International Statistical Review

International Statistical Review (2015), 0, 0, 1–21 doi:10.1111/insr.12109

A Review on Modal Clustering

Giovanna Menardi

Dipartimento di Scienze Statistiche, Università di Padova, via C. Battisti, 241, 35121 Padova, Italy
E-mail: menardi@stat.unipd.it

Summary

In spite of the current availability of numerous methods of cluster analysis, evaluating a clustering configuration is questionable without the definition of a true population structure, representing the ideal partition that clustering methods should try to approximate. A precise statistical notion of cluster, unshared by most of the mainstream methods, is provided by the density-based approach, which assumes that clusters are associated to some specific features of the probability distribution underlying the data. The non-parametric formulation of this approach, known as modal clustering, draws a correspondence between the groups and the modes of the density function. An appealing implication is that the ill-specified task of cluster detection can be regarded to as a more circumscribed problem of estimation, and the number of clusters is also conceptually well defined. In this work, modal clustering is critically reviewed from both conceptual and operational standpoints. The main directions of current research are outlined as well as some challenges and directions of further research.

Key words: Density-based clustering; kernel density estimation; mode hunting; non-parametric estimation.

1 Background and motivation

The task of partitioning a set of data into a number of groups, which goes under the heading of cluster analysis, has been pervasively pursued in many fields, both as a preliminary step to explore data and as a targeted focus of the data analysis, for example, for classification or data reduction. Hundreds of techniques have been proposed over the years, most of them building on some notion of distance or dissimilarity. The soundness of these techniques is questionable, as they rely on a heuristic notion of cluster and this lack of a 'ground truth' prevents us from evaluating a clustering configuration or comparing alternatives.

A precise statistical notion is provided by the density-based approach, where clusters are associated to some specific features of the probability distribution assumed to underlie the data. This idea has been developed in two distinct directions. The model-based approach, whose roots date back at least in the seventies (e.g. Wolfe, 1970) regards the probability distribution underlying the data as a mixture of subpopulations, assumed to have some parametric form. A cluster is associated to each component of the mixture, and the observations are allocated to the cluster with maximal density among the components. This approach is nowadays well established, and it is not considered here. Standard accounts are the book of McLachlan & Basford (1988) as well as the seminal works of Fraley and Raftery (1998, 2002).

The focus is on a second, less widespread, density-based clustering formulation, often referred to as *modal* or *nonparametric* clustering. It finds its roots in Carmichael *et al.* (1968)

who first attempted to define a concept of clusters close to the natural intuition, as 'continuous, relatively densely populated regions of the space, surrounded by continuous, relatively empty regions'. This, still vague, definition of clusters as high-density regions was then somewhat developed by Wishart (1969) who explicitly stated that clustering methods should be able to identify 'distinct data modes, independently of their shapes and variance' and put forward an operational procedure consistent with this notion. A few years later, Hartigan (1975) introduced the concept of *density-contour clusters* as the maximal connected subsets of the density level sets. In a similar perspective, Fukunaga & Hostetler (1975) discussed the use of a density gradient estimator to detect the modes of a density function for clustering purposes.

The outlined notion of cluster claims several reasons of attractiveness. It is not bound to a particular shape, and its complying with geometric intuition makes it close to a 'natural' grouping of data. Also, the number of clusters is an intrinsic property of the data generator mechanism, thereby well defined, at least conceptually, and its determination is itself an integral part of the estimation procedure.

Despite its attractiveness, the study of modal clustering came to a standstill after the aforementioned early works and was resumed only recently, owing to the current computational advances. However, its development has been quite scattered, suffering from a substantial lack of interconnection among the inherent works. The purpose of this paper is to review critically and systematise the state of the art about modal clustering. A taxonomy of the directions of the current developments is presented, also enlightening the connections with some alternative methods (Section 2). We discuss some aspects related to modal clustering, which we see as the most conceptually and practically relevant (Section 3) and review the software available (Section 4). Final remarks and an outline of the current and future directions of research conclude the work.

2 An overview of modal clustering methods

2.1 Formalisation

To recast clustering into the frame of a standard statistical problem, we refer to the observed data $(x_1, \ldots, x_n)'$ as a sample of n realisations of a d-dimensional random vector from a probability density function $f : \mathcal{X} \subseteq \mathbb{R}^d \mapsto \mathbb{R}^+$. Typically, interest focuses on some characteristics of the unknown density function, as a functional or a set of parameters on which f depends. Based on the sample, we wish to make inference on f and, then, on its characteristics. In a clustering problem, the characteristic of interest is a partition of $\mathcal{X}$ into a number of (*population*) clusters induced by f. Therefore, in principle, a clustering method should be able to estimate a partition of the sample space. In practice, we often get content of detecting the *empirical* clusters, which just represent a partition of the observed data.

Modal clustering builds on the idea that population clusters correspond to the 'domain of attraction' of the modes of f (Stuetzle, 2003). While still rather vague, this cluster notion already looks sufficient for circumscribing the clustering task and relieving its ill posedness. The modes of the density function are regarded as the archetypes of the clusters, which are in turn represented by the surrounding regions.

An elegant attempt of translating these concepts into a formal definition has been made by Chacón (2014). When $d = 1$, this is rather immediate, at least for regular densities: points at which f has a local maximum are representative of the clusters, whose boundaries are the points where f has local minima. However, the extension for $d > 1$ is not obvious. By exploiting some notions from differential geometry, Chacón (2014) defines a cluster as the unstable manifold of the negative gradient flow corresponding to the local maxima of

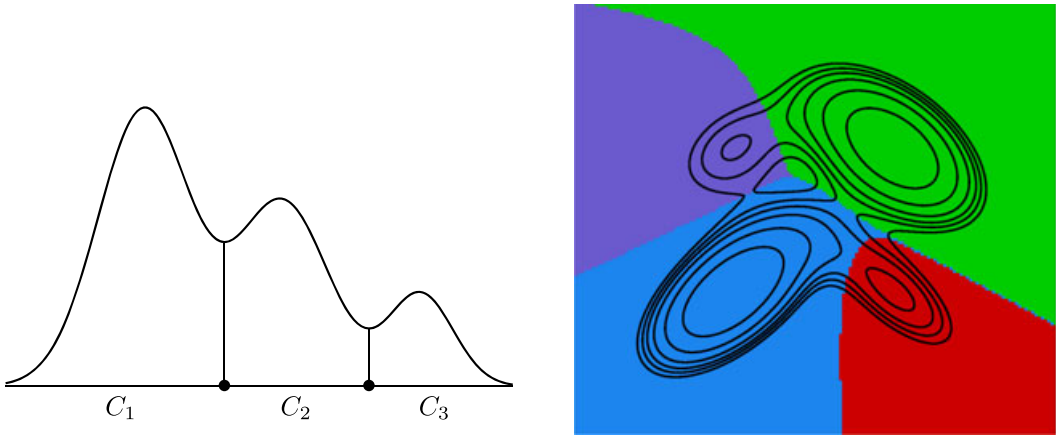

Figure 1. *An example of univariate and bivariate density that enlightened the domain of attraction of each mode, as presented by Chacón (2014).*

f. Intuitively, if f is figured as a mountainous landscape, and modes are its peaks, clusters are the 'regions that would be flooded by a fountain emanating from a peak of the mountain range'. These notions are illustrated for univariate and bivariate examples in **F1**Figure 1. It is worth to mention that similar ideas had been earlier studied by Ray & Lindsay (2005), who analysed the number and location of the modes of a mixture of normal densities by introducing the *ridgeline manifold*, having a strong potential impact on the clustering problem (Li *et al.*, 2007).

The mentioned definition of cluster, albeit attractive in its precision, meets our expectations for rather simple situations only, as it requires the density to be a Morse function (see, e.g. Matsumoto, 2002). Indeed, the Morse theory relies on strong assumptions about the smoothness of f and the non-degeneracy of its critical points, which appear often unrealistic. Also, one assumption is that critical points are isolated and have distinct critical values. This assumption fails when considering non-standard densities as, for example, functions with plateaux. In fact, the concept of modality is itself not univoque (Klemelä, 2009) as, for example, the uniform distribution could be regarded as both unimodal and without modes. In order to account for such situations, we need to allow for a general definition of a mode. With some abuse of notation, we stem from Klemelä's (2009) definition of level-set unimodality and consider the following:

Definition 1. *Let* $f : \mathcal{X} \subseteq \mathbb{R}^d \mapsto \mathbb{R}$ *a probability density function. A connected*[†] *set of points* $\mathcal{M} \subseteq \mathcal{X}$ *is a* mode *of* f *if* $\forall\, m \in \mathcal{M}$, $f(m) = h > 0$ *and there exists a connected set* $C_{\mathcal{M}} \supset \mathcal{M}$ *such that* $f(m) \geq f(x)\ \forall x \in C_{\mathcal{M}}, \forall m \in \mathcal{M}$.

Given the modes $\mathcal{M}_1, \ldots, \mathcal{M}_K$, their domain of attraction has to be regarded as the maximum associated sets $C_{\mathcal{M}_1}, \ldots, C_{\mathcal{M}_K}$ such as those $C_{\mathcal{M}_{k_1}} \not\supset \mathcal{M}_{k_2}$ and $C_{\mathcal{M}_{k_1}} \cap C_{\mathcal{M}_{k_2}} = \emptyset$ for $k_1 \neq k_2$.

It should be noticed that this definition is also subject to some limitations since because it does not guarantee that clusters form a partition of the sample space.

In the next sections, we outline two main strands of methods performing modal clustering, in addition to some scattered contributions, which are more difficult to be framed.

[†] Correction added on 15 July 2015, after first online publication: the word "connected" was added to Definition 1.

