Statistical Science
2015, Vol. 30, No. 4, 518–532
DOI: 10.1214/15-STS526

A Population Background for Nonparametric Density-Based Clustering

José E. Chacón

Abstract. Despite its popularity, it is widely recognized that the investigation of some theoretical aspects of clustering has been relatively sparse. One of the main reasons for this lack of theoretical results is surely the fact that, whereas for other statistical problems the theoretical population goal is clearly defined (as in regression or classification), for some of the clustering methodologies it is difficult to specify the population goal to which the data-based clustering algorithms should try to get close. This paper aims to provide some insight into the theoretical foundations of clustering by focusing on two main objectives: to provide an explicit formulation for the ideal population goal of the modal clustering methodology, which understands clusters as regions of high density; and to present two new loss functions, applicable in fact to any clustering methodology, to evaluate the performance of a data-based clustering algorithm with respect to the ideal population goal. In particular, it is shown that only mild conditions on a sequence of density estimators are needed to ensure that the sequence of modal clusterings that they induce is consistent.

Key words and phrases: Clustering consistency, distance in measure, Hausdorff distance, modal clustering, Morse theory.

1. INTRODUCTION

Clustering is one of the branches of Statistics with more research activity in recent years. As noted by Meilă (2007), "clustering is a young domain of research, where rigorous methodology is still striving to emerge." Indeed, some authors have recently expressed their concerns about the lack of theoretical or formal developments for clustering, as, for instance, von Luxburg and Ben-David (2005), Ben-David, von Luxburg and Pál (2006), Ackerman and Ben-David (2009), Zadeh and Ben-David (2009). This paper aims to contribute to this regularization (or, say, rigorousization).

Stated in its most simple form, cluster analysis consists in "partitioning a data set into groups so that the points in one group are similar to each other and are as different as possible from the points in other groups" (Hand, Mannila and Smyth, 2001, page 293). Posed as such, the problem does not even seem to have a statistical meaning. In fact, in concordance with Li, Ray and Lindsay (2007), it is possible to roughly classify clustering methods into three categories, depending on the amount of statistical information that they involve. These categories are very basically depicted in the following three paragraphs.

Some clustering techniques are solely based on the distances between the observations. Close observations are joined together to form a group, and extending the notion of inter-point distance to distance between groups, the resulting groups are gradually merged until all the initial observations are contained into a single group. This represents, of course, the notion of agglomerative *hierarchical clustering* (Izenman, 2008, Section 12.3). The graphical outcome depicting the successive agglomeration of data points up to a single group is the well-known dendrogram, and depending on the notion of inter-group distance used along the merging process, the most common procedures of this type are known as single linkage, complete linkage or average linkage (see also Hastie, Tibshirani and Friedman, 2009, page 523).

José E. Chacón is Profesor Titular (Associate Professor), Departamento de Matemáticas, Universidad de Extremadura, 06006 Badajoz, Spain (e-mail: jechacon@unex.es).

A first statistical flavor is noticed when dealing with those clustering methodologies that represent each cluster by a central point, such as the mean, the median or, more generally, a trimmed mean. This class of techniques is usually referred to as *partitioning methods*, and surely the most popular of its representatives is K-means (MacQueen, 1967). For a prespecified number K of groups, these algorithms seek for K centers with the goal of optimizing a certain score function representing the quality of the clustering (Everitt et al., 2011, Chapter 5).

When a more extended set of features of the data-generating probability distribution is used to determine the clustering procedure, it is usual to refer to these techniques as distribution-based clustering or, for the common case of continuous distributions, as *density-based clustering*. This approach is strongly supported by some authors, like Carlsson and Mémoli (2013), who explicitly state that "density needs to be incorporated in the clustering procedures."

As with all the statistical procedures, there exist parametric and nonparametric methodologies for density-based clustering. Surely the gold standard of parametric density-based clustering is achieved through mixture modeling, as clearly described in Fraley and Raftery (2002). It is assumed that the distribution generating the data is a mixture of simple parametric distributions, for example, multivariate normal distributions, and each component of the mixture is associated to a different population cluster. Maximum likelihood is used to fit a mixture model and then each data point is assigned to the most likely component using the Bayes rule.

The nonparametric methodology is based on identifying clusters as regions of high density separated from each other by regions of lower density (Wishart, 1969, Hartigan, 1975). Thus, a cluster is seen as a zone of concentration of probability mass. In this sense, population clusters are naturally associated with the modes (i.e., local maxima) of the probability density function, and this nonparametric approach is denominated mode-based clustering or *modal clustering* (Li, Ray and Lindsay, 2007). Precisely, each cluster is usually understood as the "domain of attraction" of a mode (Stuetzle, 2003).

The concept of domain of attraction is not that simple to specify, and providing a precise definition for that is one of the main goals of this paper. The first attempt to make the goal of modal clustering precise was introduced through the notion of level sets (Hartigan, 1975). If the distribution of the data has a density f, given $\lambda \geq 0$, the λ-level set of f is defined as $L(\lambda) = \{\mathbf{x}: f(\mathbf{x}) \geq \lambda\}$. Then, population λ-clusters are defined as the connected components of $L(\lambda)$, a definition that clearly captures the notion of groups having a high density. An extensive account of the usefulness of level sets in applications is given in Mason and Polonik (2009).

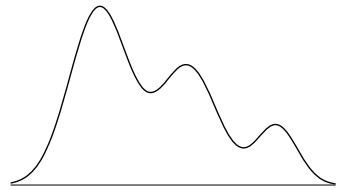

FIG. 1. *Univariate trimodal density for which it is not possible to capture its whole cluster structure using a level set analysis based on a single level.*

One of the advantages of clustering based on level sets is that the population target is clearly identified (the connected components of the λ-level set). However, the main drawback of this approach is perhaps the fact that the notion of population cluster depends on the level λ, as recognized by Stuetzle (2003). Nevertheless, other authors, like Cuevas, Febrero and Fraiman (2001) or Cadre, Pelletier and Pudlo (2013), affirm that the choice of λ is only a matter of resolution level of the analysis.

Still, it is easy to think of many examples in which it is impossible to observe the whole cluster structure on the basis of a single level λ. Essentially as in Rinaldo et al. (2012), page 906, Figure 1 shows a simple univariate example of this phenomenon: three different modal groups are visually identifiable, yet none of the level sets of the density has three connected components. To amend this, the usual recommendation is to analyze the cluster structure for several values of the level λ. Graphical tools oriented to this goal are the cluster tree (Stuetzle, 2003) or the mode function (Azzalini and Torelli, 2007, Menardi and Azzalini, 2014). Both graphics are useful to show how the clusters emerge as a function of λ. See Section 3 for a more detailed explanation.

Finally, the idea of examining the evolution of the cluster structure as the density level varies is closely related with the topic of persistent homology, a tool from Computational Topology that, since its relatively recent introduction, has attracted a great deal of interest for its applications in Topological Data Analysis; see Edelsbrunner and Harer (2008), Carlsson (2009) or Chazal et al. (2013). This tool allows to quantify which topological aspects of an object are most persistent as

the resolution level evolves, thus leading to the identification of the most important features of the object. In the context of data-based clustering based on level sets, it can be very useful to distinguish which of the discovered clusters are real and which of them are spurious (Fasy et al., 2014).

The rest of this paper is structured as follows: in Section 2 we introduce the concept of whole-space clustering as the type of object of interest in cluster analysis, and we point out the difference with the more usual notion of a clustering of the data. Later, it is explained that the population whole-space clustering depends on the adopted definition of cluster for each of the clustering methodologies. Section 3 expands on the first main contribution of the paper by providing a precise definition of the population goal of modal clustering, making use of Morse theory, leading to an equivalent yet simpler formulation (in a sense) as with the cluster tree. Once a population background for clustering has been set up, Section 4 contains the second main contribution of the paper, a proposal of two new loss functions to measure the similarity of two whole-space clusterings. These distance functions are not limited to modal clustering nor even to density-based clustering, they are applicable to any clustering methodology having a clearly identified population goal. As such, they can be used to define a notion of *clustering consistency*, and for the particular case of modal clustering it is shown that mild conditions are needed so that the data-based clustering constructed from a sequence of density estimators is consistent in this sense.

2. POPULATION CLUSTERINGS

Many different notions of cluster are possible, but no matter which one is used, it is necessary to have a clear idea of the type of object that clustering methods pursue from a population point of view. That object will be called a clustering.

Since the empirical formulation of the clustering task comprises partitioning a data set into groups, it suggests that its population analogue should involve a partition of the whole space or, at least, of the support of the distribution. Hence, a clustering of a probability distribution P on $\mathbb{R}^d$, or a *whole-space P-clustering*, should be understood as an essential partition of $\mathbb{R}^d$ into mutually disjoint measurable components, each with positive probability content (Ben-David, von Luxburg and Pál, 2006). More specifically, a whole-space P-clustering (or, simply, a clustering) is defined as a class of measurable sets $\mathcal{C} = \{C_1, \dots, C_r\}$ such that:

1. $P(C_i) > 0$ for all $i = 1, \dots, r$,
2. $P(C_i \cap C_j) = 0$ for $i \neq j$, and
3. $P(C_1 \cup \cdots \cup C_r) = 1$.

The components $C_1, \dots, C_r$ of such a partition are called clusters. Thus, two clusterings $\mathcal{C}$ and $\mathcal{D}$ are identified to be the same if they have the same number of clusters and, up to a permutation of the cluster labels, every cluster in $\mathcal{C}$ and its most similar match in $\mathcal{D}$ differ in a null-probability set (more details on this are elaborated in Section 4).

