notion of a "good" or "accurate" clustering varies between fields and applications. For example, to some computer scientists, the correct clustering of a dataset is often defined as the solution to an optimization problem (think K-means) and a good algorithm either solves or approximates a solution to this problem, ideally with some guarantees (Puzicha et al., 2000; Dasgupta, 2016; Arthur and Vassilvitskii, 2007). From this perspective, the dataset is viewed as fixed, and the cluster definition is based on the data alone (Hennig, 2015). Moreover, depending on the particular application, how good a clustering is deemed to be may be further loosened, such as in the task of image segmentation, where a good clustering need only "extract the global impression of an image" according to Shi and Malik (2000). For an early discussion and criticism of this vagueness, see the paper of Cormack (1971) surveying the field at the time, where the author says that "There are many intuitive ideas, often conflicting, of what constitutes a cluster, but few formal definitions." More recent discussions include those of Estivill-Castro (2002), von Luxburg et al. (2012) and that of Hennig (2015). As Gan et al. (2021) say in their recent book on clustering, "The clustering problem has been addressed extensively, although there is no uniform definition for data clustering and there may never be one".

Even as this algorithmic view of clustering is widespread well outside computer science, it is not satisfactory from a statistical inference perspective. Indeed, in statistics, it is typically assumed that the sample is representative of an underlying population and a clustering method, to be useful, should inform the analyst about that population. This viewpoint calls for a definition of clustering at the population level. When the sample is assumed iid from an underlying distribution representing the population, by clustering we mean a partition of the support of that distribution, and in that case, a clustering of the sample is deemed "good" or "accurate" by reference to the population clustering — and a clustering algorithm is a good one if it is consistent, meaning, exact in the large-sample limit. This reference to the population is what gives meaning to statistical inference, and to questions such as whether an observed cluster is "real" or not.

Our contribution is the following. *We propose a set of axioms for population hierarchical clustering and show that this axiomatic definition is well-founded and essentially coincides with the cluster tree introduced by Hartigan (1975).*

1.1 Definitions of Clustering Algorithms

While we take an axiomatic approach to defining the population-level hierarchical clustering, several previous works have explored axiomatic approaches to defining clustering algorithms.

1.1.1 Flat Clustering

The most famous of these works which might be that of Kleinberg (2002), who proposes three axioms (scale-invariance, richness, and consistency[1]) and establishes an impossibility theorem stating that no clustering algorithm can simultaneously satisfy all three axioms.

As pointed out by Ben-David and Ackerman (2008); Strazzeri and Sánchez-García (2022); Cohen-Addad et al. (2018) and others, including Kleinberg himself in the same article (Kleinberg, 2002, Sec 5), the consistency axiom may not be so desirable. Rather, a relaxation of this property, in which a refinement or coarsening of the clusters is allowed, may be more appropriate. Kleinberg states that clustering algorithms that satisfy scale-invariance, richness, and this relaxed notion of refinement-coarsening consistency do exist and clustering algorithms that satisfy scale-invariance, near richness, and refinement consistency also exist. This was, in a sense, confirmed by Cohen-Addad et al. (2018), who allow the number of clusters to vary with the refinement. Zadeh and Ben-David (2009) show that, if the number of clusters that a clustering algorithm can return is fixed at k, there exist clustering algorithms that satisfy scale-invariance, k-richness, and consistency in the original sense of (Kleinberg, 2002). They also show that single linkage is the unique clustering algorithm returning a fixed number of clusters simultaneously satisfying these axioms and two additional axioms.

1. The 'consistency' axiom is not in the statistical sense, but refers to the property that if within-cluster distances are decreased and between-cluster distances are enlarged, then the output clustering does not change.

Puzicha et al. (2000) consider clustering data via optimization of a suitable objective function and define a suitable objective function with a set of axioms. Though their axioms are somewhat strong, requiring the objective function have an additive structure, they show that only one of the objective functions considered satisfies all of their axioms. Ben-David and Ackerman (2008) also propose a set of axioms which strongly parallel Kleinberg's axioms for a clustering quality measure function and show the existence of functions satisfying these axioms.

1.1.2 Hierarchical Clustering

In the 1960s and 1970s, there was some interest in developing an axiomatic foundation of hierarchical clustering. For example, Jardine and Sibson (1968); Jardine et al. (1967); Sibson (1970) list axioms that, according to them, a hierarchical clustering algorithm should satisfy, and then state that single linkage is the only algorithm they are aware of that satisfies all of their axioms.

More recently, in large part motivated by the work of Kleinberg (2002), Carlsson and Mémoli (2010) propose their own sets of axioms for hierarchical clustering, and then prove that single linkage is the only algorithm that satisfies them.[2]

1.2 Definitions of Population Clustering

There is not a generally accepted definition of clustering at the population level.

1.2.1 Flat Clustering

Given a clustering algorithm based on optimizing an objective function, it is sometimes possible to extend that objective function to apply to a probability distribution, in which case the solution to the resulting optimization problem becomes a natural definition of clustering for the population defined by the probability distribution—and the clustering algorithm has good chances of being statistically consistent for that population clustering. Famously, this is the case of K-means, as first established by Pollard (1981) in the context of Euclidean spaces, later extended by Pärna (1990, 1986); Jaffe (2022) to the broader context of Banach and other metric spaces.

Another popular approach assumes that the data is drawn from a mixture model $f = \sum_{i=1}^{k} \alpha_i f_i$ and the population level clustering consists of k clusters corresponding to the mixture components (Fraley and Raftery, 2002; Bouveyron et al., 2019; McLachlan et al., 2019; Everitt, 2013; McLachlan and Peel, 2000). If the underlying density is not a mixture, it can be approximated by a mixture model (typically chosen to be a multivariate Gaussian), though this requires a modeling choice. This approximation may require a very large number of components to approximate well, resulting in an artificially large number of clusters. Moreover, even if the density is a mixture, under this definition, a unimodal mixture could have multiple clusters. We mention that the estimation of mixtures has led to a number of algorithms and methods including the famous EM algorithm (Dempster et al., 1977) and more recent spectral approaches (Anandkumar et al., 2012; Hsu and Kakade, 2013).

2. Though this result has been presented as a demonstration that Kleinberg's impossibility theorem does not hold when hierarchical clustering algorithms are considered, this connection is somewhat unclear to us, as the proposed axioms do not mirror Kleinberg's axioms very precisely.

Alternatively, in the gradient flow approach to defining the population level clustering, often attributed to Fukunaga and Hostetler (1975), each point is assigned to the nearest mode (i.e. local maximum) in the direction of the gradient. Thus, at least when the density has Morse regularity, the clusters correspond to the basins of attraction of the modes. Although this definition relies on assumptions about the smoothness of the density and does not account for arbitrarily flat densities (Menardi, 2016), it overcomes some of the described difficulties of mixture model clustering. That said, if the components in the mixture model are well-separated, this definition results in a similar clustering to the mixture-based definition (Chacón, 2020).

1.2.2 Hierarchical Clustering

Adopting a hierarchical perspective of clustering, Hartigan (1975) proposes a population-level cluster tree, where clusters correspond to the maximally connected components of density upper level sets. Though Hartigan provides minimal motivation for this definition beyond observing that each cluster C in his tree "conforms to the informal requirement that C is a high-density region surrounded by a low-density region", this definition of hierarchical clustering at the population level has led to a substantial amount of work (Eldridge et al., 2015; Wang et al., 2019; Chaudhuri et al., 2014; Kim et al., 2016; Balakrishnan et al., 2013; Steinwart, 2011). We note that Hartigan's definition of hierarchical clustering has been shown to be fully compatible with Fukunaga and Hostetler's definition of flat clustering (Arias-Castro and Qiao, 2023).

1.3 Contribution and Content

Although multiple definitions of population clustering exist (Section 1.2), we are not aware of definitions that rely on axioms. Inspired by the axiomatic approaches to defining clustering algorithms (Section 1.1), we propose a set of axioms that we believe a population hierarchical clustering ought to satisfy. We show that this axiomatic definition is well-posed and further show that it leads to the definition put forth by Hartigan (1975) under some conditions, for example, when the underlying population has a continuous density with connected support satisfying some additional mild assumptions.

We note that, although an axiomatic definition of population flat clustering is also lacking, we focus entirely on hierarchical clustering.

The organization of the paper is as follows. Section 2 provides some basic notation and definitions. In Section 3, we take an axiomatic approach to defining a hierarchical clustering for a piecewise constant density with connected support. In Section 4, we extend this definition to continuous densities, first to densities with connected support, and then to more general densities. Section 5 is a discussion section where we go over some extensions, some practical considerations, and also discuss some outlook on flat clustering. In an appendix, we provide a close examination of the merge distortion metric of Eldridge et al. (2015) (Section A), and provide further technical details for the special case of a Euclidean space (Section B).

2. Preliminaries

Throughout this paper, we will work with a metric space (Ω, d). For technical reasons, we assume it is locally connected, which is for example the case if the balls are connected. This is so that the connected components of an open set are connected.[3]

In principle, we would equip this metric space with a suitable Borel measure, and consider densities with respect to that measure. As it turns out, this equipment is not needed as we can directly work with non-negative functions. We will do so for the most part, although we will sometimes talk about densities.

For a set $A \subseteq \Omega$, we let $\text{int}(A)$ or A° denote its interior and $\text{clo}(A)$ or $\overline{A}$ denote its closure; we also let $\text{cc}(A)$ denote the collection of its connected components. For a function $f : \Omega \to \mathbb{R}$, its support is $\text{supp}(f) = \text{clo}\{f \neq 0\}$, and for $\lambda \in \mathbb{R}$, its upper λ-level set is defined as $\{f \geq \lambda\}$, denoted U_λ when there is no ambiguity.

Definition 1 (Hierarchical clustering or cluster tree) *A hierarchical clustering, also referred to as a cluster tree, of $\mathcal{X} \subseteq \Omega$ is a collection of connected subsets of $\mathcal{X}$, referred to as clusters, that has a nested structure in that two clusters are either disjoint or nested.*

A hierarchical clustering of a function f is understood as a hierarchical clustering of its support $\text{supp}(f)$. Hartigan's definition of hierarchical clustering for a density is arguably the most well-known one.