2.2 Mode Hunting Methods

Especially widespread in the computer science and engineering communities, a first strand of modal clustering methods aims at finding explicitly the modes of the density underlying the data and then associates each observation to the pertaining mode. Most of contributions that follow this direction can be considered as a refinement of the *mean-shift* clustering, proposed earlier by Fukunaga & Hostetler (1975). Consider the kernel estimator

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} K_H(x - x_i) \tag{1}$$

of the underlying density f, with H a symmetric positive definite bandwidth matrix, $K_H(x) = |H|^{-1/2}K(H^{-1/2}x)$ and K an integrating-to-one kernel function.

The authors considered the simple case where $H = h^2 I$, where h is a positive constant and I is the identity matrix. They derived conditions on K to assure that the gradient $\nabla\hat{f}$ of the estimator (1) is a well-behaved estimator of the density gradient. They developed an algorithm that, starting from a generic point $x^{(0)}$, recursively shifts it to a local weighted mean, until convergence. Denoted by $w_i(x^{(s)})$ the vector of weights of the components of x_i at step s, at the next step $s + 1$,

$$\begin{aligned} x^{(s+1)} &= \sum_{i=1}^{n} w_i(x^{(s)})x_i \\ &= x^{(s)} + \left[\sum_{i=1}^{n} w_i(x^{(s)})x_i - x^{(s)}\right] \\ &= x^{(s)} + M(x^{(s)}), \end{aligned}$$

where $M(x^{(s)})$ denotes the mean shift. Up to a normalising factor, the weights $w_i(x)$ are specified as $\nabla K_H(x - x_i)$, that is, the gradient vectors of the kernel function. Hence, the mean shift is shown to be a gradient ascent algorithm based on a normalised kernel estimator of the gradient:

$$x^{(s+1)} = x^{(s)} + a\frac{\nabla\hat{f}(x^{(s)})}{\hat{f}(x^{(s)})}$$

(in fact, the preceding interpretation holds when $\nabla\hat{f}(x^{(s)})$ is an estimate of $\nabla\hat{f}(x^{(s)})$ based on a *shadow kernel*, which is, in general, different, from K). While not affecting the direction of the shift, the use of a normalised gradient estimate is shown to claim some desirable properties as, among others, a faster convergence to the local maxima of $\hat{f}$. About the constant a, the authors give reasons for a conservative selection.

It stands to reason the use of the mean shift for clustering purposes, as at each iteration, data points are moved a small step toward the closest mode along the direction of the gradient. See Figure 2**F2** for an illustration. As the algorithm can be applied starting from any point in $\mathbb{R}^d$, not only does it cluster the observed data but eventually induces a partition of $\mathbb{R}^d$, where clusters are regions of the sample space formed by points pertaining to the same mode.

Perhaps for computational reasons, mean-shift clustering had been largely neglected until Cheng (1995) re-brought it to light and generalised the class of kernel functions for which the gradient ascend interpretation of the mean-shift algorithm is valid. After that, a number of variant forms of mean-shift clustering has been discussed (Comaniciu and Meer, 2002;

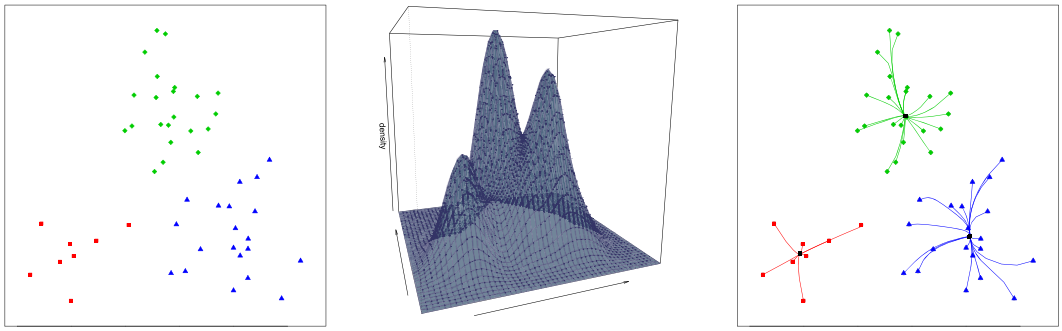

Figure 2. *A simple set of data in $\mathbb{R}^2$, which exhibits three clear clusters (left panel) and its generating three-modal density function (middle panel). The mode-climbing path of each observations identified by the mean shift is illustrated in the right panel.*

Carreira-Perpiñán, 2008; Yuan *et al.*, 2012, just to cite a few). In most of these works, the purely statistical motivation for this approach to clustering has been lost, and the focus has been mainly addressed toward algorithmic aspects, or to the study of the kernel density estimated during the mode-seeking process (Section 3.2).

Similar from a technical viewpoint, but claiming a stronger inferential motivation, is the work of Li *et al.* (2007). They observe that, by their nature, kernel densities have a mixture structure, yet of a special type, as at least the centres of the mixture components are known. This allows to develop an approach that inherits the advantages of both model-based clustering and non-parametric methods. A posterior probability can be, in fact, associated to the clusters for gaining insight into the data or computed for each observation and used for diagnostics or soft clustering. At the same time, the use of a non-parametric estimator guarantees a correspondence between the clusters (associated to the density bumps) and the geometric characteristics of the density function, which is prevented by model-based clustering. Given the mixture structure of the kernel density estimates, Li *et al.* (2007) discuss the use of an expectation–maximisation (EM)-style algorithm, *modal EM* (MEM), aimed at seeking the local maxima of the given kernel mixture. The similarity with the methods described earlier is evident, all the more because the EM algorithm has been proven to be a special type of mean shift (Carreira-Perpinan, 2007). Li *et al.* (2007) further propose a smart hierarchical extension of clustering. At level l of the hierarchy, a kernel density estimate is built with bandwidth $h_l > h_{l-1}$, and the MEM algorithm is run to obtain the parent modes of the modes detected at level $l-1$. Hence, by construction, the two corresponding partitions at levels $l-1$ and l are nested.

2.3 Level Sets-Based Methods

A second strand of density-based clustering methods, which follows the route of Hartigan (1975), does not explicitly link clusters to some representative points, that is, the modes, of the sample space, but associates the clusters with high-density regions of the sample space, defined by the density level sets.

Formally, any section of the density function at a level λ identifies the (upper) level set:

$$L(\lambda) = \{x \in \mathbb{R}^d : f(x) \geq \lambda\}, \quad 0 \leq \lambda \leq \max f. \tag{2}$$

For any choice of λ, $L(\lambda)$ may be connected or disconnected. In the latter case, it consists of a number of connected components, each of them associated with a cluster at the level λ. As $L(\lambda)$ is unknown, an estimate $\hat{L}(\lambda)$ of $L(\lambda)$ is obtained by replacing $f(x)$, in (2), with a non-parametric estimate $\hat{f}(x)$.

The rationale behind this class of methods is that any connected component of $L(\lambda)$ includes at least one mode of the density function, and, on the other hand, for each mode of the density function, there exists λ for which one of the connected components of the associated $L(\lambda)$ includes this mode at most.

Of course, there may not exist a single λ^* for which each mode of the density function belongs to a distinct connected component of $L(\lambda^*)$, or, even if it does, one should be able to identify it. We refer the reader to Section 2.3.1 to discuss how this is not, in fact, an issue. For now, to ease exposition, we assume that such λ^* exists and it is known.

An awkward aspect that has seriously limited the application of this approach is that, given $\hat{L}(\lambda^*)$, finding its connected components is both conceptually and computationally easy in the unidimensional case only, where connected sets are intervals. Conversely, in multidimensional spaces, it is immediate to state whether any given point belongs to $\hat{L}(\lambda^*)$, while saying how many connected sets comprise $\hat{L}(\lambda^*)$, and identifying them is not obvious. The issue is usually addressed by borrowing tools from graph theory. As the interest primarily focuses on allocating observations to clusters (instead of identifying a partition of the whole sample space), first, a suitable graph $\mathcal{G}$ is built, having vertices $x_1, \ldots, x_n$. Then, the subgraph $\mathcal{G}_{\lambda^*}$, induced by the sample level set

$$S(\lambda^*) = \{x_i \in (x_1, \ldots, x_n) : \hat{f}(x_i) \geq \lambda^*\},$$

is formed by removing from $\mathcal{G}$ all the vertices not in $S(\lambda^*)$ and all edges with at least one vertex among them. Finally, the connected components of $\mathcal{G}_{\lambda^*}$ are easily identified as the sets of observations that are pairwise connected through an edge; see, for example, Cormen *et al.* (2001) for an overview of the algorithms designed for this task in graph analysis. The problem of identifying connected sets of $\hat{L}(\lambda^*)$ is then simplified by shifting the target from a continuous multidimensional space to a finite and discrete set. A key matter becomes to suitably define the graph $\mathcal{G}$.

A common choice is the nearest-neighbour graph, discussed by Cuevas *et al.* (2000, 2001) and Stuetzle (2003) and, in some variant forms, by Rinaldo & Wasserman (2010). These works have a number of distinct original contributions that stand out. Cuevas *et al.* (2000, 2001) exploit bootstrap ideas to obtain more accurate estimates of the level sets and assess the variability of the estimates. Rinaldo & Wasserman (2010) provide a strong toolkit of mathematical results by establishing convergence rates for cluster estimators as well as the conditions that f must satisfy for applying modal clustering. Also, they study the stability properties of the level-set clustering. Unfortunately, these results and those of Cuevas *et al.* (2000, 2001) depend on λ^*, which limits considerably the reach of these contributions. This is not the case for Stuetzle (2003) who takes advantage of a connection between the minimum spanning tree recovered from the nearest-neighbour graph and the nearest-neighbour estimator of f. As this estimator tends to identify a large number of spurious modes, the author introduces a method for their pruning. See also the next section.