At this point it is worth distinguishing between two different, although closely related, concepts. When the probability distribution P is unknown, and a sample drawn from P is given, any procedure to obtain a data-based (essential) partition $\widehat{\mathcal{C}} = \{\widehat{C}_1, \dots, \widehat{C}_r\}$ will be called a *data-based clustering*. This simply means that $\int_{\widehat{C}_i} dP > 0$ for all $i = 1, \dots, r$, $\int_{\widehat{C}_i \cap \widehat{C}_j} dP = 0$ for $i \neq j$ and $\int_{\widehat{C}_1 \cup \cdots \cup \widehat{C}_r} dP = 1$. However, when data are available most clustering procedures focus on partitioning the data set, and, indeed, many of them do not even induce a clustering of the probability distribution. This will be referred to henceforth as a *clustering of the data*. Notice that, clearly, any data-based clustering $\widehat{\mathcal{C}} = \{\widehat{C}_1, \dots, \widehat{C}_r\}$ immediately results in a clustering of the data, by assigning the same group to data points belonging to the same component in $\widehat{\mathcal{C}}$.

2.1 The Ideal Population Clustering

The definition of (whole-space) clustering represents the type of population object that clustering methods should try to get close in general, but it is the particular employed notion of cluster that makes the theoretical goal of clustering methodologies change, focusing on different concepts of *ideal population clustering*.

For some clustering techniques, this ideal population clustering is well established. For instance, it is well known that the population clustering induced by the optimal set of K-means is a Voronoi tessellation. To be precise, let $\boldsymbol{\mu}_1^*, \dots, \boldsymbol{\mu}_K^* \in \mathbb{R}^d$ be a solution to the population K-means problem, in the sense that they minimize

$$R(\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_K) = \int \min_{k=1,\dots,K} \|\mathbf{x} - \boldsymbol{\mu}_k\| \, dP(\mathbf{x}),$$

where $\|\cdot\|$ denotes the usual Euclidean norm. Then, the K-means algorithm assigns an arbitrary point in $\mathbb{R}^d$ to the group whose center is closer, so that the ideal population clustering is given by $\mathcal{C} = \{C_1, \dots, C_K\}$, where

$$C_k = \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x} - \boldsymbol{\mu}_k^*\| \leq \|\mathbf{x} - \boldsymbol{\mu}_j^*\| \text{ for all } j \neq k\}$$

is the Voronoi cell corresponding to $\boldsymbol{\mu}_k^*$, for $k = 1, \ldots, K$ (see Graf and Luschgy, 2000, Chapter 4).

The ideal population clustering for mixture model clustering can be derived in a similar way. Assume that the underlying density is a mixture $f(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \cdot f_k(\mathbf{x})$, where π_k denotes the prior probability of the kth mixture component (with $\pi_k > 0$ and $\sum_{k=1}^{K} \pi_k = 1$), and $f_k(\mathbf{x})$ is the density of the kth component. In this setup, assuming also that the mixture model is identifiable, a point $\mathbf{x} \in \mathbb{R}^d$ is assigned to the group k for which the a posteriori probability $\pi_k f_k(\mathbf{x})/f(\mathbf{x})$ is maximum, so the ideal population clustering that f induces has population clusters

$$C_k = \{\mathbf{x} \in \mathbb{R}^d : \pi_k f_k(\mathbf{x}) \geq \pi_j f_j(\mathbf{x}) \text{ for all } j \neq k\}$$

for $k = 1, \ldots, K$.

For the modal approach to clustering, however, the notion of ideal population clustering is not so straightforward to formulate. Informally, if the data-generating density f has modes $\mathbf{M}_1, \ldots, \mathbf{M}_K$, then the population cluster C_k is defined as the domain of attraction of $\mathbf{M}_k$, for $k = 1, \ldots, K$. Most modal clustering algorithms are based on applying a mode-seeking numerical method to the sample points and assigning the same cluster to those data that are iteratively shifted to the same limit value. Examples of such procedures include the mean shift algorithm (Fukunaga and Hostetler, 1975), CLUES (Wang, Qiu and Zamar, 2007) or the modal EM of Li, Ray and Lindsay (2007), and further alternatives are described in a previous unpublished version of this paper (Chacón, 2012). Hence, from a practical point of view, it is clear how a clustering of the data is constructed on the basis of this notion of domain of attraction. The objective of the next section is to describe in a precise way what is the population goal that lies behind these algorithms. This aims to provide an answer, in the case of modal clustering, to Question 1 in von Luxburg and Ben-David (2005): "How does a desirable clustering look if we have complete knowledge about our data generating process?"

3. DESCRIBING THE POPULATION GOAL OF MODAL CLUSTERING THROUGH MORSE THEORY

The ideal population goal for modal clustering should reflect the notion of a partition into regions of high density separated from each other by regions of lower density. The following examples in one and two dimensions are useful to illustrate the concept that we aim to formalize.

In the one-dimensional case, it seems clear from Figure 2 how this can be achieved. To begin with, the level set methodology identifies the three clusters in the density depicted in Figure 1 by computing the cluster tree as described clearly in Nugent and Stuetzle (2010): starting from the 0-level set, which corresponds to the whole real line in this example (hence, it consists of a single connected component), λ is increased until it reaches λ_1, where two components for the λ_1-level set are found, G_1' and G_2', resulting in the

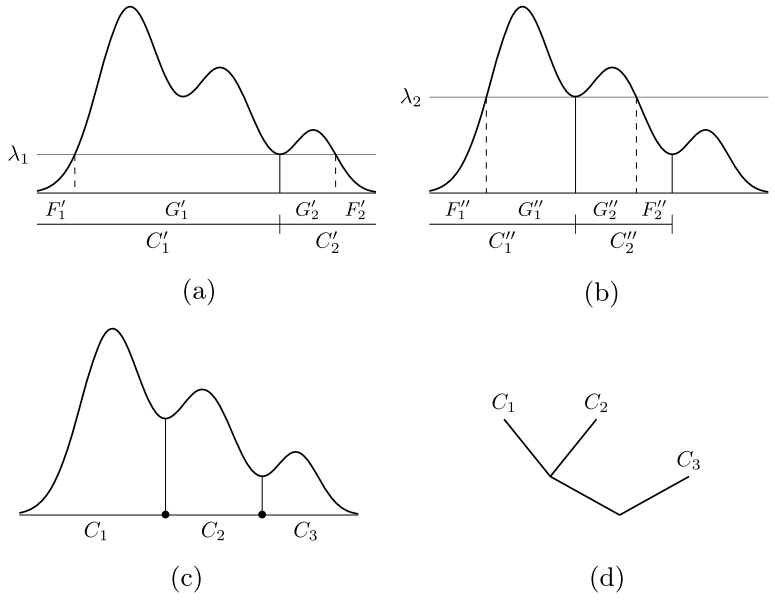

FIG. 2. *Identification of clusters for the trimodal density example using the cluster tree. Panel* (*a*): *first split*; (*b*) *second split*; (*c*) *final partition*; (*d*) *cluster tree*.

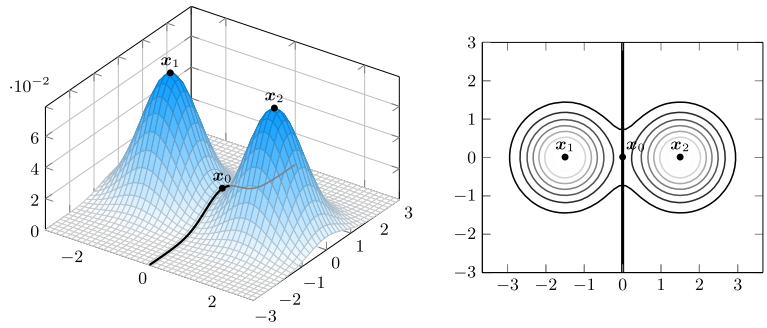

FIG. 3. *Bidimensional example, with two groups clearly identifiable at an intuitive level.*

cluster tree splitting into two different branches [see Figure 2, panel (a)]. These two components G_1' and G_2' are usually called cluster cores. They do not constitute a clustering because there is some probability mass outside $G_1' \cup G_2'$. But the remaining parts F_1' and F_2', referred to as fluff in Nugent and Stuetzle (2010), can be assigned to either the left or the right branch depending on whichever of them is closer. Thus, at level λ_1 the partition $\mathbb{R} = C_1' \cup C_2'$ is obtained. The point dividing the line into these two components can be arbitrarily assigned to either of them; this assignment makes no difference because it leads to equivalent clusterings since a singleton has null probability mass.

At level λ_2 the left branch C_1' is further divided into two branches [see panel (b) of Figure 2]. Again, the two cluster core components G_1'' and G_2'' do not form a partition of the set C_1' associated with the previous node of the tree, but it is clear how the fluff F_1'' and F_2'' can be assigned to form a partition $C_1'' \cup C_2''$ of C_1'. Since no further splitting of the cluster tree is observed as λ increases, the final population clustering is $\{C_1'', C_2'', C_2'\}$, renamed to $\{C_1, C_2, C_3\}$ in panel (c) of Figure 2.

It is immediate to observe that the levels at which a connected component breaks into two different ones correspond precisely to local minima of the density function, so an equivalent formulation consists of defining population clusters as the connected components of $\mathbb{R}$ minus the points where a local minimum is attained [the solid circles in panel (c) of Figure 2]. Notice that, unlike the cluster tree, this definition does not involve the computation of level sets for a range of levels, nor their cores and fluff, and in this sense it constitutes a more straightforward approach to the very same concept in the unidimensional setup.

To get an idea of how to generalize the previous approach to higher dimensions, consider the following extremely simple bidimensional example: an equal-proportion mixture of two normal distributions, each with identity variance matrix and centered at $\boldsymbol{\mu}_1 = (-\frac{3}{2}, 0)$ and $\boldsymbol{\mu}_2 = -\boldsymbol{\mu}_1$, respectively. At an intuitive level, it is clear from Figure 3 that the most natural border to separate the two visible groups is the black line. The problem is then: what is exactly that line? Is it identifiable in terms of the features of the density function in a precise, unequivocal way? A nice way to answer these questions is by means of Morse theory.