Definition 2 (Hartigan cluster tree) *The Hartigan cluster tree of a function f, which will be denoted $\mathcal{H}_f$, is the collection consisting of the maximally connected components of the upper λ-level sets of f for all $\lambda > 0$. $\mathcal{H}_f$ is a hierarchical clustering of supp(f).*

A dendrogram is commonly understood as the output of a hierarchical clustering algorithms such as single-linkage clustering. It turns out to be simpler to work with dendrograms instead of directly with cluster trees (Carlsson and Mémoli, 2010; Eldridge et al., 2015).

Definition 3 (Dendrogram) *A dendrogram is a cluster tree equipped with a real-valued non-increasing function defined on the cluster tree called the height function. A dendrogram is thus of the form $(\mathcal{C}, h)$ where $\mathcal{C}$ is a cluster tree and $h : \mathcal{C} \to \mathbb{R}$ is such that $h(C') \geq h(C)$ whenever $C' \subseteq C$.*

The Hartigan tree of a function f is naturally equipped with the following height function

$$h_f(C) = \inf_{x \in C} f(x). \tag{1}$$

Note that this function has the required monotonicity.

Eldridge et al. (2015) introduced the merge distortion metric to compare dendrograms. It is based on the notion of merge height, which gives the height at which two points stop belonging to the same cluster, or equivalently, the height of the smallest cluster that contains both points.

3. This is, in fact, an equivalence, the proof of which is left as an exercise in Armstrong's textbook (Armstrong, 1983, Ch 3).

Definition 4 (Merge height) *Let $(\mathcal{C}, h)$ be a dendrogram. The merge height of two points $x, y \in \Omega$ is defined as*

$$m_{(\mathcal{C},h)}(x,y) = \sup_{\substack{C \in \mathcal{C} \\ x,y \in C}} h(C). \tag{2}$$

For the special case of an Hartigan cluster tree,

$$m_f(x,y) = m_{(\mathcal{H}_f,h_f)}(x,y) = \sup_{\substack{C \text{ connected} \\ x,y \in C}} \inf_{z \in C} f(z). \tag{3}$$

Definition 5 (Merge distortion metric) *For two dendrograms, $(\mathcal{C}, h)$ and $(\mathcal{C}', h')$, their merge distortion distance is defined as*

$$d_M((\mathcal{C},h),(\mathcal{C}',h')) = \sup_{x,y \in \Omega} \big| m_{(\mathcal{C},h)}(x,y) - m_{(\mathcal{C}',h')}(x,y) \big|.$$

The merge distortion metric has the following useful property (Eldridge et al., 2015, Th 17).

Lemma 6 *For two functions f and g,*

$$d_M((\mathcal{H}_f,h_f),(\mathcal{H}_g,h_g)) \le \|f-g\|_\infty.$$

Proof The arguments in (Eldridge et al., 2015) are a little unclear (likely due to typos), but correct arguments are given in (Kim et al., 2016, Lem 1). We nonetheless provide a concise proof as it is instructive. Take $x, y \in \Omega$, and let $s = m_f(x,y)$ and $t = m_g(x,y)$. We need to show that $|s-t| \le \eta := \|f-g\|_\infty$. For any $\epsilon > 0$, by (3), there is a connected set C containing x and y such that $f(z) \ge s - \epsilon$ for all $z \in C$. Since this implies that $g(z) \ge s - \epsilon - \eta$ for all $z \in C$, by (3) again, this yields $t \ge s - \epsilon - \eta$. We have thus shown that $s \le t + \eta + \epsilon$, and can show that $t \le s + \eta + \epsilon$ in exactly the same way, which combined allows us to obtain that $|s-t| \le \eta+\epsilon$. With $\epsilon > 0$ arbitrary, we conclude. ■

The merge distortion metric has gained some popularity in subsequent works that discuss the consistency of hierarchical methods (Kim et al., 2016; Wang et al., 2019). In Section A we discuss some limitations and issues with the merge distortion metric, which is in fact a pseudometric on general cluster trees. However, in the context in which we use the metric, these issues are not significant.

We also introduce the notion of neighboring sets. Throughout, we adopt the convention that the empty set is disconnected.

Definition 7 (Neighboring regions) *Given a collection of sets $\mathcal{A} = \{A_i\}$, we define the neighborhood of A_i as*

$$\mathcal{N}(A_i) = \bigcup \big\{ A_j : int\big(\overline{A_i} \cup \overline{A_j}\big) \text{ is connected} \big\}. \tag{4}$$

Note that $A_j \subseteq \mathcal{N}(A_i) \;\Leftrightarrow\; A_i \subseteq \mathcal{N}(A_j)$, so that we may speak of A_i and A_j as being neighbors, which we will denote by $A_i \sim A_j$.

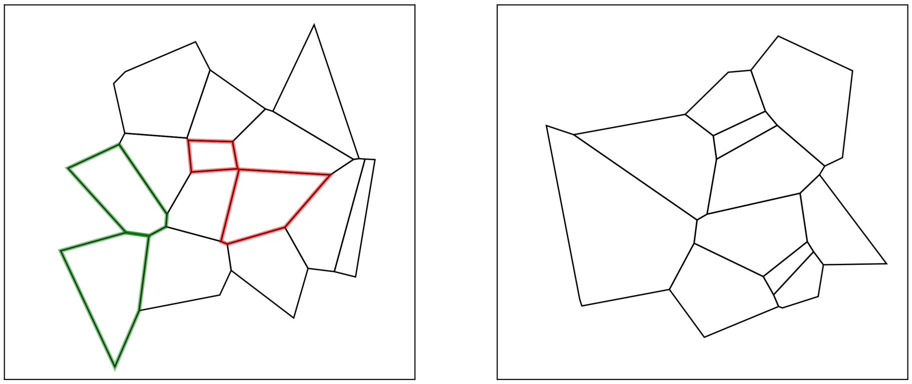

Figure 1: Left: A collection of sets with neighboring regions (green) and non-neighboring regions (red). This collection of sets does not have the internally connected property. Right: A collection of sets with the internally connected property.

Under this definition, in a Euclidean space, balls that only meet at one point are not neighbors, and neither are rectangles in dimension three that intersect only along an edge. Our discussion will be simplified in the case where we consider collections where all sets that intersect are neighbors.

Definition 8 (Internally connected property) *Let $\mathcal{A} = \{A_i\}$ be a collection of sets. We say $\mathcal{A}$ has the internally connected property if*

$$\overline{A_i} \cup \overline{A_j} \ \ \textit{connected} \ \ \Rightarrow \ int\big(\overline{A_i} \cup \overline{A_j}\big) \ \textit{connected} \ .$$

Figure 1 illustrates these two definitions.

3. Axioms

In this section, we develop a definition of the population cluster tree for a density f. Inspired by previous axiomatic approaches to clustering algorithms and in the spirit of Lebesgue integration, we propose a set of axioms for a population cluster tree when the density is piecewise constant with connected support. We then extend this definition to more general densities, and arrive at a definition that is equivalent to Hartigan's tree (Definition 2) for continuous densities with multiple connected components, under some mild assumptions.

3.1 Axioms for Piecewise Constant Functions

Previous work has discussed difficulties in defining what the "true" clusters are (Cormack, 1971; Hartigan, 1975; Hennig, 2015; von Luxburg et al., 2012), observing that there may not be a single definition for all intents and purposes. So as to simplify the situation as much as possible so that a definition may arise as natural, we first consider piecewise constant functions with connected, bounded support. These piecewise constant functions can be

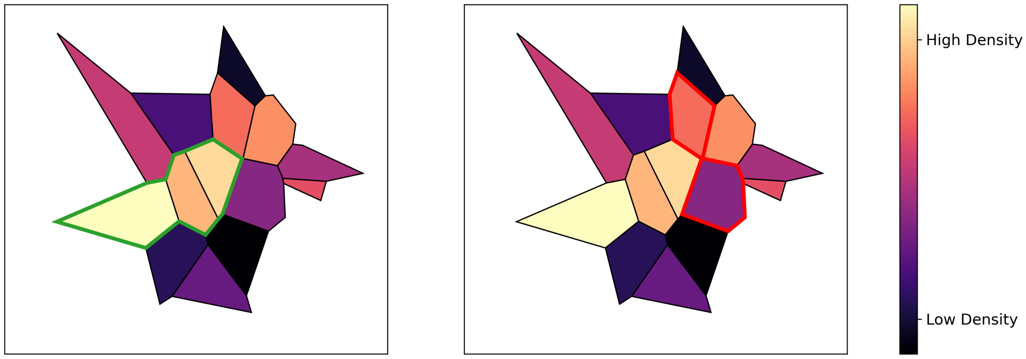

Figure 2: A piecewise constant density in $\mathcal{F}$. On the left, the highlighted region may be a cluster under Axiom 1 and on the right, the highlighted region is not a cluster under Axiom 1 as the interior is not connected.

viewed as a mixture of uniform distributions. A function in that class is of the form

$$f = \sum_{i=1}^{m} \lambda_i \, \mathbb{I}_{A_i}, \tag{5}$$

where, for all i, $\lambda_i > 0$ and A_i is a connected, bounded region with connected interior, and we also require that $\operatorname{supp}(f) = \bigcup_{i=1}^{m} \overline{A_i}$ has connected interior. Additionally, without loss of generality, assume the A_i are disjoint. Let $\mathcal{F}$ denote the class of all such functions.

Remark 9 *We require each region A_i and the entire support to not only be connected, but have connected interior, and the same is true of the clusters (Axiom 1). It is well-known that the closure of a connected set is always connected, so that this is a stronger requirement, and is meant to avoid ambiguities.*

For $f \in \mathcal{F}$ we propose that a hierarchical clustering $\mathcal{C}$ should satisfy the following three axioms. For what it's worth, Axiom 1 and Axiom 3 were put forth early on by Carmichael et al. (1968) and, most famously although not as directly, by Hartigan (1975), and also correspond to the 7th item on the list of "desirable characteristics of clusters" suggested by Hennig (2015), and Axiom 2 can be motivated by the 13th item on Hennig's list.