Azzalini & Torelli (2007) make use of the Delaunay triangulation, the graph associated to the Voronoi tessellation of the sample space. The Voronoi diagram partitions $\mathbb{R}^d$ in n regions such that each region contains exactly one observation and every point in that region is closer to its generating observation than to any other. The Delaunay graph connects through an edge a pair of observations when the corresponding regions of the Voronoi tessellation share a portion of their boundary facets.

Slightly different in spirit are the works of Stuetzle & Nugent (2010) and Menardi & Azzalini (2014), which represent an advancement of Stuetzle (2003) and Azzalini & Torelli (2007), respectively but exploit the density features to build $\mathcal{G}$. Both works build on the idea of setting

edges between vertices when $\hat{f}$ does not exhibit any valley along the segment joining them. A similar idea was first examined by Burman & Polonik (2009).

Another approach to detect high-density clusters is the one of Ooi (2002), who exploits the local density to recursively partition the sample space in homogeneous regions in the guise of classification and regression tree methodology. The regions are then merged back to form clusters by combining high-density adjacent regions.

2.3.1 The cluster tree

As already outlined, there may not exist a single value of λ^* such that each mode of the density function belongs to a distinct connected subset of $L(\lambda^*)$. Figure 3**F3** shows a simple bivariate example of this phenomenon: while the density function identifies three modes (i.e. three clusters), none of the associated level sets are formed by three connected components.

In fact, there is no need to focus on a single λ, as outlined by Hartigan (1975) by introducing the notion of *cluster tree*. This is a hierarchical structure that counts the number of connected components of the level sets as λ varies. Recalling the example in Figure 3, denote by m_1, m_2 and m_3 the three modes of the density function, ordered from the lowest to highest density ones. Starting from $\lambda = \lambda_0 = 0$, the level set consists of a single connected component (the whole support of f), and the cluster tree then shows one branch. Then, λ is increased up to λ_1, at which the level set splits into two components and then results in the cluster tree with two different branches. One of these components, C_1' say, is univocally associated to m_1. It represents a first *core* of a cluster, while the two highest modes are both included in the other component C_2'. At a higher level $\lambda = \lambda_2 (= f(m_1))$, C_2' remains, while C_1' vanishes. Thus, the cluster tree counts one branch only for all $\lambda_2 \leq \lambda \leq \lambda_3$. At $\lambda = \lambda_3$, C_2' splits into the two components C_2'' and C_3'', each of them finally associated with m_2 and, respectively, m_3. This is reflected in the cluster tree, which has again two branches.

In building the cluster tree, three cluster cores, C_1', C_2'' and C_3'', are finally detected. These are defined as the largest level-set connected components, which include one mode only. As these regions do not represent a partition of the sample space, unallocated points, referred to as *fluff* in Stuetzle & Nugent (2010), can be assigned to cluster cores by some kind of supervised classification. Stuetzle & Nugent (2010) and Azzalini & Torelli (2007), for instance, adopt a recursive, yet different, procedure in both cases based on assigning each fluff point to the most likely cluster in terms of density.

Estimates of the cluster tree are usually obtained by replaying the aforementioned procedures to estimate the connected components of $L(\lambda)$, for a grid of λ values.

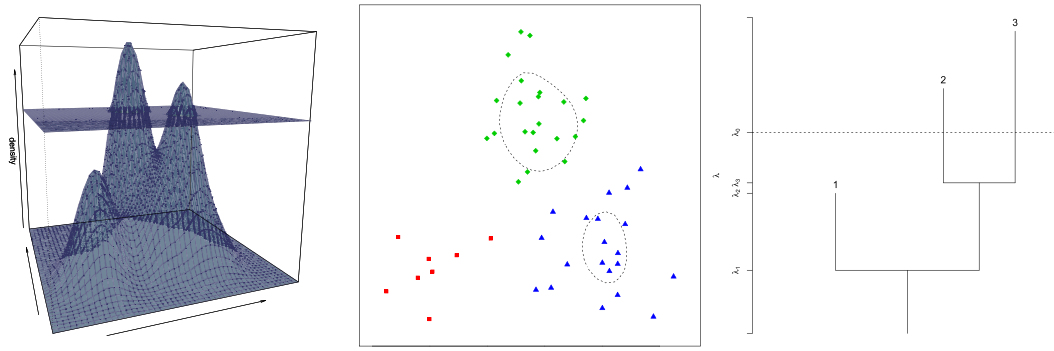

Figure 3. *A section of the density function in Figure 2 at a level λ_0 and the identified level set (middle panel), formed by two disconnected regions. In the right panel, the associated cluster tree. The horizontal dashed line corresponds to the level λ_0.*

So far, some work has been carried out to understand the properties of a cluster tree estimator. A natural definition of consistency was first given by Hartigan (1981). He defined as consistent a cluster tree estimator such that, for every pair of disjoint sets C_1 and C_2 in the true underlying cluster tree, the smallest estimated clusters containing all the sample points in C_1 and C_2 tend in probability to be themselves disjoint. Wong & Lane (1983) derived this property for the plug-in estimator $\hat{L}(\lambda)$, provided that a uniformly strongly consistent estimate $\hat{f}$ is adopted. Further references are Tsybakov (1997), Walther (1997), Cuevas *et al.* (2000), Baíllo *et al.* (2000), Baíllo *et al.* (2001), Baíllo (2003), Rigollet & Vert (2009) and Mason & Polonik (2009), and references therein, which delineate conditions, rates of convergence and further asymptotic results for a level-set estimation problem. These results extend to the estimate of the cluster tree of $\hat{f}$. However, owing to the difficulties in recovering the connected components of a density level set in $\mathbb{R}^d$, we must resort to an approximation of the cluster tree. Hence, in general, we do not have guarantee about the preservation of the consistency property. Recently, some authors have addressed this question for a single-level set (Maier *et al.*, 2009; Rinaldo and Wasserman, 2010). Unfortunately, their results depend on some parameters related to density estimate and whose optimal value varies for different λ.

Some works addressing the problem of determining the properties of the whole tree have focused on the joint use of a k-nearest neighbour $\hat{f}$ and the associated k-nearest neighbour graph ($k \geq 1$). Hartigan (1981) argues the non-consistency for $k = 1$ and recalls the conditions for a weaker notion of '*fractional*' consistency. Chaudhuri *et al.* (2014) continue on the same direction showing the strong consistency of a generalised version of the estimator, derived from Wishart (1969), and prove consistency for $k > 1$.

At the end of this section, we mention some variant forms of the cluster tree, which help to gain insights into the features of a multivariate density function. These are the *mode function* (Azzalini & Torelli, 2007), the *volume* and the *barycentre* plots (Klemelä, 2004).

2.4 Related Works

It might have surprised that, so far, there has been no mention of density-based spatial clustering of applications with noise (DBSCAN) (Ester *et al.*, 1996), often considered as the density-based clustering method par excellence. In fact, the notion of density adopted in the original work of Ester *et al.* (1996) is not explicitly specified as the one of probability theory considered here, neither conceptually nor operationally. The main idea behind DBSCAN is to cluster together points lying within a given distance, when there is at least a given number data points in this neighbourhood. Thus, we rather frame DBSCAN, and all its followers (among many others, Guha *et al.*, 1998; Ankerst *et al.*, 1999) in the class of methods *related* with the notion of density cluster. DBSCAN, and many of its descendants, was recasted to the probability framework only with hindsight, as it is a revisitation of the single-level method prompted by Wishart (1969). Density-Based Clustering (DENCLUE) (Hinneburg & Keim, 1998), for instance, extends the idea of DBSCAN by introducing the concept of influence function, which comprehensively appears as a rediscovery of kernel estimators. For a review of DBSCAN-related methods, see, for instance, Kriegel *et al.* (2011).

A further approach having close connections with mode hunting methods is the so-called scale-space clustering, developed in the area of computer vision to study the properties of blurred images. Scale-space theory represents a signal with a function f whose features of interest, for example, the extrema, are perceived as salient if they are stable over different scales. Even if framework and motivations are different, it is natural to associate these features with the clusters detected by mode hunting, based on density f (Chakravarthy & Ghosh, 1996; Leung *et al.*, 2000). The stability over blurring can be evaluated as the maintenance of the modes of

f when the smoothing parameter adopted for its estimation varies. Note the resemblance with the idea of hierarchy of clusters mentioned by (Li *et al.*, 2007).

Valley seeking methods, whom Koontz & Fukunaga (1972) represent the forerunners, follow the reverse perspective of modal clustering and allocate data into different clusters when there is evidence of a density valley between them. In Ertoz *et al.* (2002), this idea is implemented by counting the number of shared neighbours of two observations.

A further approach having a connection with modal clustering is *support vector* clustering (Ben-Hur *et al.*, 2001). According to this formulation, a function is defined, which maps the data into a higher dimensional space where the smallest sphere including all the observations is sought. When mapped back to the sample space, the sphere takes on a non-linear shape, and its connected components are regarded as clusters. By construction, any path connecting two observations in different clusters goes through a region of the sample space not belonging to the support of the density underlying the data.

Another class of methods sharing some (operational) ideas with modal clustering is *spectral* clustering (Von Luxburg, 2007), which constructs a similarity graph on the observations, for example, a Gaussian similarity function or one based on nearest neighbours. Then, the Laplacian matrix of the similarity graph is built and its eigen-decomposition computed. The eigenvectors corresponding to the k-largest eigenvalues are representatives of the clusters, to which data are allocated according to a measure of distance.