Morse theory is a branch of Differential Topology that provides tools for analyzing the topology of a manifold $M \subseteq \mathbb{R}^d$ by studying the critical points of a smooth enough function $f: M \to \mathbb{R}$. A classical reference book on this subject is Milnor (1963) and enjoyable introductions to the topic can be found in Matsumoto (2002) and Jost (2011), Chapter 7. A useful application of Morse theory is for terrain analysis, as nicely developed in Vitalli (2010). In terrain analysis, a mountain range can be regarded as the graph of a function $f: M \to \mathbb{R}$, representing the terrain elevation, over a terrain $M \subseteq \mathbb{R}^2$, just as in the left graphic of Figure 3. The goal of terrain analysis is to provide a partition of M through watersheds indicating the different regions, or catchment basins, where water flows under the effect of gravity.

The fundamentals of Morse theory can be extremely summarized as follows. A smooth enough function $f: M \to \mathbb{R}$ is called a *Morse function* if all its critical points are nondegenerate. Precisely, for our purposes, f can be considered smooth enough if it is three times continuously differentiable. Here, the critical points of f are understood as those $\mathbf{x}_0 \in M$ for which the gradient $\mathsf{D} f(\mathbf{x}_0)$ is null, and nondegeneracy means that the determinant of the Hessian matrix $\mathsf{H} f(\mathbf{x}_0)$ is not zero. For such points the *Morse index* $m(\mathbf{x}_0)$ is defined as the number of negative eigenvalues of $\mathsf{H} f(\mathbf{x}_0)$.

Morse functions can be expressed in a fairly simple form in a neighborhood of a critical point $\mathbf{x}_0$, as the result known as Morse lemma shows that it is possible to find local coordinates $x_1, \ldots, x_n$ such that f can be written as $f(\mathbf{x}_0) \pm x_1^2 \pm \cdots \pm x_d^2$ around $\mathbf{x}_0$, where

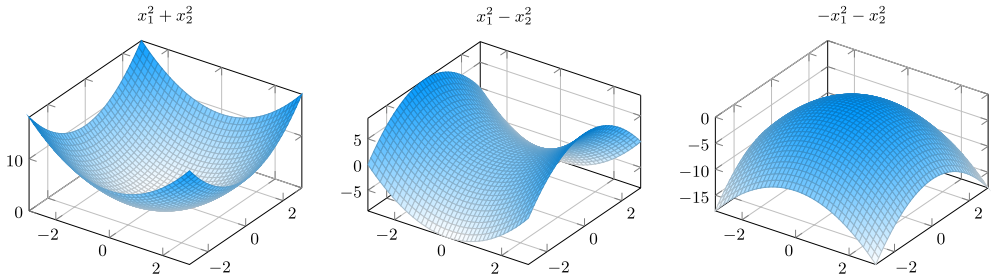

FIG. 4. *The three possible configurations around a critical point of a Morse function in the bidimensional case.*

the number of minus signs in the previous expression is precisely $m(\mathbf{x}_0)$. For example, for $d = 2$ the three possible configurations for a critical point are shown in Figure 4, corresponding to a local minimum, a saddle point and a local maximum (from left to right), with Morse indexes 0, 1 and 2, respectively.

The decomposition of M suggested by Morse theory is made in terms of the unstable and/or stable manifolds of the critical points of f as explained next. Consider the initial value problem defined by the minus gradient vector of a smooth enough function f. For a given value of $\mathbf{x} \in M$ at time $t = 0$, the integral curve $\boldsymbol{\nu}_{\mathbf{x}}: \mathbb{R} \to M$ of such an initial value problem is the one satisfying

$$\boldsymbol{\nu}'_{\mathbf{x}}(t) = -\mathsf{D}f\big(\boldsymbol{\nu}_{\mathbf{x}}(t)\big), \quad \boldsymbol{\nu}_{\mathbf{x}}(0) = \mathbf{x} \tag{3.1}$$

and the set of all these integral curves is usually referred to as the negative gradient flow. Since the minus gradient vector defines the direction of steepest descent of f, these curves (or, properly speaking, their images through f) represent the trajectories of the water flow subject to gravity.

With respect to the negative gradient flow, the *unstable manifold* of a critical point $\mathbf{x}_0$ is defined as the set of points whose integral curve starts at $\mathbf{x}_0$, that is,

$$W^u_-(\mathbf{x}_0) = \Big\{\mathbf{x} \in M: \lim_{t \to -\infty} \boldsymbol{\nu}_{\mathbf{x}}(t) = \mathbf{x}_0\Big\}.$$

Analogously, the stable manifold of $\mathbf{x}_0$ is the set of points whose integral curve finishes at $\mathbf{x}_0$, that is, $W^s_-(\mathbf{x}_0) = \{\mathbf{x} \in M: \lim_{t \to +\infty} \boldsymbol{\nu}_{\mathbf{x}}(t) = \mathbf{x}_0\}$. It was first noted by Thom (1949) that the class formed by the unstable manifolds corresponding to all the critical points of f provides a partition of M (the same is true for the stable manifolds). Furthermore, the unstable manifold $W^u_-(\mathbf{x}_0)$ has dimension $m(\mathbf{x}_0)$.

The main contribution of this section is the definition of the population modal clusters of a density f as the unstable manifolds of the negative gradient flow corresponding to local maxima of f. That is, if $\mathbf{M}_1, \ldots, \mathbf{M}_K$ denote the modes of f, then the ideal population goal for modal clustering is $\mathcal{C} = \{C_1, \ldots, C_K\}$, where $C_k = W^u_-(\mathbf{M}_k)$, for $k = 1, \ldots, K$. Or in a more prosaic way, in terms of water flows, a modal cluster is just the region of the terrain that would be flooded by a fountain emanating from a peak of the mountain range.

Although this is an admittedly cumbersome definition, going back to Figure 3, it is clear that it just describes the notion that we were looking for. The critical point $\mathbf{x}_0 = (0, 0)$ is a saddle point, thus having Morse index 1, and the black line is precisely its associated unstable manifold, $W^u_-(\mathbf{x}_0) = \{0\} \times \mathbb{R}$, which is a manifold of dimension 1. The remaining two critical points are local maxima, and their respective unstable manifolds are $W^u_-(\mathbf{x}_1) = (-\infty, 0) \times \mathbb{R}$ and $W^u_-(\mathbf{x}_2) = (0, \infty) \times \mathbb{R}$, manifolds of dimension 2 so that we can partition $\mathbb{R}^2 = W^u_-(\mathbf{x}_0) \cup W^u_-(\mathbf{x}_1) \cup W^u_-(\mathbf{x}_2)$, showing $W^u_-(\mathbf{x}_1)$ and $W^u_-(\mathbf{x}_2)$ as two population clusters separated by the border $W^u_-(\mathbf{x}_0)$, which is a null-probability set.

Notice that this definition also applies to the previous univariate example in Figure 2: the clusters C_1, C_2 and C_3 are just the unstable manifolds of the three local maxima (they are manifolds of dimension 1), and for the two local minima their unstable manifolds have dimension 0, so they include only the respective points of local minima.

Moreover, if we focus on the gradient flow, instead of the negative gradient flow, then its integral curves satisfy

$$\boldsymbol{\gamma}'_{\mathbf{x}}(t) = \mathsf{D}f\big(\boldsymbol{\gamma}_{\mathbf{x}}(t)\big), \quad \boldsymbol{\gamma}_{\mathbf{x}}(0) = \mathbf{x};$$

the unstable manifold for the negative gradient flow becomes the stable manifold for the gradient flow and viceversa. Therefore, we could equivalently define the cluster associated to a mode $\mathbf{x}_0$ of the density as its stable manifold with respect to the gradient flow, that is, $W^s_+(\mathbf{x}_0) = \{\mathbf{x} \in M: \lim_{t \to \infty} \boldsymbol{\gamma}_{\mathbf{x}}(t) = \mathbf{x}_0\} = W^u_-(\mathbf{x}_0)$. This is a precise formulation of the notion of domain of attraction of the mode $\mathbf{x}_0$, since $W^s_+(\mathbf{x}_0)$ represents the set of all the points that climb to $\mathbf{x}_0$ when they fol-

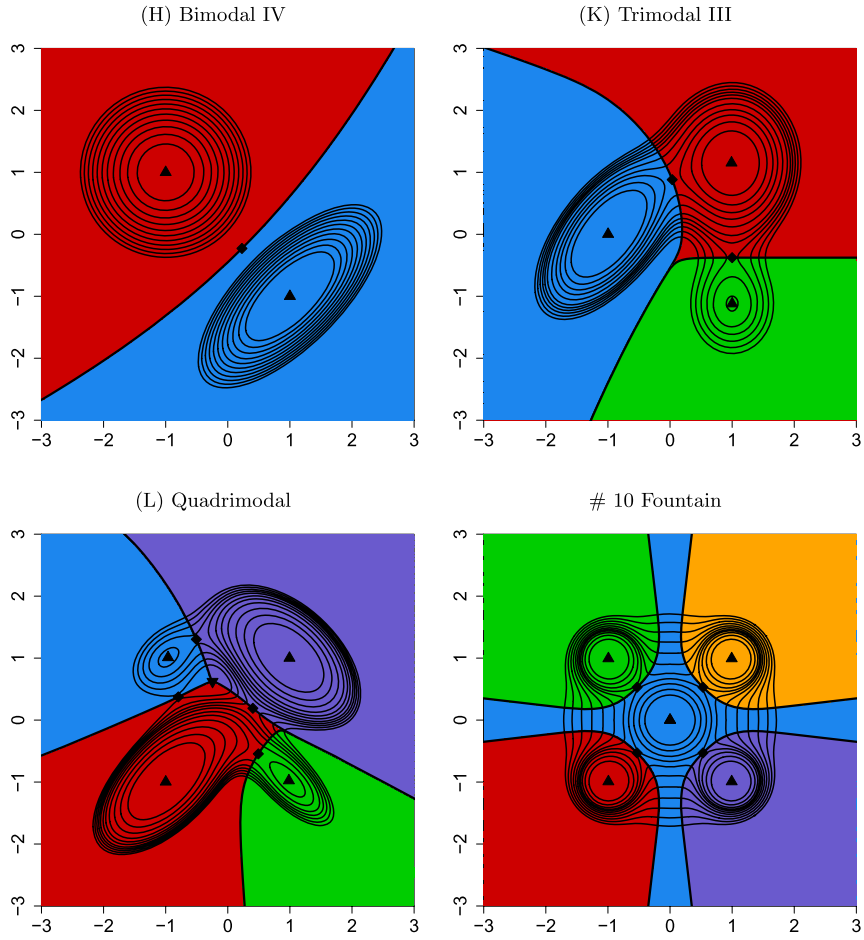

FIG. 5. *Ideal modal population clustering for some normal mixtures densities.*

low the steepest ascent path defined by the gradient direction. Moreover, estimating this path is precisely the goal of the mean shift algorithm (see Arias-Castro, Mason and Pelletier, 2013).