3.1.1 Axiom 1: Clusters have connected interior

We propose that any cluster in $\mathcal{C}$ should not only be connected, but have a connected interior. With Axiom 2 below in place, see (A2), we may express Axiom 1 as follows:

$$\text{If } C \in \mathcal{C} \text{ and } A_i, A_j \subseteq C, \text{ then there are } A_{k_1}, \dots, A_{k_n} \subseteq C \\ \text{such that } A_i \sim A_{k_1} \sim \cdots A_{k_n} \sim A_j. \tag{A1}$$

For example, for the density in Figure 2, the highlighted region in the right hand figure should not be a cluster in $\mathcal{C}$, but the highlighted region in the left hand figure could be a

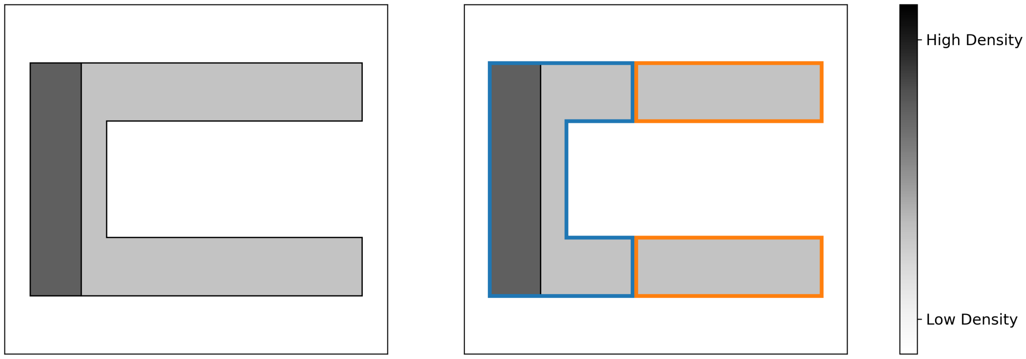

Figure 3: Left: A simple example of a piecewise constant density built on two regions. Right: The clustering output of K-means with number of clusters $K = 2$. One of the clusters is disconnected.

cluster in $\mathcal{C}$. This reflects the idea that elements of a cluster should in some sense be similar to each other, without imposing additional assumptions on the within-cluster distances, between-cluster distances, the relative sizes of clusters, or the shape of clusters.

The condition that a cluster be a connected region was considered early on in the literature as it was part of the postulates put forth by Carmichael et al. (1968). However, it is important to note that this condition is not enforced in other definitions of what a cluster is. Most prominently, K-means can return disconnected clusters– see Figure 3 for an example.

3.1.2 Axiom 2: Clusters do not partition connected regions of constant density

We propose that a connected region with constant density should not be broken up into smaller clusters as this would impose an additional structure that is not present in the density. We may write this axiom as:

$$\text{Any } C \in \mathcal{C} \text{ is of the form } C = \bigcup_{i \in I} A_i \text{ for some } I \subseteq \{1, 2, \dots, m\}. \tag{A2}$$

Figure 4 depicts and example of a valid and invalid cluster under this axiom. Note that as a consequence of this axiom, the within-cluster distances may be larger than the between-cluster distances, depending on the relative widths and separations between regions.

We find this condition to be particularly natural in the present situation where the density is piecewise constant. It is in essence already present in the concept of relatedness introduced by Carmichael et al. (1968). But it is important to note that other definitions do not enforce this property. This is the case of K-means, which can split connected regions of constant density—see, again, Figure 3 for an example.

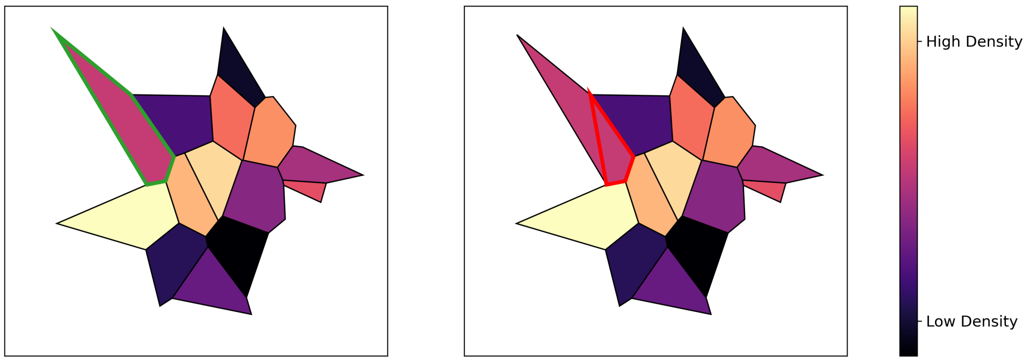

Figure 4: On the left, the highlighted region could be a cluster under Axiom 2, but the highlighted region on the right oversegments a region of constant density, and should not be a cluster.

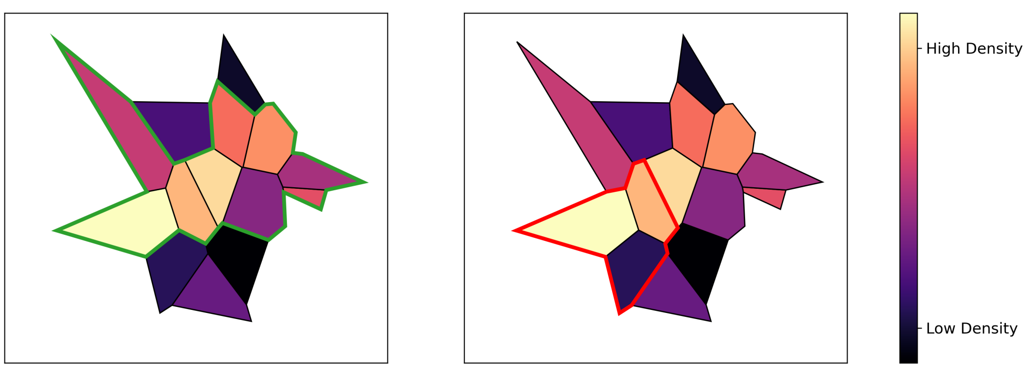

Figure 5: On the left, the lowest density in highlighted cluster exceeds the largest density in a neighboring set. On the right, the highlighted cluster contains a region with lower density than a neighbor, and thus this should not be a cluster.

3.1.3 Axiom 3: Clusters are surrounded by regions of lower density

We propose that a cluster should be surrounded by regions of lower density, meaning that:

$$\text{For any } C \in \mathcal{C}, \text{ it holds that } \inf_{x \in C} f(x) > \sup_{x \in \mathcal{N}(C) \setminus C} f(x), \tag{A3}$$

where, if $C = \bigcup_{i \in I} A_i$, then $\mathcal{N}(C) = \bigcup_{i \in I} \mathcal{N}(A_i)$ denotes the neighbor of C, extending the definition given in (4). Figure 5 includes an example.

This is one of the postulates of Carmichael et al. (1968), although it was perhaps most popularized by Hartigan in his well-known book (Hartigan, 1975, Ch 11). Although it is not part of most other approaches to clustering—K-means being among those as Figure 3 shows—we find that this condition is rather compatible with the colloquial understanding of 'points clustering together'.

3.2 Finest hierarchical clustering

Definition 10 (Finer cluster tree) *We say that a cluster tree $\mathcal{C}$ is finer than (or a refinement of) another cluster tree $\mathcal{C}'$ if $\mathcal{C}$ includes all the clusters of $\mathcal{C}'$, namely, $C \in \mathcal{C}' \Rightarrow C \in \mathcal{C}$.*

As it turns out, given a nonnegative function, there is one, and only one, finest cluster tree among those satisfying the axioms above.

Proposition 11 *For any $f \in \mathcal{F}$, there exists a unique finest hierarchical clustering of f among those satisfying the axioms.*

Proof Let f be as in (5). The proof is by construction. Let $\mathcal{C}_*$ denote the collection of every cluster that satisfies (A1), (A2), and (A3). Clearly, it suffices to show that $\mathcal{C}_*$ is a hierarchical clustering (Definition 1). Take two clusters in $\mathcal{C}_*$, say $C_1 = \bigcup_{i \in I_1} A_i$ and $C_2 = \bigcup_{i \in I_2} A_i$. We need to show that C_1 and C_2 are either disjoint or nested. Suppose for contradiction that this is not the case, so that C_1 and C_2 are neither disjoint nor nested. Since they are not disjoint, there is $i \in I_1 \cap I_2$, so that $A_i \subseteq C_1 \cap C_2$. And since they are not nested, there is $j \in I_1 \setminus I_2$, so that $A_j \subseteq C_1 \setminus C_2$. By (A1), there are $i_1, \dots, i_s \in I_1$ such that $A_i \sim A_{i_1} \sim \cdots \sim A_{i_s} \sim A_j$. Let $t = \max\{q : A_{i_q} \subseteq C_2\}$, so that $A_{i_t} \subseteq C_2$ while $A_{i_{t+1}} \nsubseteq C_2$, and in particular $A_{i_{t+1}} \subseteq \mathcal{N}(C_2) \setminus C_2$, and applying (A3), we get

$$\min_{C_2} f > \max_{\mathcal{N}(C_2) \setminus C_2} f \geq \lambda_{i_{t+1}} \geq \min_{C_1} f,$$

using at the end the fact that $A_{i_{t+1}} \subseteq C_1$. However, we could also get the reverse inequality, $\min_{C_1} f > \min_{C_2} f$, in the same exact way, which would result in a contradiction. ∎

Proposition 11 justifies the following definition.