3 Discussion

3.1 Why Modal Clustering? Why not a Good Reason?

Perhaps the main strength of modal clustering lies in its attempt to circumscribe the ill posedness of the clustering problem. Assuming a true (albeit unknown) population structure allows for the definition of a 'ground truth', representing the ideal partition that clustering methods should try to approximate and can be used as a benchmark to evaluate a clustering configuration or to compare alternatives.

An appealing implication of linking the notion of cluster to a characteristic of the underlying probability distribution is that the number of clusters is an intrinsic property of the data generator mechanism, thereby well defined, at least conceptually, yet unknown. Hence, its determination is an integral part of the estimation procedure. Unlike distance-based methods, no detection of clusters is also possible, when the density is unimodal.

The adopted notion of cluster is also particularly attractive as it is not bound to a particular shape. Its complying with geometric intuition makes it close to a 'natural' grouping of data. Operationally, the use of a non-parametric estimator of the density function allows to maintain this flexibility. This property is not shared by other methods: in the model-based paradigm, there is no bijection between the components of a mixture model and the bumps of a density function (Ray & Lindsay, 2005); an example of this behaviour is illustrated in Figure 4(e). Additionally, assuming a parametric form for each cluster biases the groups to have a predetermined shape. In the traditional distance-based framework, there is no *a priori* limitation about the shape of the clusters, but problems as the chaining effect of single-linkage or the bias toward spherical clusters typical of k-means are well known.

A fortunate side effect of the adopted notion of cluster derives from the structure of the cluster tree, which allows for catching possible different degrees of resolution of the clustering structure. In the example illustrated in Figure 3, for instance, the number of modes is three, but two of them (previously denoted as m_2 and m_3) pertain to the same macro-group, at a lower level of resolution. The cluster tree describes this situation as the leaves associated to the two

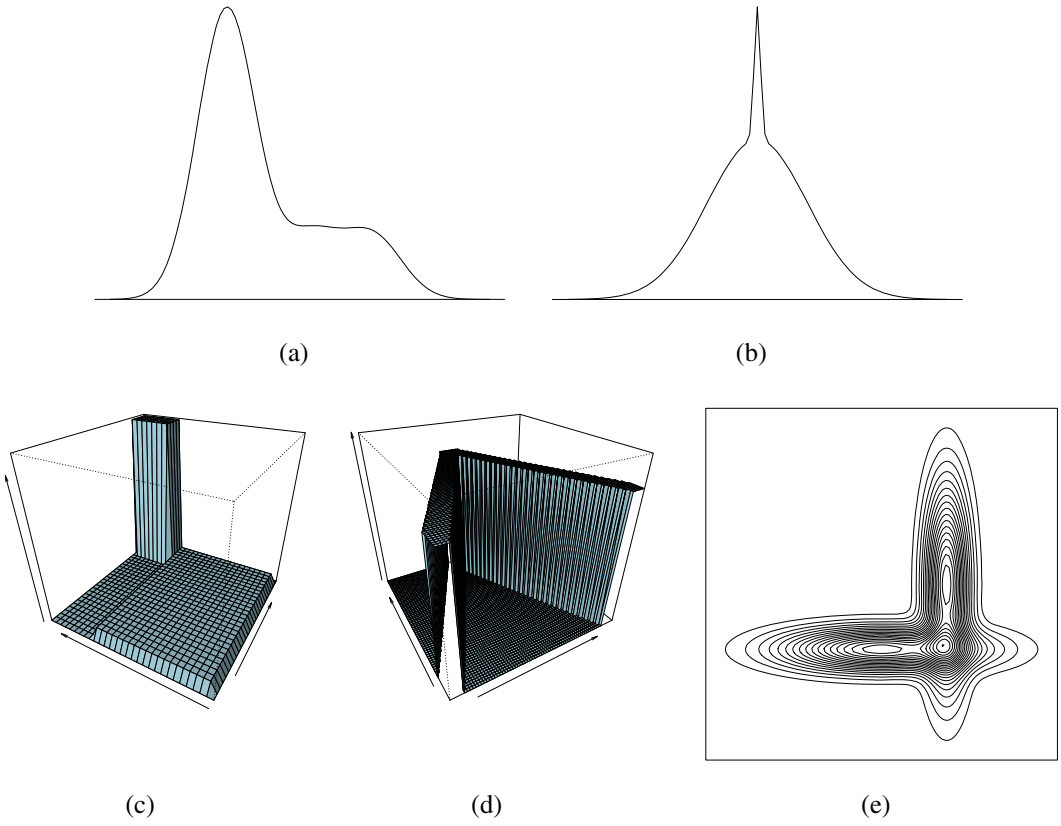

Figure 4. *Examples drawn from Hennig (2010) (a, b, c), Stuetzle (2003) (d) and Ray and Lindsay (2005) (e) (up to slight modifications) where the modal notion of cluster mismatches the human intuition.*

groups collapse to a single branch. In this sense, the cluster tree represents a somewhat formal instrument to emulate the human eye, and then once more, it strengthens the complying of the cluster concept with a 'natural' grouping of data.

Finally, a probabilistic notion of cluster allows for providing each observation with a degree of confidence of belonging to the clusters. Maximum confidence is associated to high-density observations around the detected modes, while observations lying at the tails or along the valleys of the density function have a lower confidence. Evaluating such confidence may result in the opportunity of soft clustering schemes or may be exploited for diagnostic purposes.

However, it must be bore in mind that modal clustering cannot be the panacea for the clustering problem. Despite its attractiveness, there is a number of situations where this approach is not appropriate, even conceptually. This occurs essentially when either the density function is not defined or the domain of attraction of the modes of f looks inadequate to describe the concept of cluster one has in mind.

Regarding the conditions that the density function must satisfy in order to allow the application of modal clustering, often, some regularity assumptions (as continuity or differentiability) are made in order to avoid complications or to guarantee accurate results. In fact, Rinaldo & Wasserman (2010) show that it is possible to deal not only with discontinuous densities but also with distributions with atoms, thereby not admitting a Lebesgue density at all.

On the other hand, a common situation that cannot be considered in the range of modal clustering applications occurs when data are of non-continuous type as some or all the observed

variables are highly discretised or even categorical. A possible extension of modal clustering to the case of categorical data has been discussed by Giordan & Diana (2011), which, unfortunately, consider the case of categorical ordered data only. So far, it seems that the only viable route is to apply some preliminary transformations to the data prior to the use of modal clustering, in the guise of tandem analysis.

Further situations where modal clustering is not necessarily appropriate are those where the modes do not describe adequately the idea of clusters one may have in mind. Some examples are illustrated in Figure 4. In all these examples, there might not be agreement about the clustering configuration, as the perceived clusters do not manifest themselves as modes. In Figure 4(a–d), someone could recognise clusters that are not associated with a mode of the density: the hitch is that there is no separating 'gap' between dense patterns, which is required to define a mode. Conversely, in Figure 4(e), the three-modal example that was first advocated as a reason for choosing modal clustering over the model-based approach might, in fact, suggest the presence of two groups, as the human eye perceives the two mixture components at a first glance. However, it is worth to notice that these situations are rather peculiar, and difficult to be framed whatever cluster notion one decides to adhere to.

3.2 Density Estimation

Modal clustering is governed by the estimate of the density function, on which detection of high-density regions relies. Depending on the clustering method, the choice of the specific density estimator is typically driven by some kind of convenience, either conceptual or operational. The mean-shift algorithm, for example, is based on the kernel density estimator because of its interpretation in terms of gradient ascent algorithm. In fact, the kernel density estimator is the most diffusely adopted in modal clustering, having the advantage of being intuitive and relatively simple to analyse mathematically. The nearest-neighbour estimator is also frequently adopted in the level set-based methods, in conjunction with the use of a nearest-neighbour graph for the connected components detection.

Disregarding the specific choice adopted, a non-parametric estimator is usually governed by some parameters defining the amount of smoothing. In kernel density estimation, this task is performed by the bandwidth matrix; the number of nearest neighbours plays a similar role in k-nearest neighbour estimates. Similarly, the number of summands determines the amount of smoothing in orthogonal series estimators. In the following discussion, we focus on kernel estimator (1), unless differently specified.

One claim of modal clustering is that the number of clusters, that is, the number of modes, is conceptually well defined and then, usually, an integral part of the estimation process. In fact, one could object that the number modes of the estimated density depends on the selected amount of smoothing. How to set the amount of smoothing is then an issue to be tailored, in principle not that different from choosing the number of clusters, as required by distance-based methods.

In fact, the choice of the smoothing parameters is less arbitrary than the corresponding choice of the number of clusters. In principle, the problem could be addressed by taking advantage of the rich literature about bandwidth selection in density estimation. Consolidated methods for multivariate densities are based on plug-in (Wand & Jones, 1994; Duong & Hazelton, 2003; Chacón & Duong, 2010) or cross-validation rules (Sain *et al.*, 1994; Duong & Hazelton, 2005). On the other hand, the final aim is not density estimation, and some authors (among others, Cuevas *et al.*, 2001) argue that the choice of the bandwidth should be tailored specifically for the task of clustering. In fact, only a few attempts have been made in this direction. Somewhat related to this topic is the work of Hall *et al.* (2004) where theoretical and numerical evidence

is provided about the extent to which the number of modes is a non-monotone function of bandwidth in the case of general compactly supported densities. However, it is shown that the main effects of non-monotonicity occur for relatively small bandwidths and have negligible impact on many aspects of bump hunting. For the more specific goal of providing suggestions on the choice of the bandwidth, we shall distinguish once more between mode hunting and level set-based methods. In the former case, the focus is on determining the modes of the density function and the gradient ascent paths of each observation. Then, the bandwidth should be optimal for derivative estimation. In the latter case, an optimal smoothing matrix should be able to provide an accurate approximation of the density level sets.