3.1 Examples

In Figure 5 we give further examples of how the ideal population goal of modal clustering looks for three of the bivariate normal mixture densities included in Wand and Jones (1993), namely, with their terminology, densities (H) Bimodal IV, (K) Trimodal III and (L) Quadrimodal, plus the normal mixture #10 Fountain from Chacón (2009). These densities have a number of modes ranging from two to five, respectively, and hence that is the true number of population clusters for each of these models, in the sense of modal clustering.

Each graph contains a contour plot of the density function; the location of the modes is marked with a triangle pointing upward (▲), the saddle points with a rotated square (⧫), and the only local minimum, appearing in the plot of the Quadrimodal density, is marked with a triangle pointing downward (▼). The thick lines passing through the saddle points are their corresponding unstable manifolds and represent the border between the different population clusters.

All these features have been computed numerically, making use of some results from the thorough analysis of normal mixture densities given in Ray and Lindsay (2005). For instance, the Newton–Raphson method has been used for the location of the modes by finding a zero gradient point starting from the component means, taking into account that both the location of the modes and component means are different, but very close. Next, the saddle points are searched along the ridgeline that connects every two component means, since all the critical points of the density must lie on this curve, by Theorem 1 in Ray and Lindsay (2005). Finally, the borders between the population clusters are obtained by numerically solving the initial value problem (3.1), starting from a point slightly shifted from each saddle point, along the direction of the eigenvector of its Hessian corresponding to a negative eigenvalue.

4. COMPARING CLUSTERINGS

Whatever the notion of ideal population clustering the researcher may use, in practice, this population goal has to be approximated from the data. Therefore, to evaluate the performance of a clustering method, it is necessary to introduce a loss function to measure the distance between a data-based clustering and the population goal or, more generally, to have a notion of distance between two whole-space clusterings. In this section, two proposals are derived by extending two well-known notions of distance between sets to distances between clusterings.

Recall that some clustering methods do not produce a partition of the whole feature space, but only a clustering of the data. A good deal of measures to evaluate the distance between two clusterings of the data have been proposed in the literature. The work of Meilă (2007) provides both a comprehensive survey of the most used existing measures as well as a deep technical study of their main properties, and, for instance, Arabie and Boorman (1973) or Day (1980/81) include further alternatives. But it should be stressed that all these proposals concern only partitions of a finite set. Here, on the contrary, our interest lies on developing two new notions of distance between whole-space clusterings.

Let $\mathcal{C}$ and $\mathcal{D}$ be two clusterings of a probability distribution P, and assume for the moment that both have the same number of clusters, say, $\mathcal{C} = \{C_1, \ldots, C_r\}$ and $\mathcal{D} = \{D_1, \ldots, D_r\}$. The first step to introduce a distance between $\mathcal{C}$ and $\mathcal{D}$ is to consider a distance between sets. Surely the two distances between sets most used in practice are the Hausdorff distance and the distance in measure; see Cuevas and Fraiman (2010). The Hausdorff distance is specially useful when dealing with compact sets (it defines a metric in the space of all compact sets of a metric space), as it tries to capture the notion of physical proximity between two sets (Rodríguez-Casal, 2003). In contrast, given a measure μ, the distance in μ-measure between two sets C and D refers to $\mu(C \triangle D)$, that is, to the content of their symmetric difference $C \triangle D = (C \cap D^c) \cup (C^c \cap D)$. It defines a metric on the set of all measurable subsets of a measure space, once two sets differing in a null-measure set are identified to be the same.

4.1 A Distance in Measure Between Clusterings

Although we will return to the Hausdorff distance later, our first approach to the notion of distance between $\mathcal{C}$ and $\mathcal{D}$ relies primarily on the concept of distance in μ-measure, and the measure involved is precisely the probability measure P. From a practical point of view, it does not seem so important that the clusters of a data-based partition get physically close to those of the ideal clustering. Instead, it is desirable that the points that are incorrectly assigned do not represent a very significant portion of the distribution. This corresponds to the idea of perceiving two clusters $C \in \mathcal{C}$ and $D \in \mathcal{D}$ (resulting from different clusterings) as close when $P(C \triangle D)$ is low. In this sense, the closeness between C and D is quantified by their distance in μ-measure for the particular choice $\mu = P$.

Therefore, for two clusterings $\mathcal{C}$ and $\mathcal{D}$ with the same number of clusters, a sensible notion of distance is obtained by adding up the contributions of the pairwise distances between their components once they have been relabeled, so that every cluster in $\mathcal{C}$ is compared with its most similar counterpart in $\mathcal{D}$. In mathematical terms, the distance between $\mathcal{C}$ and $\mathcal{D}$ can be measured by

$$d_1(\mathcal{C}, \mathcal{D}) = \min_{\sigma \in \mathcal{P}_r} \sum_{i=1}^{r} P(C_i \triangle D_{\sigma(i)}), \tag{4.1}$$

where $\mathcal{P}_r$ denotes the set of permutations of $\{1, 2, \ldots, r\}$.

It can be shown that d_1 defines a metric in the space of all the partitions with the same number of components, once two such partitions are identified to be the same if they differ only in a relabeling of their components. Moreover, the minimization problem in (4.1) is usually known as the *linear sum assignment problem* in the literature of Combinatorial Optimization, and it represents a particular case of the well-known Monge–Kantorovich transportation problem. A comprehensive treatment of assignment problems can be found in Burkard, Dell'Amico and Martello (2009).

If a partition is understood as a vector in the product space of measurable sets, with the components as its coordinates, then d_1 resembles the L_1 product distance, only adapted to take into account the possibility of relabeling the components. This seems a logical choice given the additive nature of measures, as it adds up the contribution of each distance between the partition components as described before. However, it would be equally possible to consider any other L_p distance, leading to define

$$d_p(\mathcal{C}, \mathcal{D}) = \min_{\sigma \in \mathcal{P}_r} \left\{ \sum_{i=1}^{r} P(C_i \triangle D_{\sigma(i)})^p \right\}^{1/p}$$

for $p \geq 1$ and also $d_\infty(\mathcal{C}, \mathcal{D}) = \min_{\sigma \in \mathcal{P}_r} \max\{P(C_i \cdot \triangle D_{\sigma(i)}) : i = 1, \ldots, r\}$. The minimization problem defining d_∞ is also well known under the name of the

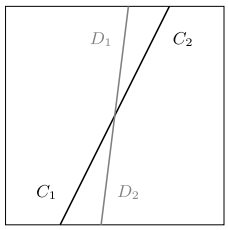

FIG. 6. *When computing the distance d_1 between the two clusterings $\mathcal{C} = \{C_1, C_2\}$ (black) and $\mathcal{D} = \{D_1, D_2\}$ (grey), it is found that $C_1 \cap D_1^c = C_2^c \cap D_2$ and $C_1^c \cap D_1 = C_2 \cap D_2^c$, so the content of each of these two discrepancy regions is added twice in $d_1(\mathcal{C}, \mathcal{D})$.*

linear bottleneck assignment problem, and its objective function is usually employed if the interest is to minimize the latest completion time in parallel computing (see Burkard, Dell'Amico and Martello, 2009, Section 6.2). Still, in the context of clustering, surely the d_1 distance seems the most natural choice among all the d_p possibilities, due to its clear interpretation.

Nevertheless, the definition of the d_1 distance involves some kind of redundancy, due to the fact that $\mathcal{C}$ and $\mathcal{D}$ are (essential) partitions of $\mathbb{R}^d$, because the two disjoint sets that form every symmetric difference in fact appear twice in each of the sums in (4.1); see Figure 6. More precisely, taking into account that $P(C \triangle D) = P(C) + P(D) - 2P(C \cap D)$, it follows that for every $\sigma \in \mathcal{P}_r$

$$\begin{aligned}\sum_{i=1}^{r} P(C_i \triangle D_{\sigma(i)}) &= 2 - 2\sum_{i=1}^{r} P(C_i \cap D_{\sigma(i)}) \\ &= 2P\left(\left\{\bigcup_{i=1}^{r}(C_i \cap D_{\sigma(i)})\right\}^c\right).\end{aligned} \tag{4.2}$$

To avoid this redundancy, our eventual suggestion to measure the distance between $\mathcal{C}$ and $\mathcal{D}$, based on the set distance in P-measure, is $d_P(\mathcal{C}, \mathcal{D}) = \frac{1}{2} d_1(\mathcal{C}, \mathcal{D})$.