Definition 12 (Finest axiom cluster tree) *For $f \in \mathcal{F}$, we denote by $\mathcal{C}_f^*$ the finest cluster tree of f among those satisfying the axioms.*

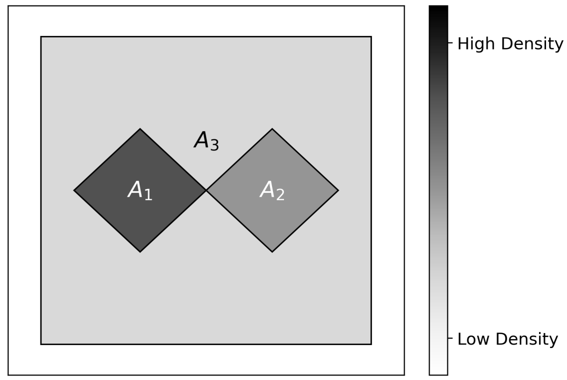

Figure 6: An example where Hartigan's cluster tree does not satisfy the axioms, so that $\mathcal{C}_f^* \neq \mathcal{H}_f$. Indeed, $\mathcal{H}_f = \{A_1, A_1 \cup A_2, A_1 \cup A_2 \cup A_3\}$ but $A_1 \cup A_2 \notin \mathcal{C}_f^*$ because $\mathrm{int}(A_1 \cup A_2)$ is not connected. Instead, we have $\mathcal{C}_f^* = \{A_1, A_2, A_1 \cup A_2 \cup A_3\}$.

3.3 Comparison with Hartigan's Cluster Tree

It is natural to compare the finest axiom cluster tree of Definition 12 with the Hartigan cluster tree of Definition 2. First, observe that for $f \in \mathcal{F}$, $\mathcal{H}_f$ satisfies (A2) and (A3). However, $\mathcal{H}_f$ need not satisfy (A1), as clusters in $\mathcal{H}_f$ are only required to be connected. As a result, in general, the Hartigan tree $\mathcal{H}_f$ is not the same as the finest axiom cluster tree $\mathcal{C}_f^*$. A counter example is given in Figure 6.

We define $\mathcal{F}_{\mathrm{int}}$ as the class of functions in $\mathcal{F}$ with $\{A_i\}$ in (5) having the internally connected property (Definition 8).

Theorem 13 *For any $f \in \mathcal{F}_{\mathrm{int}}$, it holds that $\mathcal{C}_f^* = \mathcal{H}_f$.*

Proof First, observe that under our assumption, $\mathcal{H}_f$ satisfies all axioms (A1), (A2), and (A3). Thus, because $\mathcal{C}_f^*$ is the finest cluster tree among those satisfying the axioms (Proposition 11), it must be the case that $\mathcal{H}_f \subseteq \mathcal{C}_f^*$.

For the reverse inclusion, take any $C \in \mathcal{C}_f^*$. We want to show that $C \in \mathcal{H}_f$. Recalling the definition of h_f in (1), define $\lambda = h_f(C)$ and let M be the maximally connected subset of $\{f \geq \lambda\}$ that contains C. We need to show that $C = M$. Noting that C is of the form $\bigcup_{i \in I} A_i$ because of (A2), and that M must be of the form $\bigcup_{j \in J} A_j$ because f is of the form (5), and that M contains C by definition, it is the case that $I \subseteq J$.

Suppose for contradiction that $C \neq M$, so that $I \neq J$. Since M is connected, there must be A_i in C and A_j in $M \setminus C$ such that $A_i \cup A_j$ is connected. As is well-known, this implies that $\overline{A_i \cup A_j} = \overline{A_i} \cup \overline{A_j}$ is connected, and since $f \in \mathcal{F}_{\mathrm{int}}$, $\mathrm{int}(\overline{A_j} \cup \overline{A_i})$ is also connected, in turn implying that $A_i \sim A_j$. Applying (A3), we get that $\lambda > f(A_j)$, and this is a contradiction since $A_j \subseteq M$ and M is part of the upper λ-level set. ■

Remark 14 *As a relaxation of Axiom 1, we could simply require a cluster to be connected, and allow it to have disconnected interior. If the definition of a neighboring region were* *also relaxed so that if the closure of the union of two sets is connected, then the sets are considered neighbors, then the relaxed Axiom 1, original Axiom 2, and original Axiom 3 would yield an axiomatic definition of a cluster tree that is identical to the Hartigan tree for* $f \in \mathcal{F}$*. All that said, we find the requirement that the interior be connected in our original Axiom 1 (and in Definition 7) to be more natural and robust.*

4. Extension to Continuous Functions

Having defined the finest axiom cluster tree for a piecewise constant function (Definition 12), we now examine its implication when piecewise constant functions are used to approximate continuous functions. More specifically, we consider sequences of piecewise constant functions in $\mathcal{F}_{\text{int}}$ converging to a continuous function, and show that, under some conditions, the corresponding finest axiom cluster trees converge to the Hartigan cluster tree of the limit function in merge distortion metric (Definition 5).

4.1 Functions with Connected Support

We start with continuous functions whose support has connected interior.

Definition 15 *Given a continuous function* f *with connected support, we say that* $\mathcal{C}$ *is an axiom cluster tree for* f *if there is a sequence* $(f_n) \subseteq \mathcal{F}_{\text{int}}$ *that uniformly approximates* f *such that*

$$\lim_{n\to\infty} d_M((\mathcal{C}^*_{f_n}, h_{f_n}), (\mathcal{C}, h_f)) = 0. \tag{6}$$

At this point it is not clear whether a continuous function admits an axiom cluster tree. However, if it does, then its Hartigan cluster tree is one of them and, moreover, all other axiom cluster trees are zero merge distortion distance away.

Theorem 16 *Suppose* f *is a continuous function that admits an axiom cluster tree. Then its Hartigan tree* $\mathcal{H}_f$ *is an axiom cluster tree for* f*. Moreover, if* $\mathcal{C}$ *is an axiom cluster tree for* f*, then* $d_M((\mathcal{C}, h_f), (\mathcal{H}_f, h_f)) = 0$*.*

Proof Let $\mathcal{C}$ be an axiom cluster tree for f. By Definition 15, there is a sequence (f_n) in $\mathcal{F}_{\text{int}}$ that converges uniformly to f such that (6) holds. By the triangle inequality,

$$d_M((\mathcal{C}, h_f), (\mathcal{H}_f, h_f)) \le d_M((\mathcal{C}, h_f), (\mathcal{C}^*_{f_n}, h_{f_n})) + d_M((\mathcal{C}^*_{f_n}, h_{f_n}), (\mathcal{H}_f, h_f)).$$

We already know that the first term on the RHS tends to zero. For the second term, using Theorem 13 and Lemma 6,

$$d_M((\mathcal{C}^*_{f_n}, h_{f_n}), (\mathcal{H}_f, h_f)) = d_M((\mathcal{H}_{f_n}, h_{f_n}), (\mathcal{H}_f, h_f)) \le \|f_n - f\|_\infty \to 0, \quad n \to \infty. \tag{7}$$

We thus have that $d_M((\mathcal{C}, h_f), (\mathcal{H}_f, h_f)) = 0$ — this being true for any axiom cluster tree $\mathcal{C}$. In the process, we have also shown in (7) that $\mathcal{H}_f$ is axiomatic. ■

The remaining of this subsection is dedicated to providing sufficient conditions on a function f for the existence of sequence $(f_n) \subseteq \mathcal{F}_{\text{int}}$ that converges uniformly to f. In formalizing this, we will utilize the following terminology and results.

Definition 17 (Internally connected partition property) *We say that Ω has the internally connected partition property if it is connected, and for any $r > 0$, there exists a locally finite partition $\{A_i\}$ of Ω that has the internally connected property and is such that, for all i, A_i is connected with connected interior and diameter at most r.*

We establish in Proposition 25 that any Euclidean space (and, consequently, of any finite-dimensional normed space) has the internally connected partition property. And we conjecture that this extends to some Riemannian manifolds.

Proposition 18 *Suppose (Ω, d) is a metric space where all closed and bounded sets are compact*[4]*, and that has the internally connected partition property. Let $f : \Omega \to [0, \infty)$ be continuous with all upper level sets bounded, and such that the upper λ-level set is connected when $\lambda > 0$ is small enough. Then, there is a sequence $(f_n) \in \mathcal{F}_{\rm int}$ that converges uniformly to f.*

Proof It is enough to show that, for any $\eta > 0$, there is a function in $\mathcal{F}_{\rm int}$ within η of f in supnorm. Therefore, fix $\eta > 0$, and take it small enough that the upper η-level set, $K = \{x : f(x) \geq \eta\}$, is connected. Consider

$$K_1 = \big\{y : d(y, x) \leq 1, \text{ for some } x \in K\big\}.$$

In particular, K_1 is compact, and since f is continuous on K_1, it is uniformly so, and therefore there exists $0 < \epsilon < 1$ such that, if $x, y \in K_1$ are such that $d(x, y) \leq \epsilon$, then $|f(x) - f(y)| \leq \eta$.

By the fact that Ω has the internally connected partition property, it admits a locally finite partition $\{A_i\}$ with the internally connected property and such that, for all i, A_i has connected interior and diameter at most ϵ. Let

$$I = \{i : A_i \cap K \neq \emptyset\},$$

and note that I is finite and that $K \subseteq \bigcup_{i \in I} A_i \subseteq K_1$. For $i \in I$, let $\lambda_i = \sup_{x \in A_i} f(x)$. Because $A_i \cap K \neq \emptyset$, we have $\lambda_i \geq \eta$. Finally, we define the piecewise constant function $g = \sum_{i \in I} \lambda_i \mathbb{I}_{A_i}$. We claim that $g \in \mathcal{F}_{\rm int}$. Since $\{A_i : i \in I\}$ inherits the internally connected property from $\{A_i\}$, all we need to check is that $\bigcup_{i \in I} \overline{A_i}$ is connected. To see this, first note that it is enough that $\bigcup_{i \in I} A_i$ be connected (since the closure of a connected set is connected). Suppose for contradiction that $\bigcup_{i \in I} A_i$ is disconnected, so that we can write it as a disjoint union of $\bigcup_{i \in I_1} A_i$ and $\bigcup_{i \in I_2} A_i$, where I_1 and I_2 are non-empty disjoint subsets of I. Because $K \subseteq \bigcup_{i \in I} A_i$, then we have that K is the disjoint union of $K_1 := \bigcup_{i \in I_1} A_i$ and $K_2 := \bigcup_{i \in I_2} A_i$, both non-empty by construction, so that K is not connected—a contradiction.