In the context of mean-shift clustering, Comaniciu (2003) presents a bandwidth selection method producing a stable estimate of the second-order moments of the data distribution across different scales. Einbeck (2011) considers favourably a bandwidth if a high proportion of data points falls within circles centred at the modes. Chacón & Duong (2013) present three methods for unconstrained smoothing matrix selection aimed at derivative estimation.

In the context of level set-based methods, the situation is less favourable, at least from an operational point of view, as most of literature focuses on theoretical results. Among them, we like to mention the contribution of Samworth & Wand (2010) who propose a plug-in bandwidth selection rule tailored for high-density regions estimation and possessing attractive asymptotic properties. Unfortunately, they focus on the univariate case only and on level sets associated to a single, given probability amount. Rinaldo & Wasserman (2010) consider the multivariate setting and present two bandwidth selection methods: one is based on the maximisation of an estimate of the excess mass functional; the other one selects the bandwidth that maximises a measure of cluster stability. In fact, the authors acknowledge a substantial limit of both the procedures, which are designed for the estimation of a single-level set and then produce values depending on λ. A different line of thought is expressed in Menardi & Azzalini (2014), believing that the problem of bandwidth selection is less influential than one could expect. As density estimation is merely an intermediate step for the subsequent detection of connected level sets, a rough indication of the location and shapes of high-density regions may be adequate, and this is provided by most sensible methods for bandwidth selection. In some way, this is acknowledged by Rinaldo & Wasserman (2010) concurring that even if a biased level-set estimator is adopted, the bias may not be of a great practical concern as it may contain a small amount of mass.

It is widely recognised that the problem is controversial, and it seems difficult to draw out some practical indications to address the selection of the smoothing amount for clustering purposes. The only attempt in this direction has been pursued by Chacón & Monfort (2013) who have compared the performance of a number of methods for bandwidth selection and alternative bandwidth parametrisations in mean-shift clustering. Unfortunately, none of the bandwidth selectors they consider turn out to outperform the others, and only cautious recommendations arise from their study. Here, we try to make a larger-scale effort in running a simulation, to study the behaviour of modal clustering as the window width varies. In the work that this program entails, we focus on the following choices:

- Clustering methods: we consider one representative for each of the two main approaches to modal clustering: the mean-shift algorithm for mode hunting and the method of Azzalini & Torelli (2007) as a level set-based approach.
- Bandwidth selection methods: owing to the high computational burden for bandwidth selection even in moderately low dimensions, we restrict on diagonal parametrisation of the smoothing matrix and consider the following: (i) a multivariate generalisation of the plug-in rule introduced by Wand & Jones (1994), HPI; (ii) a least square cross-validation selector (multivariate generalisation of Bowman, 1984), HCV; and (iii) the asymptotically optimal

bandwidth under the assumption of multivariate normality (rule of thumb adopted by Azzalini and Torelli, 2007), HN. Clustering is run by setting $H = c \cdot H_{opt}$, with $H_{opt} \in \{HPI, HCV, HN\}$ and c a constant varying in $(0, 5)$. Matrices H_{opt} are selected as optimal for derivative estimation, when running mean-shift clustering and for density estimation when running level-set based clustering.
- True models: simulation settings have been defined in $\mathbb{R}^2$ and $\mathbb{R}^5$ and describe the following situations: (i) well separated, spherical groups; (ii) partially overlapping, spherical groups; (iii) non-convex-shaped groups; and (iv) imbalanced, elliptical groups. Details are reported in the Appendix.
- Evaluation measure: the identified partitions have been compared with the true clustering in terms of *adjusted Rand index* (Hubert & Arabie, 1985).

Although we consider a smaller number of bandwidth selectors and a simpler matrix parametrisation than Chacón & Monfort (2013), our settings allow us to draw some indications about a possible different role of bandwidth selection in the two modal clustering approaches, on the stability of results across different scales of the bandwidths as well as for different clustering configurations and dimensionality of the sample space. Results are displayed in Figures 5 and 6. The overall comment is that modal clustering can generally learn adequately the structure in data. Results are not invariant to the choice of the amount of smoothing, yet they are rather stable to different choices. While in all the considered settings data are definitely not Gaussian, the optimal bandwidth for Gaussian density seems to be quite a safe selection. For all the considered settings, the maximum average adjusted Rand index is obtained for c close to one in $H = c \cdot HN$, and the robustness of this result as c departs from one depends on the complexity of the clustering structure. Bandwidth selection based on plug-in or cross-validation rules clearly produces undersmoothed densities. Mean-shift and levels set-based clustering tend to produce comparable results, with a preference for the former approach in the imbalanced setting.

A further aspect related to density estimation, worth to be accounted for, is the dimensionality of the problem at hand. Nowadays, the data gathered in many scientific domains are high dimensional, or even the number of available observations is smaller than the number of variables. The curse of dimensionality is known to have a strong impact on non-parametric density estimators, and this explains a worsened behaviour of modal clustering. In high dimensions, much of the probability mass flows to the tails of the data density, possibly giving rise to the birth of spurious clusters and averaging away features in the highest density regions.

Some authors believe that reliable estimates of the density can be obtained in as many as six dimensions (Scott & Sain, 2005), and Scott (1992, p. 196) warns that 'in order to include sufficient data, the smoothing parameters must be so large that no local behaviour of the function can be reasonably approximated without astronomical sample sizes'. While these arguments should discourage from the application of density-based clustering on high-dimensional data, the situation is not as critical as it appears at first. In high dimensions, non-parametric estimates can still be used to coarsely describe the data structure, and often, allowing different amounts of smoothing depending on the characteristics of the data is advisable (Menardi & Azzalini, 2014; Stuetzle & Nugent, 2010). Statistical tests can be applied to establish significance of the density modes (Burman & Polonik, 2009; Genovese *et al.*, 2013; Cheng & Ray, 2014a), and mechanisms for mode pruning (Stuetzle, 2003; Stuetzle & Nugent, 2010) or cluster merging (Li *et al.*, 2007) can be helpful for addressing the appearance of spurious clusters.

Indeed, in spite of a quite low expectation about the behaviour of modal clustering in high-dimensional spaces, fair results have been reported in the application of modal clustering in several tens or even hundreds of dimensions (see, e.g. Stuetzle and Nugent, 2010).

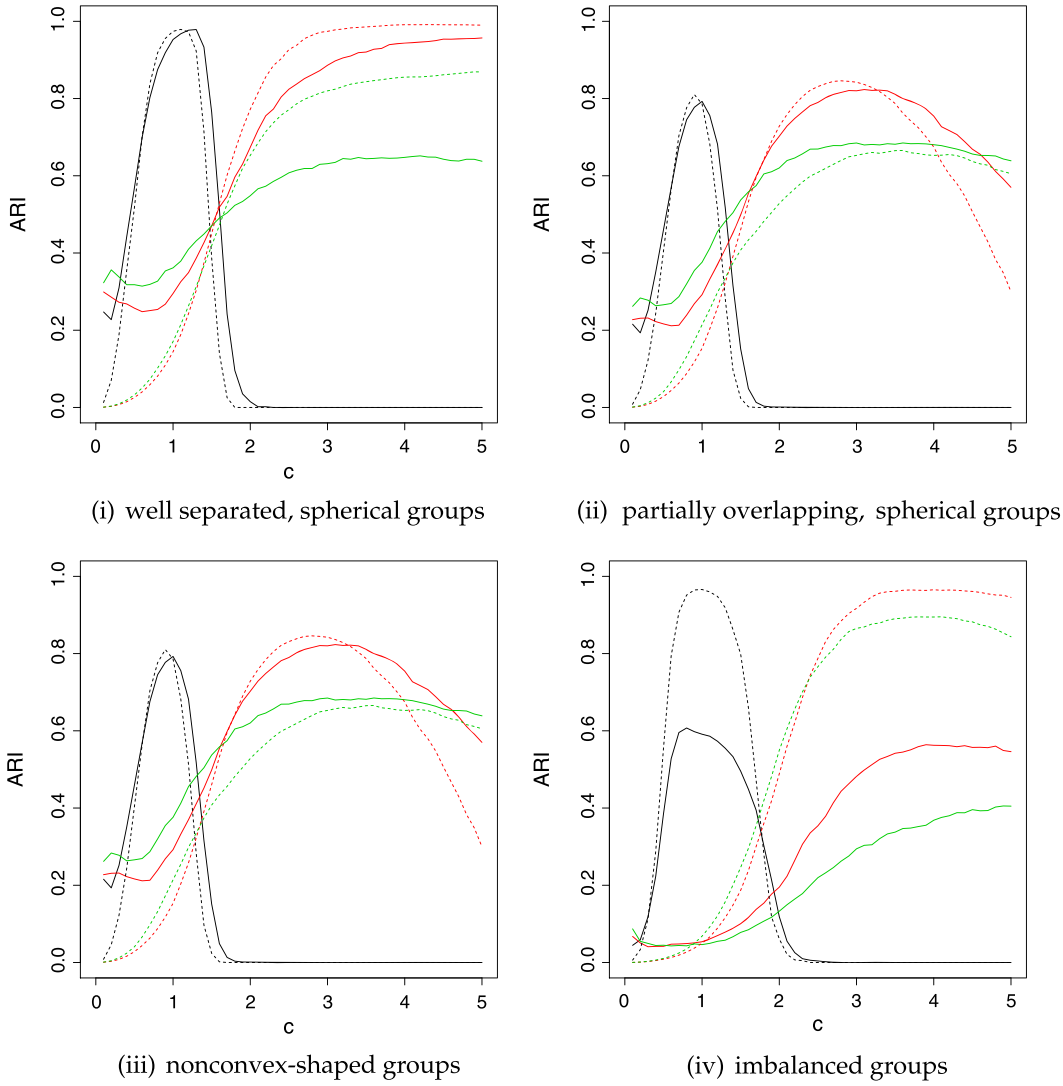

Figure 5. *Mean ARI across simulations as the smoothing matrix is multiplied for* $c \in (0, 5)$. *Solid and dashed lines distinguish between level set-based clustering and, respectively, mode-seeking clustering. Different colours correspond to different bandwidth selectors: black for HN, red for HPI and green for HCV. Data have been generated in* $\mathbb{R}^2$.