If the partitions $\mathcal{C}$ and $\mathcal{D}$ do not have the same number of clusters, then as many empty set components as needed are added so that both partitions include the same number of components, as in Charon et al. (2006), and the distance between the extended partitions is computed as before. Explicitly, if $\mathcal{C} = \{C_1, \ldots, C_r\}$ and $\mathcal{D} = \{D_1, \ldots, D_s\}$ with $r < s$, then, writing $C_i = \varnothing$ for $i = r+1, \ldots, s$, we set

$$\begin{aligned} &d_P(\mathcal{C}, \mathcal{D}) \\ &\quad = \frac{1}{2} \min_{\sigma \in \mathcal{P}_s} \sum_{i=1}^{s} P(C_i \triangle D_{\sigma(i)}) \\ &\quad = \frac{1}{2} \min_{\sigma \in \mathcal{P}_s} \left\{ \sum_{i=1}^{r} P(C_i \triangle D_{\sigma(i)}) + \sum_{i=r+1}^{s} P(D_{\sigma(i)}) \right\}. \end{aligned}$$

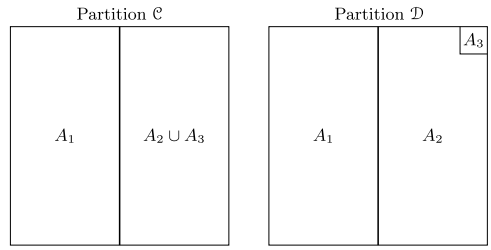

FIG. 7. *Two partitions of the unit square that do not differ much if A_3 has low probability.*

Thus, the term $\sum_{i=r+1}^{s} P(D_{\sigma(i)})$ can be interpreted as a penalization for unmatched probability mass.

The idea is that two partitions such as those shown in Figure 7 do not differ much if A_3 has low probability, even if they do not have the same number of clusters. For the partitions in Figure 7, denote $\mathcal{C} = \{C_1, C_2\}$ and $\mathcal{D} = \{D_1, D_2, D_3\}$ with $C_1 = D_1 = A_1$, $C_2 = A_2 \cup A_3$, $D_2 = A_2$, $D_3 = A_3$, and assume that $P(A_1) = 0.5$, $P(A_2) = 0.45$ and $P(A_3) = 0.05$. Then, it can be shown that $d_P(\mathcal{C}, \mathcal{D}) = 0.05$. In (4.1) every cluster of $\mathcal{C}$ is matched to some cluster in $\mathcal{D}$, depending on the permutation for which the minimum is achieved. When $\mathcal{C}$ has less clusters than $\mathcal{D}$, some of the components of $\mathcal{D}$ will be matched with the empty set, indicating that they do not have an obvious match in $\mathcal{C}$s or that they are unimportant. In the previous example, the minimum is achieved when C_1 is matched with D_1, C_2 with D_2 and D_3 is matched with the empty set.

Indeed, if the existence of unmatched probability mass is considered to be of greater concern, it is always possible to modify the distance in P-measure by introducing a tuning parameter $\lambda \geq 0$ to assign a different weight to the penalization, thus mimicking other existing procedures as penalized regression or pruning of decision trees. In this case, the distance would be defined as

$$\begin{aligned} &d_{P,\lambda}(\mathcal{C}, \mathcal{D}) \\ &\quad = \frac{1}{2} \min_{\sigma \in \mathcal{P}_s} \left\{ \sum_{i=1}^{r} P(C_i \triangle D_{\sigma(i)}) + \lambda \sum_{i=r+1}^{s} P(D_{\sigma(i)}) \right\}, \end{aligned}$$

so that $d_P(\mathcal{C}, \mathcal{D}) = d_{P,1}(\mathcal{C}, \mathcal{D})$.

It is interesting to note that $d_P(\mathcal{C}, \mathcal{D})$ can be estimated in a natural way by replacing P with the empirical measure based on the data $\mathbf{X}_1, \ldots, \mathbf{X}_n$, leading to

$$\begin{aligned} &\widehat{d_P}(\mathcal{C}, \mathcal{D}) \\ &\quad = \frac{1}{2n} \min_{\sigma \in \mathcal{P}_s} \left\{ \sum_{i=1}^{r} \sum_{j=1}^{n} I_{C_i \triangle D_{\sigma(i)}}(\mathbf{X}_j) \right. \\ &\qquad \left. + \sum_{i=r+1}^{s} \sum_{j=1}^{n} I_{D_{\sigma(i)}}(\mathbf{X}_j) \right\}, \end{aligned}$$

where I_A denotes the indicator function of the set A. When $r = s$, it follows from (4.2) that an alternative expression for $d_P(\mathcal{C}, \mathcal{D})$ is

$$d_P(\mathcal{C}, \mathcal{D}) = 1 - \max_{\sigma \in \mathcal{P}_r} \sum_{i=1}^{r} P(C_i \cap D_{\sigma(i)})$$

and, therefore, its sample analogue,

$$\widehat{d_P}(\mathcal{C}, \mathcal{D}) = 1 - \frac{1}{n} \max_{\sigma \in \mathcal{P}_r} \sum_{i=1}^{r} \sum_{j=1}^{n} I_{C_i \cap D_{\sigma(i)}}(\mathbf{X}_j),$$

coincides with the so-called classification distance between two clusterings of the data, whose properties are explored in Meilă (2005, 2007, 2012). For $r < s$, however, $\widehat{d_P}$ differs from the classification distance (which does not include the penalty term), but it corresponds exactly with the transfer distance, studied in detail in Charon et al. (2006) (see also Denœud, 2008). Extending the properties of the transfer distance to its population counterpart suggests an interpretation of $d_P(\mathcal{C}, \mathcal{D})$ as the minimal probability mass that needs to be moved to transform the partition $\mathcal{C}$ into $\mathcal{D}$, hence the connection with the optimal transportation problem.

The above argument allows to recognize $d_P(\mathcal{C}, \mathcal{D})$ as the population version of some commonly used empirical distances between partitions of a data set. However, it should be noted that the estimate $\widehat{d_P}(\mathcal{C}, \mathcal{D})$ requires the two clusterings to be fully known and, hence, it may not be very useful if the goal is to approximate the distance between the ideal population clustering and a data-based clustering.

4.2 A Hausdorff Distance Between Clusterings

An alternative notion of distance between two clusterings based on the Hausdorff metric has been kindly suggested by Professor Antonio Cuevas, noting that precisely this distance was used in Pollard (1981) to measure the discrepancy between the set of sample K-means and the set of population K-means. If (X, ρ) is a metric space and $A, B \subseteq X$ are two nonempty subsets of X, the Hausdorff distance between A and B is defined as

$$d_H(A, B) = \max\Big\{\sup_{a \in A} \inf_{b \in B} \rho(a, b), \sup_{b \in B} \inf_{a \in A} \rho(a, b)\Big\}$$

or, equivalently, as

$$d_H(A, B) = \inf\big\{\varepsilon > 0 \colon A \subseteq B^{\varepsilon} \text{ and } B \subseteq A^{\varepsilon}\big\},$$

where $A^{\varepsilon} = \bigcup_{a \in A} \{x \in X \colon \rho(x, a) \le \varepsilon\}$, and B^{ε} is defined analogously.

In the context of clustering, X can be taken to be the metric space consisting of all the sets of $\mathbb{R}^d$ equipped with the distance $\rho(C, D) = P(C \triangle D)$, once two sets with P-null symmetric difference have been identified to be the same. Then any two clusterings $\mathcal{C} = \{C_1, \ldots, C_r\}$ and $\mathcal{D} = \{D_1, \ldots, D_s\}$ can be viewed as (finite) subsets of X and, therefore, the Hausdorff distance between $\mathcal{C}$ and $\mathcal{D}$ is defined as

$$\begin{aligned} &d_H(\mathcal{C}, \mathcal{D}) \\ &\quad = \max\Big\{\max_{i=1,\ldots,r} \min_{j=1,\ldots,s} P(C_i \triangle D_j), \\ &\qquad \max_{j=1,\ldots,s} \min_{i=1,\ldots,r} P(C_i \triangle D_j)\Big\} \\ &\quad = \inf\big\{\varepsilon > 0 \colon \mathcal{C} \subseteq \mathcal{D}^{\varepsilon} \text{ and } \mathcal{D} \subseteq \mathcal{C}^{\varepsilon}\big\}. \end{aligned}$$

To express it in words, $d_H(\mathcal{C}, \mathcal{D}) \le \varepsilon$ whenever for every $C_i \in \mathcal{C}$ there is some $D_j \in \mathcal{D}$ such that $P(C_i \cdot \triangle D_j) \le \varepsilon$ and vice versa. Hence, as noted by Pollard (1981), if ε is taken to be less than one half of the minimum of distance between the clusters within $\mathcal{C}$ and also less than one half of the minimum distance between the clusters within $\mathcal{D}$, then $d_H(\mathcal{C}, \mathcal{D}) \le \varepsilon$ implies that $\mathcal{C}$ and $\mathcal{D}$ must necessarily have the same number of clusters.

The Hausdorff distance can be regarded as a uniform distance between sets. It is not hard to show, using standard techniques from the Theory of Normed Spaces, that when $r = s$ we have

$$d_H(\mathcal{C}, \mathcal{D}) \le 2d_P(\mathcal{C}, \mathcal{D}) \le r d_H(\mathcal{C}, \mathcal{D}).$$

However, when $r < s$ the distance d_H can be more demanding than d_P, meaning that both partitions have to be really close so that their Hausdorff distance results in a small value. For instance, it can be checked that for the two clusterings of the previous example, shown in Figure 7, the Hausdorff distance between them is $d_H(\mathcal{C}, \mathcal{D}) = 0.45$, mainly due to the fact that C_2 and D_3 are far from each other, since $P(C_2 \triangle D_3) = P(A_2) = 0.45$.