We now show that $\|f - g\|_\infty \leq \eta$. For $x \notin \bigcup_{i \in I} A_i$, $g(x) = 0$ and since $x \notin K$, $f(x) < \eta$, so that $|f(x) - g(x)| \leq \eta$. For $x \in A_i$, for some $i \in I$, $g(x) = f(y)$ for some $y \in \overline{A_i}$, and because $x, y \in K_1$ and $d(x, y) \leq \epsilon$, we have $|f(y) - f(x)| \leq \eta$. ∎

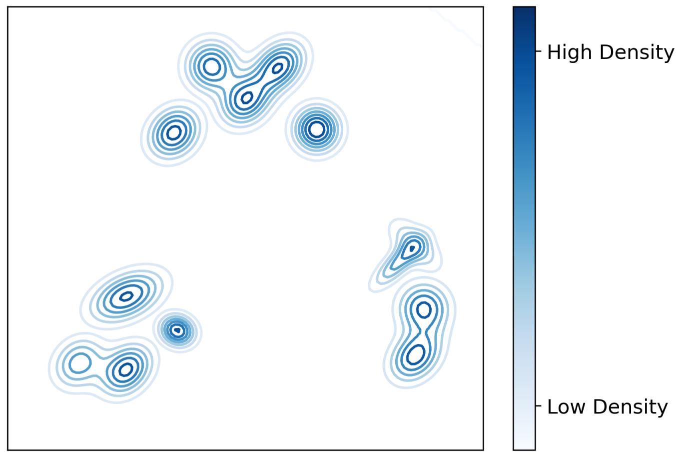

Figure 7: An example of density with a support that has disconnected interior which appears to exhibit some clustering structure beyond that happening inside each of its eight connected components.

4.2 Functions with Disconnected Support

So far, we have focused our attention on densities whose support has connected interior. However, there is no real difficulty in extending our approach to more general densities. Indeed, given a function with support having disconnected interior, our approach can define a hierarchical clustering of each connected component of $\{f > 0\}$.

In more detail, let f be a function of the form

$$f = \sum_{j=1}^{N} f_j, \tag{8}$$

where $\mathrm{int}(\mathrm{supp}(f_j)) \cap \mathrm{int}(\mathrm{supp}(f_k)) = \emptyset$ when $j \neq k$. First, suppose that each $f_j \in \mathcal{F}$. If we apply the axioms of Section 3, we obtain that C is a cluster for f if and only if it is a cluster for one of the f_j, and consequently that the finest axiom cluster tree for f is simply the union of the finest axiom cluster trees for the f_j, i.e.,

$$\mathcal{C}_f^* = \bigcup_{j=1}^{N} \mathcal{C}_{f_j}^*.$$

If f is continuous, that is, if each f_j in (8) is continuous, we may proceed exactly as in Section 4.1 and, based on the facts that $\mathcal{H}_f = \bigcup_j \mathcal{H}_{f_j}$, $h_f(C) = h_{f_j}(C)$ when $C \in \mathcal{C}_{f_j}$, and

$$d_M((\mathcal{C}, h_f), (\mathcal{C}', h_f)) = \max_{j=1,\dots,N} d_M((\mathcal{C}_j, h_{f_j}), (\mathcal{C}_j', h_{f_j})),$$

for any two axiom cluster trees for f, $\mathcal{C} = \bigcup_j \mathcal{C}_j$ and $\mathcal{C}' = \bigcup_j \mathcal{C}_j'$ (all cluster trees for f are of that form), we find that Theorem 16 applies verbatim.

4. This is sometimes called the Heine–Borel property.

This is as far as our approach goes. The end result is Hartigan's cluster tree, with the same caveats that come from using the merge distortion metric detailed in Section A. In particular, instead of a tree we have a forest with N trees in general, one for each f_j. We find this end result natural, but if it is desired to further group regions (see Figure 7 for an illustration), one possibility is to apply a form of agglomerative hierarchical clustering to the 'clusters', $\text{supp}(f_1), \dots, \text{supp}(f_N)$. (In our definition, these are not clusters of $\mathcal{H}_f$, but this is immaterial.) Doing this presents the usual question of what clustering procedure to use, but given what we discuss in Section 5.3.1, single-linkage clustering would be a very natural choice.

5. Discussion

5.1 Extensions

We speculate that our axiomatic definition of hierarchical clustering can be extended beyond continuous functions (Section 4) to piecewise continuous functions with connected support by the same process of taking a limit of sequences in $\mathcal{F}_{\text{int}}$ that uniformly approximate the function of interest f.

The natural approach is to work within each region where f is continuous, say R, and to consider there a partition of R that would allow the definition of a piecewise constant function approximating f uniformly on R. The main technical hurdle is the construction of such a partition with the internally connected property, as a region R may not be regular enough to allow for that. Additionally, even if there is a partition with the internally connected property on each region, taken together, these partitions may not have the internally connected property. We see some possible workarounds, but their implementation may be complicated.

5.2 Limitations

We consider an example where the second axiom may be undesirable. Suppose the support of a piecewise constant density includes a region of constant density consisting of two balls connected by a narrow bottleneck. Under the axiom of connectedness, this forms one cluster that cannot be segmented into any smaller clusters. However, it may be preferred to consider the two balls as two separate clusters.

We concede that if the bottleneck is narrow enough, this is reasonable, though if the width of the bottleneck increases, eventually the two balls should not be considered two separate clusters. If we were to modify the axiom and consider the degree of connectedness, we would need to decide how to measure connectedness (e.g., via the Cheeger constant) and choose a threshold value for connectedness above which a region of constant density would not be divisible. Although such a variant of the second axiom could be desirable in practical situations, at the fundamental level, the specification of a measure of connectedness and the choice of a threshold would involve a certain degree of arbitrariness, and it would make the axiom quantitative and no longer be qualitative, thus resulting in a substantial loss of elegance and naturalness.

5.3 Practical Implications

We first examine some implications of adopting the axioms defining clusters in Section 3.

5.3.1 Algorithms

A large majority of existing algorithms for hierarchical clustering do not return cluster tree estimators that are asymptotically consistent with the proposed population hierarchical clustering. This is true of all agglomerative hierarchical clustering algorithms that we know of, with the partial exception of single-linkage clustering, as repeatedly pointed out by Hartigan (1977, 1981, 1985). Interestingly, single-linkage clustering arises out of various axiomatic discussions of flat clustering[5] such as (Kleinberg, 2002; Ben-David and Ackerman, 2008; Zadeh and Ben-David, 2009; Cohen-Addad et al., 2018), as well as in axiomatic discussions of hierarchical clustering algorithms (Jardine and Sibson, 1968; Carlsson and Mémoli, 2010).

The fact that single linkage clustering arises as a viable candidate in multiple axiomatic definitions of clustering is despite the heavy criticism in the literature for its 'chaining' tendencies. Indeed, in practice this behavior can be a concern, and regularized variants of single-linkage clustering are often preferred. Most prominently, this includes the "robust" variant of single-linkage clustering proposed in Chaudhuri et al. (2014), and a hierarchical extension of DBSCAN (Ester et al., 1996), as described in Wang et al. (2019). Both of these estimators have been shown to be consistent in the merge distortion metric for the estimation of Hartigan's cluster tree, by Eldridge et al. (2015) and Wang et al. (2019), respectively.

5.3.2 Hierarchical Clustering in High Dimensions

Wang et al. (2019) derive minimax rates for the estimation of the Hartigan cluster tree, which turn out to match the corresponding minimax rates for density estimation in the L_∞ norm under assumptions of Hölder smoothness on the density. In particular, these rates exhibit the usual behavior in that they require that the sample size grow exponentially with the dimension. This is a real limitation of adopting the axiomatic definition that we propose.

That been said, the usual caveats apply in that the curse of dimensionality is with respect to the intrinsic dimension if the density is in fact with respect to a measure supported on a lower-dimensional manifold (Balakrishnan et al., 2013); and assuming more structure can help circumvent the curse of dimensionality, as done for example in (Chacón, 2019), where a mixture is fitted to the data before applying modal clustering.

Acknowledgements

We would like to thank Sanjoy Dasgupta for suggesting the potential issue with the second axiom discussed in Section 5.2.

5. The hierarchical single-linkage algorithm can be terminated early to obtain a flat clustering.

Appendix A. Merge Distortion Metric

In this section we discuss some limitations and issues of the merge distortion metric. We restrict our attention to the situation considered in (Eldridge et al., 2015) where the height of a tree is defined by the density itself as in (1). We denote the density by f and the corresponding height function by h, and we identify a cluster tree $\mathcal{C}$ with the dendrogram $(\mathcal{C}, h)$ whenever needed. We only consider cluster trees $\mathcal{C}$ made of clusters $C \in \mathcal{C}$ satisfying $h(C) > 0$. Our discussion applies to non-negative functions, and throughout this section, f will be non-negative.

The main issue that we want to highlight is that the merge distortion metric is only a pseudometric, and not a metric, on general cluster trees, as it is possible to have $d_M(\mathcal{C}, \mathcal{C}') = 0$ even when $\mathcal{C}$ and $\mathcal{C}'$ are not isomorphic. (To be clear, we take the partially ordered sets $\mathcal{C}$ and $\mathcal{C}'$ to be isomorphic if they are order isomorphic.) Two examples of this follow.