That said, the feeling is that modal clustering is not designed for working in high dimensions. On the other hand, as a typical aspect of high-dimensional data is the tendency to fall into manifolds of lower dimension, dimension reduction methods are often advisable. The issue is certainly worth to be further investigated.

3.3 Computational Complexity of Modal Clustering

A major weakness of modal clustering lies in its computational complexity, which is also the reason that this approach was initially set aside after its first formulations. Nowadays, the current computational advances allow to handle rather big-data problems, but the number of required operations to run modal clustering is still of some relevance, preventing its application to various fields.

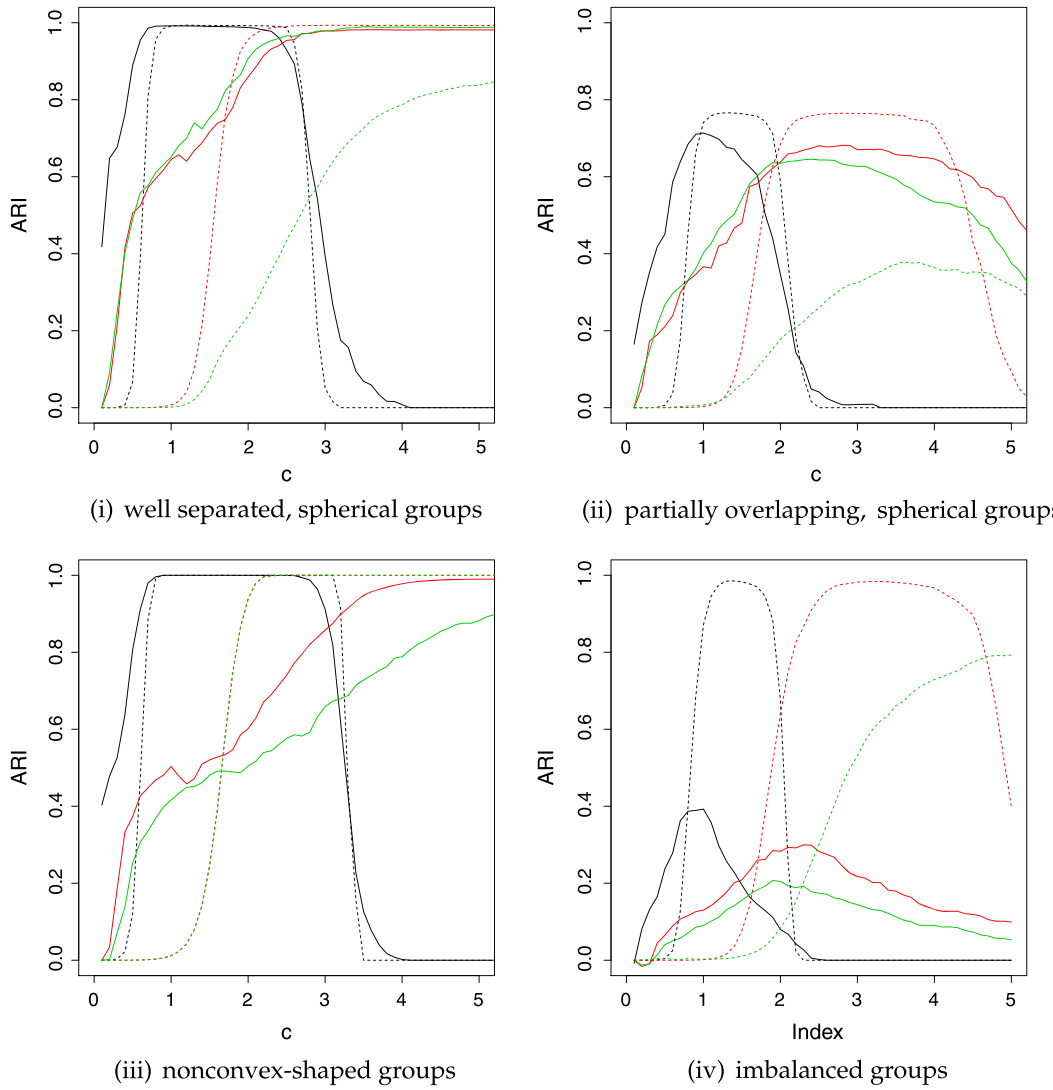

Figure 6. *Compare with Figure 5. Data have been generated in* $\mathbb{R}^5$.

Whatever specific method is selected, the minimum requirement is the evaluation of the density function at each observation. Hence, when a kernel estimator is adopted, the number of required operations grows quadratically with the sample size, because of the sum of n kernels. Depending on the choice of the multivariate kernel, each of these latter evaluations depend on the data dimension. A radially symmetric kernel requires $O(d^2n)$ operations while a product kernel requires $O(dn)$ operations.

Modal-seeking algorithms further need to be iterated until convergence to the modes. Level-set methods are even computationally higher demanding as, in addition to density estimation, detection of connected components is required for a grid of values of λ. This is strongly dependent on the selected method. The Delaunay triangulation, adopted by Azzalini & Torelli (2007), requires a number of operations, which increases exponentially with d for $d > 3$. On the other hand, it is strongly competitive in low-dimensional spaces as it requires $O(n \log n)$ operations

when $d \leq 3$. The procedures proposed by Stuetzle (2003), Stuetzle & Nugent (2010) and Menardi & Azzalini (2014) require $O(dn^2)$ operations, in addition to those mentioned earlier.

These limitations often force us to run some correctives when n and/or d are especially large. High-dimensional spaces are advisably reduced by feature selection or subspace projection methods, which also provide a safeguard against the curse of dimensionality (e.g. Menardi and Torelli, 2013, Paris and Durand, 2007, Lee and Li, 2012). A common way of proceeding to speed up computations due to a large n is to exclude, somehow, some observations from the computations. For example, a binned density estimator may be adopted; alternatively, a subset of observations may be used for cluster detection, and the remaining observations are classified to the cluster presenting the maximum likelihood. Li *et al.* (2007) propose to cluster the means identified by running k-means, for some large k, prior to clustering. Several adjustments to speed up the mean-shift procedures have been proposed (Yang *et al.*, 2003a, Yang et al., 2003b), also specifically for some applications such as image segmentation (e.g., Lebourgeois *et al.*, 2013).

4 Software

Thankfully, the main statistical environments and software offer tools to run most of existing algorithms for modal clustering. These are listed in the following.

As mean-shift clustering is rather widespread across different scientific communities, it is perhaps the most implemented by various languages and environments: we recall the R package `LPCM` (Einbeck & Evers, 2013) and function `MeanShift` in the Python library `scikit-learn`. There are also some implementations in Java, as, for example, the `Aiphial` project. Function `MeanShiftCluster` and the library `EDISON` are available for Matlab, the latter library being an interface for the homonym C++ library. Within the set of clustering methods performing mode seeking, we also recall the R package `Modalclust` (Cheng & Ray, 2014b), which performs the method of Li *et al.* (2007) and its parallelised implementation, which timely speeds up computations.

Level-set based methods have been mostly studied and developed within the statistical community and are thereby available especially in the R computing environment. Package `pdfCluster` (Azzalini & Menardi, 2014) implements the methods proposed by Azzalini & Torelli (2007) and Menardi & Azzalini (2014). Package `gslclust` implements generalised single-linkage clustering, proposed by Stuetzle (2003) and Nugent & Stuetzle (2010) and available at http://www.stat.washington.edu/wxs/software.htm. Related to these, the `hdrcde` package provides, among others, tools for computation of highest density regions in one and two dimensions (Hyndman, 2013). Python package `DeBaCl` offers functions for interactive cluster tree estimation and implements the recent algorithm of Chaudhuri & Dasgupta (2010). Some modifications of early clustering works based on levels set detection, such as the one of Wong & Lane (1983), are implemented by `PROC MODECLUS` in the SAS language.

Density-based related method are also variously implemented: DBSCAN (Ester *et al.*, 1996) is performed by the R package `fpc` (Hennig, 2014), by the Python library `scikit-learn` and by various Java and Matlab codes. The `SVC` toolbox provides tools for support vector clustering.

5 Final remarks

Clustering is a complicated task that cannot be performed without subject matter considerations and without human intervention. Bearing always in mind this, modal clustering represents a smart way to limit the arbitrariness of the problem. Beyond its conceptual attractiveness, it

is computationally competitive with standard methods, and its application to a number of real and complex domains has shown interesting results (see, among others, Jang, 2006, Tao *et al.*, 2007, De Bin and Risso, 2011, Contreras-Reyes and Azzalini, 2012). Also, inherent research has recently spread beyond the specific boundaries of the outlined clustering problem to address different, related directions. For instance, some ideas of modal clustering have been extended to find the local maxima in estimation problems of regression (Yao *et al.*, 2012) or, in the Bayesian framework, to provide a solution to the label switching problem (Yao & Lindsay, 2009).