A clear picture of the difference between d_H and d_P is obtained by arranging all the component-wise distances $P(C_i \triangle D_j)$ into an $r \times s$ matrix. Then, the Hausdorff distance is obtained by computing all the row-wise and column-wise minima and taking the maximum of all of them. In contrast, for the distance in P-measure the first step when $r < s$ is to add $s - r$ row copies of the vector $(P(D_1), \ldots, P(D_s))$ to the matrix of component-wise distances, and then compute the distance in P-measure as half the minimum possible sum obtained by adding up a different element in each row. As a further difference, note that the Hausdorff distance does not involve a matching problem;

instead, this distance is solely determined by the two components that are furthest from each other.

Obviously, a sample analogue is also obtained in this case by replacing P for the empirical probability measure, leading to

$$\widehat{d}_H(\mathcal{C},\mathcal{D}) = \frac{1}{n}\max\left\{\max_{i=1,\dots,r}\min_{j=1,\dots,s}\sum_{k=1}^{n} I_{C_i\triangle D_j}(\mathbf{X}_k), \max_{j=1,\dots,s}\min_{i=1,\dots,r}\sum_{k=1}^{n} I_{C_i\triangle D_j}(\mathbf{X}_k)\right\},$$

which seems not to have been considered previously as a distance between two clusterings of the data.

4.3 Consistency of Data-Based Clusterings

As indicated above, a data-based clustering is understood as any procedure that induces a clustering $\widehat{\mathcal{C}}_n$ of a probability distribution P based on the information obtained from a sample $\mathbf{X}_1,\dots,\mathbf{X}_n$ from P. Once a clustering methodology has been chosen, and its ideal population goal $\mathcal{C}_0$ is clearly identified, a data-based clustering $\widehat{\mathcal{C}}_n$ can be said to be consistent if it gets closer to $\mathcal{C}_0$ as the sample size increases. Formally, if $d(\widehat{\mathcal{C}}_n,\mathcal{C}_0)\to 0$ as $n\to\infty$ for some of the modes of stochastic convergence (in probability, almost surely, etc.), d represents one of the distances between clusterings defined above or any other sensible alternative. Note that a different notion of consistency, specifically intended for the cluster tree approach, is studied in Chaudhuri and Dasgupta (2010).

For density-based clustering, a plug-in strategy to obtain data-based clusterings would consist of replacing the unknown density f with an estimator $\hat{f}_n$. Obvious candidates for the role of $\hat{f}_n$ include nonparametric density estimators for modal clustering or mixture model density estimators with parameters fitted by maximum likelihood for mixture model clustering. This is a very simple approach that involves to some extent estimating the density function to solve the clustering problem (unsupervised learning).

According to von Luxburg (2004), page 21, this plug-in strategy may not be a good idea because density estimation is a very difficult problem, especially in high dimensions. However, a similar situation is found in the study of classification (supervised learning), where the optimal classifier, the Bayes rule, depends on the regression function of the random labels over the covariates. Here, even if classification can be proved to be a problem easier than regression, nevertheless, regression-based algorithms for classification play an important role in the development of supervised learning theory (see Devroye, Györfi and Lugosi, 1996, Chapter 6).

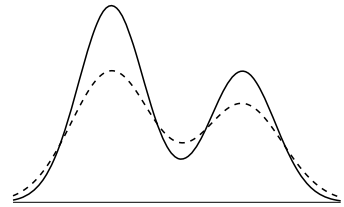

FIG. 8. *Two density functions that are not close but induce exactly the same clustering.*

Along the same lines, Figure 8 illustrates why we should not completely discard density estimation as an intermediate step for clustering. Figure 8 shows a typical situation where the solid line is the true density and the dashed line is a kernel density estimator, since an expansion of its pointwise bias shows that, on average, the kernel estimator underestimates the maxima and overestimates the minima (Wand and Jones, 1995, page 21). But even if the two density functions are not really close in any global sense, they produce exactly the same clusterings of $\mathbb{R}$.

In any case, the following result shows that the plug-in strategy leads to consistent data-based modal clusterings as long as the first and second derivatives of the sequence of density estimators converge uniformly to their true density counterparts.

THEOREM 4.1. *Let a Morse function f be the density of a univariate probability distribution P with compact support, and denote by $\mathcal{C}_0$ the ideal modal clustering that it induces, as defined in Section 3. Let $\{\hat{f}_n\}$ be a sequence of density estimators such that $\hat{f}_n^{(j)}\to f^{(j)}$ uniformly almost surely for $j=1,2$, with $^{(j)}$ standing for the jth derivative. Denote by $\widehat{\mathcal{C}}_n$ the modal clustering induced by $\hat{f}_n$. Then*:

(a) $\#\widehat{\mathcal{C}}_n\to\#\mathcal{C}_0$ *with probability one as* $n\to\infty$, *where* $\#A$ *denotes the number of elements in a set* A.

(b) *Both* $d_P(\widehat{\mathcal{C}}_n,\mathcal{C}_0)\to 0$ *and* $d_H(\widehat{\mathcal{C}}_n,\mathcal{C}_0)\to 0$ *with probability one as* $n\to\infty$.

The proof of this result is shown in the Appendix. The analysis of the proposed distances between clusterings is greatly simplified in the univariate case since the cluster boundaries are solely determined by the points of local minima of the density. The extension of this result for dimension $d\geq 2$ seems quite a challenging

open problem, since the cluster boundaries in dimension d are $(d-1)$-dimensional manifolds which may have very intricate forms.

Part (a) shows that the number of clusters in $\widehat{\mathcal{C}}_n$ converges to the true number of clusters in $\mathcal{C}_0$ almost surely. As indicated in Cuevas, Febrero and Fraiman (2000), since $\#\widehat{\mathcal{C}}_n$ and $\#\mathcal{C}_0$ are integer-valued, this convergence is equivalent to the fact that the event

$$\{\text{There exists } n_0 \in \mathbb{N} \text{ such that } \#\widehat{\mathcal{C}}_n = \#\mathcal{C}_0 \text{ for all } n \geq n_0\}$$

has probability one.

Note also that if $f^{(2)}$ is uniformly continuous and $\hat{f}_n$ are kernel estimators with bandwidth $h = h_n$ based on a sufficiently regular kernel, Silverman (1978), Theorem C, showed that a necessary and sufficient condition for the uniform convergence condition in the previous theorem to hold is just that $h \to 0$ and $nh^5/\log n \to \infty$ as $n \to \infty$ (see also Deheuvels, 1974 and Bertrand-Retali, 1978).

4.4 Asymptotic Loss Approximations

The proof of Theorem 4.1 reveals that, for big enough n, the distance in measure and the Hausdorff distance between $\widehat{\mathcal{C}}_n$ and $\mathcal{C}_0$ can be written as

$$d_P(\widehat{\mathcal{C}}_n, \mathcal{C}_0) = \sum_{j=1}^{r-1} |F(\hat{m}_{n,j}) - F(m_j)| \quad \text{and}$$

$$d_H(\widehat{\mathcal{C}}_n, \mathcal{C}_0) = \max_{j=1,\ldots,r-1} |F(\hat{m}_{n,j}) - F(m_j)|,$$

where F is the distribution function of P. Here, $m_1, \ldots, m_{r-1}$ and $\hat{m}_{n,1}, \ldots, \hat{m}_{n,r-1}$ denote the local minima of f and $\hat{f}_n$, respectively (i.e., the cluster boundaries of $\mathcal{C}_0$ and $\widehat{\mathcal{C}}_n$). From these expressions the L_1 and L_∞ nature of d_P and d_H is even more clear.

Furthermore, under the conditions of Theorem 4.1, after two Taylor expansions it is possible to obtain the approximations

$$|F(\hat{m}_{n,j}) - F(m_j)| \simeq f(m_j)|\hat{m}_{n,j} - m_j| \simeq \frac{f(m_j)}{f''(m_j)}|\hat{f}_n'(m_j)|.$$

This shows how not only the performance of $\widehat{\mathcal{C}}_n$ is closely connected to the problem of first-derivative estimation, but also that modal clustering is more difficult, as the density at the cluster boundaries is higher and/or flatter as the intuition dictates.

In the case of kernel estimators, Proposition 4.1 of Romano (1988) provides a precise description of the asymptotic behavior of $\hat{f}_n'(m_j)$. Precisely, under some smoothness conditions it can be shown that assuming that the bandwidth further satisfies $nh^7 \to \beta^2$ with $0 \leq \beta < \infty$, then $\hat{f}_n'(m_j)$ admits the representation

$$\hat{f}_n'(m_j) = (nh^3)^{-1/2}\sigma Z_n + \beta\mu$$

for some explicit constants $\sigma > 0$ and $\mu \in \mathbb{R}$, where Z_n is a sequence of asymptotically $N(0,1)$ random variables. This representation could be helpful as a starting point to tackle the problem of optimal bandwidth choice for kernel clustering, which has only been treated briefly in the previous literature (e.g., Einbeck, 2011, Chacón and Duong, 2013, Chacón and Monfort, 2014) and surely deserves further investigation. However, we will not pursue this further here.

5. DISCUSSION

At the time of comparing different clustering procedures, it is necessary to have a "ground truth," or population goal, that represents the ideal clustering to which the clustering algorithms should try to get close. The importance of having a clear population goal for clustering is nicely highlighted in Klemelä (2009), Chapter 8. Sometimes this ideal population clustering is not so easy to specify, and of course it depends on the notion of cluster in which the researcher is interested.

Whereas the population goal is clearly defined for some clustering methods, like K-means clustering or mixture model clustering, it remained less obvious for modal clustering. Here, the ideal population goal of modal clustering is accurately identified, making use of some tools from Morse theory as the partition of the space induced by the domains of attraction of the local maxima of the density function.