Example 1 *Consider $f = \frac{1}{2}\mathbb{I}_{A_1} + \frac{1}{3}\mathbb{I}_{A_2} + \frac{1}{6}\mathbb{I}_{A_3}$ where the A_i are disjoint sets with unit measure. Let $\mathcal{C} = \{A_1, A_1 \cup A_2, A_1 \cup A_2 \cup A_3\}$ and $\mathcal{C}' = \{A_1, A_2, A_1 \cup A_2, A_1 \cup A_2 \cup A_3\}$. Both $\mathcal{C}$ and $\mathcal{C}'$ are cluster trees and it can be checked that $m_{\mathcal{C}}(x, y) = m_{\mathcal{C}'}(x, y)$ for all x, y so that $d_M(\mathcal{C}, \mathcal{C}') = 0$. However, the two trees are clearly not isomorphic.*

Example 2 *Consider $f = \mathbb{I}_A$ where A has unit measure. Then any collection of subsets of A with a nested structure is a cluster tree for f, and the merge distortion distance between any pair of such cluster trees is zero.*

The issue in the preceding examples arises because a cluster tree contains nested clusters with the same cluster height. For example, in Example 1, the addition of the cluster A_2 to $\mathcal{C}$ does not change the merge height of any two points, and hence the merge distance between $\mathcal{C}$ and $\mathcal{C}'$ is zero.

Note that neither of these examples compare Hartigan trees, and we suspect in the original merge distortion metric paper (Eldridge et al., 2015), the claim (without proof) that if the merge distortion metric is zero then the trees must be isomorphic was intended in the context of comparing Hartigan trees. This is true for comparing Hartigan trees of continuous densities on $\mathbb{R}^d$, as for Hartigan trees of continuous functions f, g,

$$d_M(\mathcal{H}_f, \mathcal{H}_g) = \|f - g\|_\infty.$$

This is established in (Kim et al., 2016, Lem 1). The proof of that result can be adapted to extend the result to the case where f is continuous and g is piecewise-continuous satisfying an additional regularity condition that, for every x in its support, there exists a δ small enough such that g is continuous on a half-ball centered at x of radius δ.

In view of Theorem 16, we are particularly interested in understanding how different a cluster tree $\mathcal{C}$ such that $d_M(\mathcal{C}, \mathcal{H}_f) = 0$ can be from $\mathcal{H}_f$. The following results clarify the situation. The λ-level set of f is defined as

$$L_\lambda = \{f = \lambda\}.$$

Proposition 19 *Let f be a continuous density. Consider a collection of clusters of the form*

$$\mathcal{C} = (\mathcal{H}_f \setminus \{C_i : i \in I\}) \cup \{\mathcal{S}_j : j \in J\}, \tag{9}$$

where $C_i \in cc(U_{\lambda_i})$ for some $\lambda_i > 0$ such that $\{\lambda_i : i \in I\}$ has empty interior; and $\mathcal{S}_j$ is a cluster tree of L_{λ_j} for some $\lambda_j > 0$ such that $\{\lambda_j : j \in J\}$ are all distinct. Then $\mathcal{C}$ is a cluster tree for f satisfying $d_M(\mathcal{C}, \mathcal{H}_f) = 0$.

Proof We will use the fact that, by continuity of f, the supremum in (2) is attained, or more specifically, that if $\lambda = m_f(x, y)$, there is a connected component C of U_λ that contains x and y. The continuity of f also implies that, for any subset C, $h(C) = h(\overline{C})$.

We first show that any $\mathcal{C}$ defined as in (9) is a cluster tree. Indeed, the removal of any number of clusters preserves the nested structure. Now, consider adding $\mathcal{S}_j$, a cluster tree for L_{λ_j} for some $\lambda_j > 0$. We may clearly assume that $\mathcal{S}_j$ is a cluster tree for a connected component of L_{λ_j}, say B_j, which is itself contained in some $C_j \in \text{cc}(U_{\lambda_j})$, so that $S \subseteq B_j \subseteq C_j$ for any $S \in \mathcal{S}_j$. Take $C \in \mathcal{H}_f$ distinct from C_j. We show that either $S \cap C = \emptyset$ or $S \subseteq C$ for any $S \in \mathcal{S}_j$. Let $\lambda = h(C)$ so that C is a connected component of U_λ. If $\lambda = \lambda_j$, C_j and C are disjoint. If $\lambda < \lambda_j$, B_j is disjoint from C unless C contains C_j. If this is the case, C also contains B_j, and therefore S. If $\lambda > \lambda_j$, $B_j \subseteq L_{\lambda_j}$, $C \subseteq U_\lambda$, and $L_{\lambda_j} \cap U_\lambda = \emptyset$. Take $S' \in \mathcal{S}_k$. We show that S and S' are either disjoint or nested. This is the case if $j = k$ by assumption that $\mathcal{S}_j$ is a cluster tree. For $j \neq k$, B_j and B_k are disjoint since, by assumption, $\lambda_j \neq \lambda_k$ in that case. (We have used the fact that two distinct clusters in $\mathcal{H}_f$ have disjoint boundaries.)

To go further, we use the assumption that $\Lambda := \{\lambda_i : i \in I\}$ has empty interior. We want to show that $m_\mathcal{C}(x, y) = m_f(x, y)$ for any pair of points x and y. First, consider $\mathcal{C}_1 = \mathcal{H}_f \setminus \{\text{cc}(U_{\lambda_i}) : i \in I\}$. Clearly, because the merge height is defined based on a supremum, $m_{\mathcal{C}_1}(x, y) \leq m_\mathcal{C}(x, y)$. Let $\lambda = m_f(x, y)$, so that there is $C \in \text{cc}(U_\lambda)$ such that $x, y \in C$. If $\lambda \neq \lambda_i$ for all $i \in I$, then $m_{\mathcal{C}_1}(x, y) \geq \lambda$. If $\lambda = \lambda_i$ for some $i \in I$, we reason as follows. For $t < \lambda$, let C_t be the connected component of U_t that contains C. Then $x, y \in C_t$ for all $t < \lambda$, and therefore $m_{\mathcal{C}_1}(x, y) \geq t$ for any $t < \lambda$ not in Λ. Since Λ has empty interior, its complement is dense in Λ, and by continuity of f this implies that $m_{\mathcal{C}_1}(x, y) \geq \lambda$. We have thus established that $m_{\mathcal{C}_1}(x, y) \geq \lambda = m_f(x, y)$, which then implies $m_\mathcal{C}(x, y) \geq m_f(x, y)$. Next, consider $\mathcal{C}_2 = \mathcal{H}_f \cup \{\mathcal{S}_j : i \in J\}$, so that $m_{\mathcal{C}_2}(x, y) \geq m_\mathcal{C}(x, y)$. Consider $S \in \mathcal{S}_j$, so that $S \subseteq C_j$ for some $C_j \in \text{cc}(U_{\lambda_j})$. Because $h(S) \leq h(C_j)$ and $C_j \in \mathcal{H}_f$, the merge height of x and y is not increased by adding S to $\mathcal{H}_f$. Therefore, $m_{\mathcal{C}_2}(x, y) \leq m_f(x, y)$, which then implies that $m_\mathcal{C}(x, y) \leq m_f(x, y)$. ■

It turns out that the condition (9) is not necessary for a cluster tree $\mathcal{C}$ to satisfy $d_M(\mathcal{C}, \mathcal{H}_f) = 0$ — although we believe it is not far from that. To deal with the possible removal of clusters from $\mathcal{H}_f$, we only consider cluster trees satisfying the following regularity condition. We say that a cluster tree $\mathcal{C}$ is closed (for $h = h_f$) if it is closed under intersection and union in the sense that, for any sub-collection of nested clusters $\mathcal{S} \subseteq \mathcal{C}$, $\bigcap_{C \in \mathcal{S}} C \in \mathcal{C}$ and, if $\inf_{C \in \mathcal{S}} h(C) > 0$, $\bigcup_{C \in \mathcal{S}} C \in \mathcal{C}$. (Note that this is automatic when $\mathcal{S}$ is finite, but below we will consider infinite sub-collections.)

Lemma 20 *Suppose $\mathcal{C}$ is a closed cluster tree. Then the supremum defining the merge height in (2) is attained, meaning that for any x, y there is $C \in \mathcal{C}$ containing x, y such that $m_\mathcal{C}(x, y) = h(C)$.*

Proof Fix x, y and let $\lambda = m_{\mathcal{C}}(x, y)$, assumed to be strictly positive. It suffices to show that there is a cluster that contains these points with height at least λ.

By the definition in (2), for any $m \geq 1$ integer, there is $C_m \in \mathcal{C}$ that contains x, y such that $h(C_m) > \lambda(1 - 1/m)$. Note that C_m and C_n have at least x, y in common, so that they must be nested. Therefore the sub-collection $\{C_m : m \geq 1\}$ is nested, and by the fact that $\mathcal{C}$ is closed, $C = \bigcap_{m \geq 1} C_m$ is a cluster in $\mathcal{C}$. By monotonicity of h, $h(C) \geq h(C_m)$ for all m, so that $h(C) \geq \lambda$. ■

To simplify things further, we just avoid talking about what happens within level sets. We will use the following results.

Lemma 21 *For any $s, t > 0$, the connected components of $\{f > s\}$ and those of $\{f \geq t\}$ are either disjoint or nested.*

Proof Let R be a connected component of $\{f > s\}$ and let C be a connected component of $\{f \geq t\}$. Assume they intersect, i.e., $C \cap R \neq \emptyset$. First, assume that $s < t$. In that case $C \subseteq \{f > s\}$, and being connected, there is a unique connected component of $\{f > s\}$ that contains it, which is necessarily R. The reasoning is similar if $s \geq t$. Indeed, in that case $R \subseteq \{f \geq t\}$, and being connected, there is a unique connected component of $\{f \geq t\}$ that contains it, which is necessarily C. ■

Recall that a mode is simply a local maximum with strictly positive value, i.e., it is a point x such that $f(x) > 0$ and $f(x) \geq f(y)$ whenever $d(x, y) \leq r$ for some $r > 0$.