Many questions about modal clustering are still open and require further research. A more rigorous mathematical framework would be hoped, able to define a wide range of density functions that make sense for clustering purposes, and also including those with a non-standard behaviour (this would overcome, for instance, the limits of Morse theory as a background for this setting). Operationally, interesting challenges regard the choice of smoothing parameters for the density estimate as well as a better understanding of modal clustering behaviour in high dimensions. Related to this, there is a need to address the problem of dimension reduction specifically for modal clustering. From a computational point of view, the new challenges one has to face with big data have to be adequately considered to improve the efficiency of clustering.

Acknowledgements

Some points discussed in this paper were first presented at the International Federation of Classification Societies (IFCS) conference held in Tilburg in 2013. I would like to thank Christian Hennig for the useful remarks he prompted in that occasion: most of them have been incorporated in this paper. I am also grateful to Adelchi Azzalini for reading the manuscript and providing causes for reflection.

References

Ankerst, M., Breunig, M.M., Kriegel, H. & Sander, J. (1999). Optics: Ordering points to identify the clustering structure. In *ACM Sigmod Record*, Vol. 28: ACM; 49–60.

Azzalini, A. & Menardi, G. (2014). Clustering via nonparametric density estimation: The R package pdfCluster. *J. Stat. Softw.*, **57**(11), 1–26.

Azzalini, A. & Torelli, N. (2007). Clustering via nonparametric density estimation. *Stat. Comput.*, **17**, 71–80.

Baíllo, A. (2003). Total error in a plug-in estimator of level sets. *Statist. Probab. Lett.*, **65**(4), 411–417.

Baíllo, A., Cuevas, A. & Justel, A. (2000). Set estimation and nonparametric detection. *Canad. J. Statist.*, **28**(4), 765–782.

Baíllo, A., Cuesta-Albertos, J.A. & Cuevas, A. (2001). Convergence rates in nonparametric estimation of level sets. *Statist. Probab. Lett.*, **53**(1), 27–35.

Ben-Hur, A., Horn, D., Siegelmann, H.T. & Vapnik, V. (2001). Support vector clustering. *J. Mach. Learn. Res.*, **2**, 125–137.

Bowman, A.W. (1984). An alternative method of cross-validation for the smoothing of density estimates. *Biometrika*, **71**(2), 353–360.

Burman, P. & Polonik, W. (2009). Multivariate mode hunting: Data analytic tools with measures of significance. *J. Multivar. Anal.*, **100**(6), 1198–1218.

Carmichael, J.W., George, J.A. & Julius, R.S. (1968). Finding natural clusters. *Syst. Zool.*, **17**(2), 144–150.

Carreira-Perpinan, M.A. (2007). Gaussian mean-shift is an EM algorithm. *IEEE Trans. Pattern Anal. Mach. Intell.*, **29**(5), 767–776.

Carreira-Perpinan, M.A. (2008). Generalised blurring mean-shift algorithms for nonparametric clustering. In *CVPR*: 2013 IEEE Conference on Computer Vision and Pattern Recognition; 1–8.

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

Chacón, J.E. (2014). A population background for nonparametric density-based clustering. ArXiv e-prints: 1408.1381.

Chacón, J.E. & Duong, T. (2010). Multivariate plug-in bandwidth selection with unconstrained pilot bandwidth matrices. *Test*, **19**(2), 375–398.

Chacón, J.E & Monfort, P. (2013). A comparison of bandwidth selectors for mean shift clustering. arXiv preprint:1310.7855.

Chakravarthy, S.V. & Ghosh, J. (1996). Scale-based clustering using the radial basis function network. *IEEE Trans. Neural Netw.*, **7**(5), 1250–1261.

Chaudhuri, K. & Dasgupta, S. (2010). Rates of convergence for the cluster tree. In *Advances in Neural Information Processing Systems,* pp. 343–351. USA: Curran Associates.

Chaudhuri, K., Dasgupta, S., Kpotufe, S. & von Luxburg, U. (2014). Consistent procedures for cluster tree estimation and pruning. arXiv preprint math.PR/0000000.

Cheng, Y. (1995). Mean shift, mode seeking, and clustering. *IEEE Trans. Pattern Anal. Mach. Intell.*, **17**(8), 790–799.

Cheng, Y. & Ray, S. (2014a). Multivariate modality inference using Gaussian kernel. *Open J. Statist.*, **4**(05), 419–434.

Cheng, Y. & Ray, S. (2014b). Parallel and hierarchical mode association clustering with an r package modalclust. *Open J. Statist.*, **4**(10), 826–836.

Comaniciu, D. (2003). An algorithm for data-driven bandwidth selection. *IEEE Trans. Pattern Anal. Mach. Intell.*, **25**(2), 281–288.

Comaniciu, D. & Meer, P. (2002). Mean shift: a robust approach toward feature space analysis. *IEEE Trans. Pattern Anal. Mach. Intell.*, **24**(5), 603–619.

Contreras-Reyes, J.E. & Azzalini, A. (2012). *On the spatial correlation between areas of high coseismic slip and aftershock clusters of the Maule earthquake Mw=8.8*. ArXiv e-prints: 1208.1517.

Cormen, T.H., Leiserson, C.E., Rivest, R.L. & Stein, C. (2001). *Introduction to Algorithms*, Vol. 2 Cambridge: MIT Press.

Cuevas, A., Febrero, M. & Fraiman, R. (2000). Estimating the number of clusters. *Canad. J. Statist.*, **28**, 367–382.

Cuevas, A., Febrero, M. & Fraiman, R. (2001). Cluster analysis: a further approach based on density estimation. *Comput. Stat. Data Anal.*, **36**, 441–459.

De Bin, R. & Risso, D. (2011). A novel approach to the clustering of microarray data via nonparametric density estimation. *BMC Bioinformatics*, **12**(49), 1–8.

Duong, T. & Hazelton, M. (2003). Plug-in bandwidth matrices for bivariate kernel density estimation. *J. Nonparametr. Stat.*, **15**(1), 17–30.

Duong, T. & Hazelton, M.L. (2005). Cross-validation bandwidth matrices for multivariate kernel density estimation. *Scand. J. Stat.*, **32**(3), 485–506.

Einbeck, J. (2011). Bandwidth selection for mean-shift based unsupervised learning techniques: a unified approach via self-coverage. *J. Pattern. Recogn. Res.*, **6**(2), 175–192.

Einbeck, J. & Evers, L. (2013). Lpcm: Local principal curve methods. Available at http://CRAN.R-project.org/package=LPCM. R package version 0.44-8.

Ertoz, L., Steinbach, M. & Kumar, V. (2002). A new shared nearest neighbor clustering algorithm and its applications. In *Workshop on Clustering High Dimensional Data and its Applications at 2nd Siam International Conference on Data Mining,* pp. 105–115. Arlington, USA.

Ester, M., Kriegel, H.P., Sander, J. & Xu, X. (1996). A density based algorithm for discovering clusters in large spatial databases withe noise. In *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96)*, pp. 226–231. Portland: AAAI Press.

Fraley, C. & Raftery, A. (1998). How many cluster? Which clustering method? Answers via model-based cluster analysis. *Comput. J.*, **41**, 578–588.

Fraley, C. & Raftery, A.E. (2002). Model-based clustering, discriminant analysis and density estimation. *J. Am. Stat. Assoc.*, **97**, 611–631.

Fukunaga, K. & Hostetler, L.D. (1975). The estimation of the gradient of a density function, with application in pattern recognition. *IEEE Trans. Inform. Theory*, **21**(1), 32–40.

Genovese, C., Perone-Pacifico, M., Verdinelli, I. & Wasserman, L. (2013). Nonparametric inference for density modes. ArXiv e-prints.

Giordan, M. & Diana, G. (2011). A clustering method for categorical ordinal data. *Commun. Stat.-Theor. Methods*, **40**(7), 1315–1334.

Guha, S., Rastogi, R. & Shim, K. (1998). Cure: an efficient clustering algorithm for large databases. In *ACM SIGMOD Record*, Vol. 27, pp. 73–84: Seattle: ACM.

Hall, P., Minnotte, M.C. & Zhang, C. (2004). Bump hunting with non-Gaussian kernels. *Ann. Stat.*, **32**(5), 2124–2141.

Hartigan, J.A. (1975). *Clustering Algorithms.* New York: J. Wiley & Sons.

Hartigan, J.A. (1981). Consistency of single linkage for high-density clusters. *J. Am. Stat. Assoc.*, **76**(374), 388–394.

Hennig, C. (2010). Methods for merging Gaussian mixture components. *Adv. Data Anal. Classif.*, **4**(1), 3–34.

Hennig, C. (2014). fpc: Flexible procedures for clustering. Available at http://CRAN.R-project.org/package=fpc. R package version 2.1-7.

Hinneburg, A. & Keim, D.A. (1998). An efficient approach to clustering in large multimedia databases with noise. In *Proceedings of the 4th International Conference of Knowledge Discovery and Data Mining (KDD',98)*, Vol. 98, pp. 58–65. New York.

Hubert, L. & Arabie, P. (1985). Comparing partitions. *J. Classif.*, **2**, 193–218.

Hyndman, R.J. (2013). *hdrcde. Highest Density Regions and Conditional Density Estimation. R Package*. Available at http://cran.r-project.org/package=pdfCluster.

Jang, W. (2006). Nonparametric density estimation and clustering in astronomical sky surveys. *Comput. Stat. Data Anal.*, **50**(3), 760–774.

Klemelä, J. (2004). Visualization of multivariate density estimates with level set trees. *J. Comput. Graph. Stat.*, **13**(3), 599–620.

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

Koontz, W. & Fukunaga, K. (1972). A nonparametric valley-seeking technique for cluster analysis. *IEEE Trans. Comput.*, **100**(2), 171–178.

Kriegel, H., Kröger, P., Sander, J. & Zimek, A. (2011). Density-based clustering. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, **1**(3), 231–240.