This definition of the modal clusters needs the probability density to be smooth to a certain degree, specifically it must be a 3-times continuously differentiable Morse function. It would be appealing to extend this notion to density functions that are not Morse functions, meaning either that they are smooth but have degenerate critical points or even that they are not differentiable to such extent. To treat the first case, it might be useful to resort to the theory of singularities of differential mappings, which is exhaustively covered in the book by Arnold et al. (1998), for instance. On the other hand, the study of the nonsmooth case might start from Agrachev, Pallaschke and Scholtes (1997), where Morse theory for piecewise smooth functions is presented. Here, the key role would be played by the subgradient, which generalizes the concept of the gradient for nonsmooth functions.

Alternatively, as in Donoho (1988), a nonsmooth density f could be convolved with a mollifier ϕ_h to obtain a smoother version $\phi_h * f$, so that the population modal clustering $\mathcal{C}_h$ of $\phi_h * f$ is determined as in the smooth case, and then define the population modal clustering of f as the limit (in some sense) of $\mathcal{C}_h$ as $h \to 0$. Of course, further investigation on how to properly formalize this notion would be required.

Once a clustering methodology with a clearly defined population goal has been chosen, it is necessary to have a distance to measure the accuracy of data-based clusterings as approximations of the ideal goal. A second contribution of this paper is the introduction of two new loss functions for this aim, which are valid for any clustering methodology. Particularly, when applied to modal clustering, it is shown that the plug-in approach leads to clustering consistency under mild assumptions.

A further interesting challenge for future research consists of studying the choice of the parameters for the density estimators (the bandwidth for kernel estimators, the mixture parameters for mixture model estimators) that minimize the distance between the corresponding data-based clustering and the true population clustering, as measured by any of the distances between clusterings discussed in Section 4. Or, maybe even better, to develop methods aimed to perform modal clustering that do not necessarily rely on a pilot density estimate, perhaps by somehow adapting those classification methods whose construction is not based on a regression estimate.

APPENDIX: PROOF OF THE CONSISTENCY THEOREM

The proof uses some arguments from Theorem 3 in Cuevas and González Manteiga (1991); see also Lemma 3 in Genovese et al. (2015).

First, since f is a Morse function with compact support, it has only finitely many isolated critical points (Matsumoto, 2002, Corollary 2.19). Assume that f has r local maxima and let $m_1 < \cdots < m_{r-1}$ denote the local minima of f so that the modal population clustering induced by f is defined as $\mathcal{C}_0 = \{C_1, \ldots, C_r\}$ with $C_j = (m_{j-1}, m_j)$ for $j = 1, \ldots, r$, where $m_0 = -\infty$ and $m_r = \infty$ (if f has no local minimum, then $r = 1$ and $\mathcal{C}_0 = \{C_1\} = \{\mathbb{R}\}$).

We claim the following: with probability one, there exists $n_0 \in \mathbb{N}$ such that $\hat{f}_n$ has exactly $r-1$ local minima for all $n \geq n_0$; moreover, there exists $\varepsilon > 0$ such that every $\hat{f}_n$ with $n \geq n_0$ has exactly one local minimum $\hat{m}_{n,j}$ in $[m_j - \varepsilon, m_j + \varepsilon]$ for all $j = 1, \ldots, r-1$. To prove this claim, notice that since $f''(m_j) > 0$ for all j, and f'' is continuous, it is possible to find some $\varepsilon > 0$ such that $f''(x) > 0$ on $[m_j - \varepsilon, m_j + \varepsilon]$, for all j. The almost sure uniform convergence of $\hat{f}_n''$ to f'' implies that there is some $n_0 \in \mathbb{N}$ such that, with a possibly smaller ε, all $\hat{f}_n''$ with $n \geq n_0$ are strictly positive on those intervals as well. On the other hand, on each of these intervals f' is strictly increasing and since $f'(m_j) = 0$, it must go from negative to positive. But the uniform convergence of $\hat{f}_n'$ to f' implies that also $\hat{f}_n'$ must go from negative to positive (perhaps with a smaller ε) for big enough n. Therefore, all of them must have a critical point there, and since we previously showed that $\hat{f}_n'' > 0$, this means both that the critical point is a local minimum and that there cannot be any more of them in such neighborhoods of the local minima. A similar argument shows that, for big enough n, all the $\hat{f}_n$ with $n \geq n_0$ must also have a local maximum in a small enough neighborhood around the modes of f, and that there cannot be other critical points of $\hat{f}_n$ outside these neighborhoods.

Furthermore, using standard arguments in M-estimation theory, under these conditions it follows that also $\hat{m}_{n,j}$ converges to m_j as $n \to \infty$: to show this, notice that given an arbitrary $\eta > 0$, small enough so that $\eta < \varepsilon$, the value of $\delta := \inf\{|f'(x)| : \eta \leq |x - m_j| \leq \varepsilon\}$ is strictly positive. Hence, from the almost sure uniform convergence $\hat{f}_n' \to f'$ it follows that, with probability one, for all big enough n we have $|\hat{f}_n'(x)| > \delta/2 > 0$ whenever $\eta \leq |x - m_j| \leq \varepsilon$. Since $|\hat{m}_{n,j} - m_j| \leq \varepsilon$ and $\hat{f}_n'(\hat{m}_{n,j}) = 0$, this implies that $|\hat{m}_{n,j} - m_j| < \eta$.

In this situation, for the clustering $\widehat{\mathcal{C}}_n = \{\widehat{C}_{n1}, \ldots, \widehat{C}_{n,r}\}$ induced by $\hat{f}_n$ [with $\widehat{C}_{n,j} = (\hat{m}_{n,j-1}, \hat{m}_{n,j})$, $\hat{m}_{n,0} = -\infty$ and $\hat{m}_{n,r} = \infty$], taking a small enough ε, the distance in P-measure and the Hausdorff distance between $\widehat{\mathcal{C}}_n$ and $\mathcal{C}$ can be simply written as

$$d_P(\widehat{\mathcal{C}}_n, \mathcal{C}_0) = \sum_{j=1}^{r-1} \left|F(\hat{m}_{n,j}) - F(m_j)\right|,$$

$$d_H(\widehat{\mathcal{C}}_n, \mathcal{C}_0) = \max_{j=1,\ldots,r-1} \left|F(\hat{m}_{n,j}) - F(m_j)\right|,$$

respectively, where F is the distribution function of P. Therefore, the convergence of the estimated local minima to the true local minima of f yields the result.

ACKNOWLEDGMENTS

The author wishes to thank Professor Antonio Cuevas from Universidad Autónoma de Madrid as well

as Professor Ricardo Faro from Universidad de Extremadura for insightful conversations and suggestions concerning the material of Section 4. The paper by Ray and Lindsay (2005), in which interesting connections between Morse theory and the topography of multivariate normal mixtures are illustrated, was thought-provoking enough to inspire part of this paper.

Supported in part by Spanish Ministerio de Ciencia y Tecnología projects MTM2010-16660, MTM2010-17366 and MTM2013-44045-P, and by the Gobierno de Extremadura Grant GR10064.

REFERENCES

ACKERMAN, M. and BEN-DAVID, S. (2009). Measures of clustering quality: A working set of axioms for clustering. In *Advances in Neural Information Processing Systems* **21** (D. Koller, D. Schuurmans, Y. Bengio and L. Bottou, eds.) 121–128. Curran Associates, Red Hook, NY. Available at http://papers.nips.cc/paper/3491-measures-of-clustering-quality-a-working-set-of-axioms-for-clustering.pdf.

AGRACHEV, A. A., PALLASCHKE, D. and SCHOLTES, S. (1997). On Morse theory for piecewise smooth functions. *J. Dyn. Control Syst.* **3** 449–469. MR1481622

ARABIE, P. and BOORMAN, S. A. (1973). Multidimensional scaling of measures of distance between partitions. *J. Math. Psych.* **10** 148–203. MR0321559

ARIAS-CASTRO, E., MASON, D. and PELLETIER, B. (2013). On the estimation of the gradient lines of a density and the consistency of the mean-shift algorithm. Preprint.

ARNOLD, V. I., GORYUNOV, V. V., LYASHKO, O. V. and VASIL'EV, V. A. (1998). *Singularity Theory. I.* Springer, Berlin. MR1660090

AZZALINI, A. and TORELLI, N. (2007). Clustering via nonparametric density estimation. *Stat. Comput.* **17** 71–80. MR2370969

BEN-DAVID, S., VON LUXBURG, U. and PÁL, D. (2006). A sober look at clustering stability. In *Learning Theory* (G. Lugosi and H.-U. Simon, eds.). *Lecture Notes in Computer Science* **4005** 5–19. Springer, Berlin. MR2277915

BERTRAND-RETALI, M. (1978). Convergence uniforme d'un estimateur de la densité par la méthode du noyau. *Rev. Roumaine Math. Pures Appl.* **23** 361–385. MR0494658

BURKARD, R., DELL'AMICO, M. and MARTELLO, S. (2009). *Assignment Problems*. SIAM, Philadelphia, PA. MR2488749

CADRE, B., PELLETIER, B. and PUDLO, P. (2013). Estimation of density level sets with a given probability content. *J. Nonparametr. Stat.* **25** 261–272. MR3039981

CARLSSON, G. (2009). Topology and data. *Bull. Amer. Math. Soc. (N.S.)* **46** 255–308. MR2476414

CARLSSON, G. and MÉMOLI, F. (2013). Classifying clustering schemes. *Found. Comput. Math.* **13** 221–252. MR3032681

CHACÓN, J. E. (2009). Data-driven choice of the smoothing parametrization for kernel density estimators. *Canad. J. Statist.* **37** 249–265. MR2531830

CHACÓN, J. E. (2012). Clusters and water flows: A novel approach to modal clustering through Morse theory. Preprint. Available at arXiv:1212.1384.