Lemma 22 *Consider a continuous function f with bounded upper level sets. Then each connected component of any of its upper level sets contains at least one mode.*

Proof Take $C \in \text{cc}(U_\lambda)$ for some $\lambda > 0$. Because f is continuous and U_λ is compact, C is compact, so that there is $x_0 \in C$ such that $f(x_0) = \max_C f$. Because the distance function is continuous,[6] and the fact that all connected components of U_λ are compact, we have that $d(C, C') := \min_{x \in C, x' \in C'} d(x, x') > 0$ for all $C' \in \text{cc}(U_\lambda)$, so that there is $\eta > 0$ such that $d(C, C') > \eta$ for all $C' \in \text{cc}(U_\lambda)$. Now, consider y within distance η of x_0. If $y \in C$, then $f(y) \leq \max_C f = f(x_0)$; and if $y \notin C$, then $y \notin U_\lambda$, and therefore $f(y) < \lambda \leq f(x_0)$. We can conclude that x_0 is a mode. ■

Lemma 23 *Consider a continuous function f with bounded upper level sets and locally finitely many modes. Then, f satisfies the following property:*

$$\text{For every } \lambda > 0, \text{ if } \epsilon > 0 \text{ is small enough, each connected component of } \{f > \lambda\} \text{ contains exactly one connected component of } \{f > \lambda + \epsilon\}. \tag{10}$$

6. As is well-known, $|d(x, y) - d(x', y')| \leq d(x, x') + d(y, y')$ by a simple use of the triangle inequality, so that $d : \Omega \times \Omega \to \mathbb{R}$ is Lipschitz and continuous when equipping $\Omega \times \Omega$ with the product topology.

Proof Take any upper level set U. Since U is bounded, it can only include a finite number of modes. And since each of its connected components contains at least one mode by Lemma 22, it must be the case that U has at most as many connected components as it contains modes.

We now assume that the upper level sets all have finitely many components, and show that (10) holds. We do so by contradiction. Therefore, assume that (10) does not hold so that there is $\lambda > 0$ and R a connected component of $\{f > \lambda\}$, and a sequence (ϵ_n), which we can take to be decreasing and converging to zero, such that for each n, R contains at least two connected components of $\{f \geq \lambda + \epsilon_n\}$. Because R is a bounded region, applying the first part of the statement we find that R can only contain finitely many components of $\{f > \lambda + \epsilon_n\}$, denoted $A_1^n, \dots, A_{m_n}^n$, with $2 \leq m_n \leq M$ for all n, where M is the number of modes within U_λ. By taking a subsequence if needed, we may further assume that $m_n = m \geq 2$ for all n. By the usual nesting property, at every n, for each i, there is exactly one j such that $A_i^n \subseteq A_j^{n+1}$, and so that we may choose the indexing in such a way that $A_i^n \subseteq A_i^{n+1}$ for all n and all i. This allows us to define $A_i = \bigcup_n A_i^n$ for $i = 1, \dots, m$. Since f is continuous, $\{f > \lambda + \epsilon_n\}$ is open, and therefore so are its connected components (since we assume throughout that Ω is locally connected), and therefore each A_i^n is open, which then carries over to each A_i being open. The A_i are disjoint because $A_i^n \cap A_{i'}^{n'} = \emptyset$ unless $i = i'$. Therefore, because $R = \bigcup_i A_i$, R must be disconnected — a contradiction. ■

Proposition 24 *Let f be a continuous density. Assume that $\mathcal{C}$ is a closed cluster tree such that $d_M(\mathcal{C}, \mathcal{H}_f) = 0$. Then $\mathcal{C}$ contains $\mathcal{H}_f$. If, in addition,* (10) *holds, then, for every $C \in \mathcal{C}$, $\{f > h(C)\} \cap C$ is some union of connected components of $\{f > h(C)\}$.*

Proof Let $V \in \mathrm{cc}(U_\lambda)$ for some $\lambda > 0$. Fix $x \in V$ such that $f(x) = \lambda$. Take $y \in V$. First, $m_f(x,y) = \lambda$, and since $m_{\mathcal{C}}(x,y) = m_f(x,y)$ and $\mathcal{C}$ is assumed closed, there is $C_y \in \mathcal{C}$ containing x, y such that $h(C_y) = \lambda$. Note that this implies that $C_y \subseteq V$ since V is the largest connected set that contains x, y such that $h(V) \geq \lambda$. If $y \neq z$ are both in V, we have that $x \in C_y \cap C_z$, so that C_y and C_z are nested. Therefore, the collection $\{C_y : y \in V\}$ is nested, and because $\mathcal{C}$ is closed, $C = \bigcup_{y \in V} C_y$ belongs to $\mathcal{C}$. Since $C_y \subseteq V$ for all y, we have $C \subseteq V$; and since C_y contains y for all y, we also have $C \supseteq V$; therefore, $C = V$, and we conclude that $V \in \mathcal{C}$.

For the second part, assume that (10) holds. Take $C \in \mathcal{C}$ with $\lambda = h(C) > 0$. We want to show that, if R is a connected component of $\{f > \lambda\}$ such that $R \cap C \neq \emptyset$, then $R \subseteq C$. For $\epsilon > 0$ small enough, R contains exactly one connected component of $\{f > \lambda + \epsilon\}$, which by way of Lemma 21 implies that R contains exactly one connected component of $\{f \geq \lambda + \epsilon\}$, which we denote by V_ϵ. By the first part of the proposition, which we have already established, V_ϵ belongs to $\mathcal{C}$, and $\mathcal{C}$ being a cluster tree, we have either $V_\epsilon \cap C = \emptyset$ or $V_\epsilon \subseteq C$. Only the latter is possible when ϵ is small enough. Indeed, take $x \in R \cap C$, so that $f(x) > \lambda$. Let $\epsilon > 0$ be small enough that $f(x) \geq \lambda + \epsilon$, so that $x \in V_\epsilon$. Hence, $V_\epsilon \subseteq C$ when $\epsilon > 0$ is small enough, and we then use the fact that $R = \bigcup_{\epsilon > 0} V_\epsilon$ to conclude that $R \subseteq C$. ■

We remark that, when f is 'flat nowhere' in the sense that

$$\overline{\{f > \lambda\}} = \{f \geq \lambda\} \text{ for any } \lambda > 0,$$

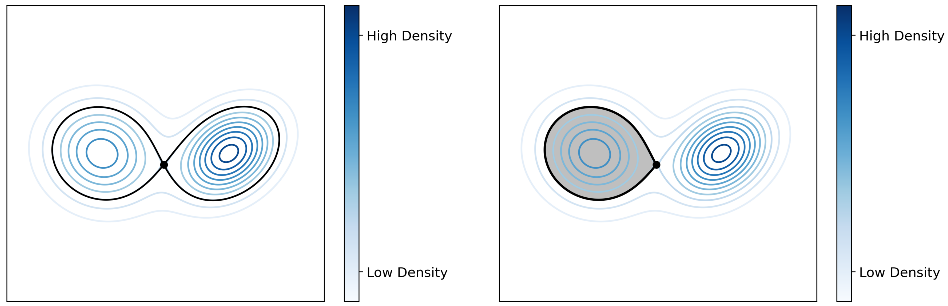

Figure 8: Left: The level sets of a bimodal Gaussian. The upper level set "splits" at the level containing the saddle point. (The level set and saddle point are highlighted.) Right: The highlighted cluster is $\overline{R_1}$. The addition of this cluster to $\mathcal{H}_f$ forms a valid and distinct cluster tree $\mathcal{C}_1$.

then, under the same conditions as in Proposition 24, any $C \in \mathcal{C}$ is closure of the union of connected components of $\{f > \lambda\}$. This still leaves the possibility that $\mathcal{C} \neq \mathcal{H}_f$, and it can indeed happen — unless f is unimodal. To see this, for simplicity, suppose that f is 'flat nowhere' and has exactly two modes. Assuming that its support has connected interior, there is exactly one level $\lambda > 0$ where the upper level set 'splits' in the sense that $\{f > \lambda\}$ has two connected components, say R_1 and R_2, while $\{f \geq \lambda\}$ is connected. Then, for $j \in \{1, 2\}$, $\overline{R_j}$ does not belong to $\mathcal{H}_f$ and $\mathcal{C}_j = \mathcal{H}_f \cup \{\overline{R_j}\}$ is a cluster tree satisfying $d_M(\mathcal{C}_j, \mathcal{H}_f) = 0$. (Note that $\mathcal{H}_f \cup \{\overline{R_1}, \overline{R_2}\}$ is not a cluster tree since $\overline{R_1}$ and $\overline{R_2}$ intersect but are not nested.) The situation is illustrated in Figure 8.

Appendix B. Euclidean Spaces

In this section we show that Euclidean spaces have the internally connected partition property by constructing a 'shifted' grid that has the required property.

Proposition 25 *Any Euclidean space has the internally connected partition property.*

Proof Consider the Euclidean space $\mathbb{R}^d$ (equipped with its Euclidean norm). It is enough to show that there is a a locally finite partition $\{A_i\}$ that has the internally connected property and is such that, for all i, $\text{int}(A_i)$ is connected and $\text{diam}(A_i) \leq \sqrt{d}$.

Let $\mathcal{L}_1 = \mathbb{Z}$, and for $d \geq 2$, define

$$\mathcal{L}_d = \Big\{(x_1, \dots, x_d) : x_d \in 2\mathbb{Z} \text{ and } (x_1, \dots, x_{d-1}) \in \mathcal{L}_{d-1}; \\ \text{or } x_d \in 2\mathbb{Z} + 1 \text{ and } (x_1 + \tfrac{1}{2}, \dots, x_{d-1} + \tfrac{1}{2}) \in \mathcal{L}_{d-1}\Big\}.$$

For $(x_1, \dots, x_d) \in \mathcal{L}_d$ define the corresponding cell

$$A_{(x_1,\dots,x_d)} = [x_1, x_1 + 1) \times \cdots \times [x_d, x_d + 1).$$

And consider the collection of these cells

$$\mathcal{A}_d = \big\{A_{(x_1,\dots,x_d)} : (x_1, \dots, x_d) \in \mathcal{L}_d\big\}.$$

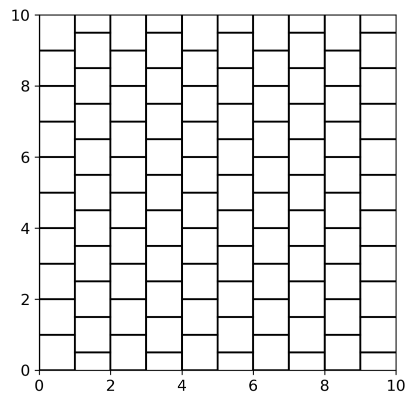

Figure 9: The existence of this shifted grid clearly shows that $\mathbb{R}^2$ has the internally connected partition property. This definition of a shifted grid can be extended to higher dimensions to show that $\mathbb{R}^d$ has the internally connected partition property for any $d \geq 2$. (This is trivially true in dimension $d = 1$ where a regular grid can be used to show that $\mathbb{R}$ has the internally connected partition property.)