Lebourgeois, F., Drira, F., Gaceb, D. & Duong, J. (2013). Fast integral meanshift: Application to color segmentation of document images. In *12th International Conference on Document Analysis and Recognition (ICDAR)*, pp. 52–56. Washinton, DC: IEEE.

Lee, H. & Li, J. (2012). Variable selection for clustering by separability based on ridgelines. *J. Comput. Graph. Stat.*, **21**(2), 315–337.

Leung, Y., Zhang, J.S. & Xu, Z. (2000). Clustering by scale-space filtering. *IEEE Trans. Pattern Anal. Mach. Intell.*, **22**(12), 1396–1410.

Li, J., Ray, S. & Lindsay, B.G. (2007). A nonparametric statistical approach to clustering via mode identification. *J. Mach. Learn. Res.*, **8**, 1687–1723.

Maier, M., Heinb, M. & von Luxburg, U. (2009). Optimal construction of -nearest-neighbor graphs for identifying noisy clusters. *Theor. Comput. Sci.*, **410**(19), 1749–1764.

Mason, D.M. & Polonik, W. (2009). Asymptotic normality of plug-in level set estimates. *Ann. Appl. Probab.*, **19**(3), 1108–1142.

Matsumoto, Y. (2002). *An Introduction to Morse Theory.* Providence: American Mathematical Society.

McLachlan, G.J. & Basford, K.E. (1988). *Mixture Models*, Vol. 74. New York: Marcel Dekker.

Menardi, G. & Azzalini, A. (2014). An advancement in clustering via nonparametric density estimation. *Stat. Comput.*, **24**(5), 753–767.

Menardi, G. & Torelli, N. (2013). Reducing data dimension for cluster detection. *J. Stat. Comput. Sim.*, **83**(11), 2047–2063.

Nugent, R. & Stuetzle, W. (2010). Clustering with confidence: A low-dimensional binning approach. In *Classification as a Tool for Research*, Eds. H. Jocarek-Junge & C. Weihs, pp. 117–125. Berlin: Springer.

Ooi, H. (2002). Density visualization and mode hunting using trees. *J. Comput. Graph. Stat.*, **11**(2), 328–347.

Paris, S. & Durand, F. (2007). A topological approach to hierarchical segmentation using meanshift. In *Proceedings of the IEEE Conference Computer Vision and Pattern Recognition CVPR '07,* pp. 1–8. Minneapolis.

Ray, S. & Lindsay, B.G. (2005). The topography of multivariate normal mixtures. *Ann. Statist.*, **33**, 2042–2065.

Rigollet, P. & Vert, R. (2009). Optimal rates for plug-in estimators of density level sets. *Bernoulli*, **15**(4), 1154–1178.

Rinaldo, A. & Wasserman, L. (2010). Generalized density clustering. *Ann. Statist.*, **38**(5), 2678–2722.

Sain, S.R., Baggerly, K.A. & Scott, D.W. (1994). Cross-validation of multivariate densities. *J. Am. Stat. Assoc.*, **89**(427), 807–817.

Samworth, R.J. & Wand, M.P. (2010). Asymptotics and optimal bandwidth selection for highest density region estimation. *Ann. Statist.*, **38**(3), 1767–1792.

Scott, D. & Sain, S. (2005). Multidimensional density estimation. *Handbook Statist.*, **24**, 229–261.

Scott, D.W. (1992). *Multivariate Density Estimation: Theory, Practice and Visualization* New York: Wiley.

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

Stuetzle, W. & Nugent, R. (2010). A generalized single linkage method for estimating the cluster tree of a density. *J. Comput. Graph. Stat.*, **19**(2), 397–418.

Tao, W., Jin, H. & Zhang, Y. (2007). Color image segmentation based on mean shift and normalized cuts. *IEEE Trans Syst Man Cybern, Part B: Cybern.*, **37**(5), 1382–1389.

Tsybakov, A.B. (1997). On nonparametric estimation of density level sets. *Ann. Statist.*, **25**(3), 948–969.
Von Luxburg, U. (2007). A tutorial on spectral clustering. *Stat. Comp.*, **17**(4), 395–416.
Walther, G. (1997). Granulometric smoothing. *Ann. Statist.*, **25**(6), 2273–2299, 12.
Wand, M.P. & Jones, M.C. (1994). Multivariate plug-in bandwidth selection. *Comput. Stat.*, **9**(2), 97–116.
Wishart, D. (1969). Mode analysis: A generalization of nearest neighbor which reduces chaining effects. In *Numerical taxonomy*, Ed. A. J. Cole, pp. 282–308. London: Academic Press.
Wolfe, J.H. (1970). Pattern clustering by multivariate mixture analysis. *Multivar. Behav. Res.*, **5**(3), 329–350.
Wong, A.M. & Lane, T. (1983). The k-th nearest neighbour clustering procedure. *J. R. Stat. Soc. Series B*, **45**, 362–368.
Yang, C., Duraiswami, R., DeMenthon, D. & Davis, L. (2003a). Mean-shift analysis using quasinewton methods. In *Proceedings of the International Conference on Image Processing, (ICIP)*, Vol. 2. pp. II-447. Nice: IEEE.
Yang, C., Duraiswami, R., Gumerov, N.A. & Davis, L. (2003b). Improved fast gauss transform and efficient kernel density estimation. In *Proceedings of the ninth IEEE International Conference on Computer Vision*: IEEE, pp. 664–671.
Yao, W. & Lindsay, B.G. (2009). Bayesian mixture labeling by highest posterior density. *J. Am. Statist. Assoc.*, **104**(486), 758–767.
Yao, W., Lindsay, B.G. & Li, R. (2012). Local modal regression. *J. Nonparametr. Stat.*, **24**(3), 647–663.
Yuan, X., Hu, B. & He, R. (2012). Agglomerative mean-shift clustering. *IEEE Trans. Knowl. Data Engin.*, **24**(2), 209–219.

[*Received January 2015, accepted May 2015*]

Appendix A

The distributions from which samples have been drawn in the empirical study of Section 3.2 are listed subsequently. The following notation is used: $N_d(\mu, \Sigma)$ is the d-dimensional normal distribution with mean μ and variance Σ, $U_{\mathcal{S}}$ is a uniform distribution defined over the set $\mathcal{S}$, 1_d is the unitary vector with d components, I_d is the identity matrix of dimension d and vech(Σ) is the vector formed by the lower triangle of a symmetric matrix Σ. Sample sizes have been set to $n = 100$ or $n = 300$ in the $\mathbb{R}^2$ and, respectively, $\mathbb{R}^5$ scenarios, owing to a greater sparsity of data.

d = 2

(i) well separated, spherical groups
$\sum_{g=1}^{2} \pi_g N(\mu_g, \Sigma_g)$
$\mu_1 = (-3, 0)'$, $\mu_2 = (3, 0)'$, $\Sigma_1 = \Sigma_2 = I_2$, $\pi_1 = \pi_2 = 0.5$

(ii) partially overlapping, spherical groups
$\sum_{g=1}^{2} \pi_g N(\mu_g, \Sigma_g)$
$\mu_1 = (0, 0)'$, $\mu_2 = (4, 0)'$, $\Sigma_1 = \Sigma_2 = I_2$, $\pi_1 = \pi_2 = 0.5$

(iii) non-convex-shaped groups
$\sum_{g=1}^{2} \pi_g U_{\mathcal{S}_g}$
$\mathcal{S}_1 = \{z > 0 : \sum_{j=1}^{d} z_j^2 = 1\}$, $\mathcal{S}_2 = \{z < 0 : \sum_{j=1}^{d} z_j^2 = 1\}$, $\pi_1 = \pi_2 = 0.5$

(iv) imbalanced, elliptical groups
$\sum_{g=1}^{3} \pi_g N(\mu_g, \Sigma_g)$
$\mu_1 = (0, 0)'$, $\mu_2 = (3, 3)'$, $\Sigma_1 = I_2$, $\text{vech}(\Sigma_2) = (0.5, -0.5, 1)'$ $\pi_1 = 0.1$, $\pi_2 = 0.9$

d = 5

(i) well separated, spherical groups
$\sum_{g=1}^{2} \pi_g N(\mu_g, \Sigma_g)$
$\mu_1 = k \cdot 1_5$, $\mu_2 = -k \cdot 1_5$, with k such that $dist(\mu_1, \mu_2) = 6$ $\Sigma_1 = \Sigma_2 = I_5$, $\pi_1 = \pi_2 = 0.5$

(ii) partially overlapping, spherical groups
$\sum_{g=1}^{2} \pi_g N(\mu_g, \Sigma_g)$
$\mu_1 = k \cdot 1_5$, $\mu_2 = -k \cdot 1_5$, with k such that $dist(\mu_1, \mu_2) = 4$ $\Sigma_1 = \Sigma_2 = I_5$, $\pi_1 = \pi_2 = 0.5$

(iii) non-groups
$\sum_{g=1}^{2} \pi_g U_{\mathcal{S}_g}$, where $\mathcal{S}_1$ is the portion of five-dimensional unit torus having positive coordinates, $\mathcal{S}_2$ is similar with negative coordinates, $\pi_1 = \pi_2 = 0.5$.

(iv) imbalanced, elliptical groups
$\sum_{g=1}^{5} \pi_g N(\mu_g, \Sigma_g)$
$\mu_1 = 0 \cdot 1_5$, $\mu_2 = 2 \cdot 1_5$, $\Sigma_1 = 0.5 \cdot I_5$, $\text{vech}(\Sigma_2) = (0.5, -0.5, -0.5, 0, 0, 1, 0.5, 0, 0, 1, 0, 0, 1, 0, 1)'$ $\pi_1 = 0.1, \pi_2 = 0.9$