CHACÓN, J. E. and DUONG, T. (2013). Data-driven density derivative estimation, with applications to nonparametric clustering and bump hunting. *Electron. J. Stat.* **7** 499–532. MR3035264

CHACÓN, J. E. and MONFORT, P. (2014). A comparison of bandwidth selectors for mean shift clustering. In *Theoretical and Applied Issues in Statistics and Demography* (C. H. Skiadas, ed.) 47–59. International Society for the Advancement of Science and Technology (ISAST), Athens.

CHARON, I., DENŒUD, L., GUÉNOCHE, A. and HUDRY, O. (2006). Maximum transfer distance between partitions. *J. Classification* **23** 103–121. MR2280697

CHAUDHURI, K. and DASGUPTA, S. (2010). Rates of convergence for the cluster tree. In *Advances in Neural Information Processing Systems* (J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel and A. Culotta, eds.) **23** 343–351. Curran Associates, Red Hook, NY.

CHAZAL, F., GUIBAS, L. J., OUDOT, S. Y. and SKRABA, P. (2013). Persistence-based clustering in Riemannian manifolds. *J. ACM* **60** Art. 41, 38. MR3144911

CUEVAS, A., FEBRERO, M. and FRAIMAN, R. (2000). Estimating the number of clusters. *Canad. J. Statist.* **28** 367–382. MR1792055

CUEVAS, A., FEBRERO, M. and FRAIMAN, R. (2001). Cluster analysis: A further approach based on density estimation. *Comput. Statist. Data Anal.* **36** 441–459. MR1855727

CUEVAS, A. and FRAIMAN, R. (2010). Set estimation. In *New Perspectives in Stochastic Geometry* (W. Kendall and I. Molchanov, eds.) 374–397. Oxford Univ. Press, Oxford. MR2654684

CUEVAS, A. and GONZÁLEZ MANTEIGA, W. (1991). Data-driven smoothing based on convexity properties. In *Nonparametric Functional Estimation and Related Topics* (*Spetses*, 1990) (G. Roussas, ed.). *NATO Adv. Sci. Inst. Ser. C Math. Phys. Sci.* **335** 225–240. Kluwer Academic, Dordrecht. MR1154331

DAY, W. H. E. (1980/81). The complexity of computing metric distances between partitions. *Math. Social Sci.* **1** 269–287. MR0616380

DEHEUVELS, P. (1974). Conditions nécessaires et suffisantes de convergence ponctuelle presque sûre et uniforme presque sûre des estimateurs de la densité. *C. R. Acad. Sci. Paris Sér. A* **278** 1217–1220. MR0345296

DENŒUD, L. (2008). Transfer distance between partitions. *Adv. Data Anal. Classif.* **2** 279–294. MR2469771

DEVROYE, L., GYÖRFI, L. and LUGOSI, G. (1996). *A Probabilistic Theory of Pattern Recognition. Applications of Mathematics* (*New York*) **31**. Springer, New York. MR1383093

DONOHO, D. L. (1988). One-sided inference about functionals of a density. *Ann. Statist.* **16** 1390–1420. MR0964930

EDELSBRUNNER, H. and HARER, J. (2008). Persistent homology—A survey. In *Surveys on Discrete and Computational Geometry. Contemp. Math.* **453** 257–282. Amer. Math. Soc., Providence, RI. MR2405684

EINBECK, J. (2011). Bandwidth selection for mean-shift based unsupervised learning techniques: A unified approach via self-coverage. *Journal of Pattern Recognition Research* **6** 175–192.

EVERITT, B. S., LANDAU, S., LESSE, M. and STAHL, D. (2011). *Cluster Analysis*, 5th ed. Wiley, Chichester.

FASY, B. T., LECCI, F., RINALDO, A., WASSERMAN, L., BALAKRISHNAN, S. and SINGH, A. (2014). Statistical inference for persistent homology: Confidence sets for persistence diagrams. Available at arXiv:1303.7117v2.

FRALEY, C. and RAFTERY, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *J. Amer. Statist. Assoc.* **97** 611–631. MR1951635

FUKUNAGA, K. and HOSTETLER, L. D. (1975). The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Trans. Inform. Theory* **IT-21** 32–40. MR0388638

GENOVESE, C. R., PERONE-PACIFICO, M., VERDINELLI, I. and WASSERMAN, L. (2015). Non-parametric inference for density modes. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* To appear. DOI:10.1111/rssb.12111.

GRAF, S. and LUSCHGY, H. (2000). *Foundations of Quantization for Probability Distributions*. *Lecture Notes in Math.* **1730**. Springer, Berlin. MR1764176

HAND, D., MANNILA, H. and SMYTH, P. (2001). *Principles of Data Mining*. MIT Press, Cambridge, MA.

HARTIGAN, J. A. (1975). *Clustering Algorithms*. Wiley, New York. MR0405726

HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2009). *The Elements of Statistical Learning*: *Data Mining*, *Inference*, *and Prediction*, 2nd ed. Springer, New York. MR2722294

IZENMAN, A. J. (2008). *Modern Multivariate Statistical Techniques*: *Regression*, *Classification*, *and Manifold Learning*. Springer, New York. MR2445017

JOST, J. (2011). *Riemannian Geometry and Geometric Analysis*, 6th ed. *Universitext*. Springer, Heidelberg. MR2829653

KLEMELÄ, J. (2009). *Smoothing of Multivariate Data*: *Density Estimation and Visualization*. Wiley, Hoboken, NJ. MR2640738

LI, J., RAY, S. and LINDSAY, B. G. (2007). A nonparametric statistical approach to clustering via mode identification. *J. Mach. Learn. Res.* **8** 1687–1723. MR2332445

MACQUEEN, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proc. Fifth Berkeley Sympos. Math. Statist. and Probability* (*Berkeley*, *Calif.*, 1965/66) 281–297. Univ. California Press, Berkeley. MR0214227

MASON, D. M. and POLONIK, W. (2009). Asymptotic normality of plug-in level set estimates. *Ann. Appl. Probab.* **19** 1108–1142. MR2537201

MATSUMOTO, Y. (2002). *An Introduction to Morse Theory*. *Translations of Mathematical Monographs* **208**. Amer. Math. Soc., Providence, RI. MR1873233

MEILĂ, M. (2005). Comparing clusterings—an axiomatic view. In *Proceedings of the International Machine Learning Conference* (*ICML*) (S. Wrobel and L. De Raedt, eds.) 577–584. ACM Press, New York.

MEILĂ, M. (2007). Comparing clusterings—an information based distance. *J. Multivariate Anal.* **98** 873–895. MR2325412

MEILĂ, M. (2012). Local equivalences of distances between clusterings—a geometric perspective. *Mach. Learn.* **86** 369–389. MR2897528

MENARDI, G. and AZZALINI, A. (2014). An advancement in clustering via nonparametric density estimation. *Stat. Comput.* **24** 753–767. MR3229695

MILNOR, J. (1963). *Morse Theory*. Princeton Univ. Press, Princeton, NJ. MR0163331

NUGENT, R. and STUETZLE, W. (2010). Clustering with confidence: A low-dimensional binning approach. In *Classification as a Tool for Research* (H. Locarek-Junge and C. Weihs, eds.) 117–125. Springer, Berlin. MR2722129

POLLARD, D. (1981). Strong consistency of k-means clustering. *Ann. Statist.* **9** 135–140. MR0600539

RAY, S. and LINDSAY, B. G. (2005). The topography of multivariate normal mixtures. *Ann. Statist.* **33** 2042–2065. MR2211079

RINALDO, A., SINGH, A., NUGENT, R. and WASSERMAN, L. (2012). Stability of density-based clustering. *J. Mach. Learn. Res.* **13** 905–948. MR2930628

RODRÍGUEZ-CASAL, A. (2003). Estimación de Conjuntos y sus Fronteras. Un Enfoque Geométrico. Ph.D. thesis, Univ. Santiago de Compostela.

ROMANO, J. P. (1988). On weak convergence and optimality of kernel density estimates of the mode. *Ann. Statist.* **16** 629–647. MR0947566

SILVERMAN, B. W. (1978). Weak and strong uniform consistency of the kernel estimate of a density and its derivatives. *Ann. Statist.* **6** 177–184. MR0471166

STUETZLE, W. (2003). Estimating the cluster type of a density by analyzing the minimal spanning tree of a sample. *J. Classification* **20** 25–47. MR1983120

THOM, R. (1949). Sur une partition en cellules associée à une fonction sur une variété. *C. R. Acad. Sci. Paris* **228** 973–975. MR0029160

TSYBAKOV, A. B. (1997). On nonparametric estimation of density level sets. *Ann. Statist.* **25** 948–969. MR1447735

VITALLI, M. (2010). Morse decomposition of geometric meshes with applications. Ph.D. thesis, Università di Genova.

VON LUXBURG, U. (2004). Statistical learning with similarity and dissimilarity functions. Ph.D. thesis, Technical Univ. Berlin.

VON LUXBURG, U. and BEN-DAVID, S. (2005). Towards a statistical theory for clustering. In *PASCAL Workshop on Statistics and Optimization of Clustering*.

WAND, M. P. and JONES, M. C. (1993). Comparison of smoothing parameterizations in bivariate kernel density estimation. *J. Amer. Statist. Assoc.* **88** 520–528. MR1224377

WAND, M. P. and JONES, M. C. (1995). *Kernel Smoothing*. *Monographs on Statistics and Applied Probability* **60**. Chapman & Hall, London. MR1319818

WANG, X., QIU, W. and ZAMAR, R. H. (2007). CLUES: A nonparametric clustering method based on local shrinking. *Comput. Statist. Data Anal.* **52** 286–298. MR2409982

WISHART, D. (1969). Mode analysis: A generalization of nearest neighbor which reduces chaining effects. In *Numerical Taxonomy* (A. J. Cole, ed.) 282–311. Academic Press, New York.

ZADEH, R. B. and BEN-DAVID, S. (2009). A uniqueness theorem for clustering. In *UAI'09 Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence* 639–646.