Each of these cells has connected interior and has diameter $\sqrt{d}$. Moreover, $\mathcal{A}_d$ is a partition of the entire space $\mathbb{R}^d$. And, as a partition, $\mathcal{A}_d$ is clearly locally finite. The partition is depicted for $d = 2$ in Figure 9.

We now prove that $\mathcal{A}_d$ has the internally connected property. We will proceed by induction on d. For $d = 1$, this is clear. For $d \geq 2$, assume that $\mathcal{A}_{d-1}$ has the internally connected property. Consider $(x_1, \ldots, x_d)$ and $(y_1, \ldots, y_d)$, both in $\mathcal{L}_d$, such that $\overline{A_{(x_1,\ldots,x_d)}} \cup \overline{A_{(y_1,\ldots,y_d)}}$ is connected. We want to show that $\operatorname{int}\big(\overline{A_{(x_1,\ldots,x_d)}} \cup \overline{A_{(y_1,\ldots,y_d)}}\big)$ is connected too. By induction, $\{A_{(z_1,\ldots,z_d)} : z_d = x_d\}$ has the internally connected property, so that it is enough to consider a situation where $y_d \neq x_d$. Suppose, for example, that $y_d > x_d$. In that case, the fact that $\overline{A_{(x_1,\ldots,x_d)}} \cup \overline{A_{(y_1,\ldots,y_d)}}$ is connected implies that $y_d = x_d + 1$ and $y_i = x_i \pm \frac{1}{2}$ for $1 \leq i \leq d - 1$. Further,

$$\operatorname{int}(\overline{A_{(x_1,\ldots,x_d)}} \cup \overline{A_{(y_1,\ldots,y_d)}}) = \operatorname{int}(A_{(x_1,\ldots,x_d)}) \cup \operatorname{int}(A_{(y_1,\ldots,y_d)}) \cup C,$$

where

$$C = \Big\{(z_1, z_2, \ldots, z_{d-1}, x_d + 1) : x_i + \tfrac{1}{4} + \tfrac{1}{4}\operatorname{sign}(y_i - x_i) \leq z_i \leq x_i + \tfrac{3}{4} + \tfrac{1}{4}\operatorname{sign}(y_i - x_i)\Big\}.$$

The fact that $C \subseteq \partial A_{(x_1,\ldots,x_d)} \cap \partial A_{(y_1,\ldots,y_d)}$ proves that the union above is connected. ■

References

A. Anandkumar, D. Hsu, and S. M. Kakade. A method of moments for mixture models and hidden markov models. In *Conference on Learning Theory*, 2012.

E. Arias-Castro and W. Qiao. A unifying view of modal clustering. *Information and Inference: A Journal of the IMA*, 12(2):897–920, 2023.

M. A. Armstrong. *Basic Topology*. Undergraduate Texts in Mathematics. Springer New York, 1983.

D. Arthur and S. Vassilvitskii. k-means++: the advantages of careful seeding. In *ACM-SIAM Symposium on Discrete Algorithms*, pages 1027–1035. Society for Industrial and Applied Mathematics, 2007.

S. Balakrishnan, S. Narayanan, A. Rinaldo, A. Singh, and L. Wasserman. Cluster trees on manifolds. *Advances in Neural Information Processing Systems*, 26, 2013.

S. Ben-David and M. Ackerman. Measures of clustering quality: A working set of axioms for clustering. *Advances in Neural Information Processing Systems*, 21, 2008.

C. Bouveyron, G. Celeux, T. B. Murphy, and A. E. Raftery. *Model-Based Clustering and Classification for Data Science: With Applications in R*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019.

G. Carlsson and F. Mémoli. Characterization, stability and convergence of hierarchical clustering methods. *Journal of Machine Learning Research*, 11(47):1425–1470, 2010.

J. W. Carmichael, J. A. George, and R. S. Julius. Finding natural clusters. *Systematic Zoology*, 17(2):144–150, 1968.

J. E. Chacón. Mixture model modal clustering. *Advances in Data Analysis and Classification*, 13(2):379–404, 2019.

J. E. Chacón. The modal age of statistics. *International Statistical Review*, 88(1):122–141, 2020.

K. Chaudhuri, S. Dasgupta, S. Kpotufe, and U. von Luxburg. Consistent procedures for cluster tree estimation and pruning. *IEEE Transactions on Information Theory*, 60(12): 7900–7912, 2014.

V. Cohen-Addad, V. Kanade, and F. Mallmann-Trenn. Clustering redemption–beyond the impossibility of kleinberg's axioms. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

R. M. Cormack. A review of classification. *Journal of the Royal Statistical Society: Series A (General)*, 134(3):321–353, 1971.

S. Dasgupta. A cost function for similarity-based hierarchical clustering. In *ACM Symposium on Theory of Computing*, pages 118–127, 2016.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B*, 39(1):1–22, 1977.

J. Eldridge, M. Belkin, and Y. Wang. Beyond hartigan consistency: Merge distortion metric for hierarchical clustering. In *Conference on Learning Theory*, volume 40, pages 588–606. PMLR, 2015.

M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *International Conference on Knowledge Discovery and Data Mining*, pages 226–231. AAAI Press, 1996.

V. Estivill-Castro. Why so many clustering algorithms: a position paper. *ACM SIGKDD Explorations Newsletter*, 4(1):65–75, 2002.

B. Everitt. *Finite Mixture Distributions*. Springer Science & Business Media, 2013.

C. Fraley and A. E. Raftery. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97(458):611–631, 2002.

K. Fukunaga and L. Hostetler. The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Transactions on Information Theory*, 21(1): 32–40, 1975.

G. Gan, C. Ma, and J. Wu. *Data Clustering : Theory, Algorithms, and Applications*. Society for Industrial and Applied Mathematics, 2nd edition, 2021.

J. A. Hartigan. *Clustering Algorithms*. Wiley, 1975.

J. A. Hartigan. Distribution problems in clustering. In *Classification and Clustering*, pages 45–71. Academic Press, 1977.

J. A. Hartigan. Consistency of single linkage for high-density clusters. *Journal of the American Statistical Association*, 76(374):388–394, 1981.

J. A. Hartigan. Statistical theory in clustering. *Journal of Classification*, 2:63–76, 1985.

C. Hennig. What are the true clusters? *Pattern Recognition Letters*, 64:53–62, 2015. Philosophical Aspects of Pattern Recognition.

D. Hsu and S. M. Kakade. Learning mixtures of spherical gaussians: moment methods and spectral decompositions. In *Innovations in Theoretical Computer Science*, pages 11–20, 2013.

A. Q. Jaffe. Strong consistency for a class of adaptive clustering procedures. *arXiv preprint arXiv:2202.13423*, 2022.

C. Jardine, N. Jardine, and R. Sibson. The structure and construction of taxonomic hierarchies. *Mathematical Biosciences*, 1(2):173–179, 1967.

N. Jardine and R. Sibson. The construction of hierarchic and non-hierarchic classifications. *The Computer Journal*, 11(2):177–184, 1968.

J. Kim, Y.-C. Chen, S. Balakrishnan, A. Rinaldo, and L. Wasserman. Statistical inference for cluster trees. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

J. Kleinberg. An impossibility theorem for clustering. *Advances in Neural Information Processing Systems*, 15, 2002.

G. J. McLachlan and D. Peel. *Finite Mixture Models*. John Wiley & Sons, 2000.

G. J. McLachlan, S. X. Lee, and S. I. Rathnayake. Finite mixture models. *Annual Review of Statistics and its Application*, 6(1):355–378, 2019.

G. Menardi. A review on modal clustering. *International Statistical Review*, 84(3):413–433, 2016.

K. Pärna. Strong consistency of K-means clustering criterion in separable metric spaces. *Tartu Riikl. Ul. Toimetised*, 733:86–96, 1986.

K. Pärna. On the existence and weak convergence of K-centres in Banach spaces. *Tartu Ulikooli Toimetised*, 893:17–287, 1990.

D. Pollard. Strong consistency of k-means clustering. *The annals of statistics*, pages 135–140, 1981.

J. Puzicha, T. Hofmann, and J. M. Buhmann. A theory of proximity based clustering: Structure detection by optimization. *Pattern Recognition*, 33(4):617–634, 2000.

J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000.

R. Sibson. A model for taxonomy. ii. *Mathematical Biosciences*, 6:405–430, 1970.

I. Steinwart. Adaptive density level set clustering. In *Conference on Learning Theory*, volume 19, pages 703–738. PMLR, 2011.

F. Strazzeri and R. J. Sánchez-García. Possibility results for graph clustering: A novel consistency axiom. *Pattern Recognition*, 128:108687, 2022.

U. von Luxburg, R. C. Williamson, and I. Guyon. Clustering: Science or art? In *ICML Workshop on Unsupervised and Transfer Learning*, volume 27, pages 65–79. PMLR, 2012.

D. Wang, X. Lu, and A. Rinaldo. Dbscan: Optimal rates for density-based cluster estimation. *Journal of Machine Learning Research*, 20(170):1–50, 2019.

R. B. Zadeh and S. Ben-David. A uniqueness theorem for clustering. In *Conference on Uncertainty in Artificial Intelligence*, pages 639–646, 2009.