A fresh look at mean-shift based modal clustering

Abstract

Modal clustering is an unsupervised learning technique where cluster centers are identified as the local maxima of nonparametric probability density estimates. A natural algorithmic engine for the computation of these maxima is the *mean shift procedure*, which is essentially an iteratively computed chain of local means. We revisit this technique, focusing on its link to kernel density gradient estimation, in this course proposing a novel concept for bandwidth selection based on the concept of a critical bandwidth. Furthermore, in the one-dimensional case, an inverse version of the mean shift is developed to provide a novel approach for the estimation of antimodes, which is then used to identify cluster boundaries. A simulation study is provided which assesses, in the univariate case, the classification accuracy of the mean-shift based clustering approach. Three (univariate and multivariate) examples from the fields of philately, engineering, and imaging, illustrate how modal clusterings identified through mean shift based methods relate directly and naturally to physical properties of the data-generating system. Solutions are proposed to deal computationally efficiently with large data sets.

Keywords: multimodality; mode detection; kernel density estimation; modal testing; critical bandwidth.

MSC Classification: 62G07 , 62H12

1 Introduction

We begin this exposition with a thought experiment. Assume some population is distributed into some loosely scattered clusters over some area; think for instance of animals dispersed into small herds, or partygoers at some event. Assume also that conditions are such that for each element only a specific, fixed, radius around itself is visible, which is large enough to permit view on some neighboring elements but no overall view over the distribution of the full population.

Now, pick one element of this population, and assume this element intends to approach the next cluster center (say, in case of the animals, to protect from potential predators). What would be the best strategy for the element to go about this, given that they have no overall view over the distribution, nor do they know where the next cluster is centered?

A reasonable strategy could be as follows. While no information over the overall pattern is available, the element is aware of its local neighborhood and the distribution of other elements within. So, there does exist a notion of where the elements are denser in its close vicinity. Hence, the element could compute the average of all elements in its vicinity, which could be described as a local center of mass or gravity, and move to that point. Being there, the view has slightly shifted. The element now sees a slightly different neighborhood, albeit overlapping with the previous one. It is now able to compute an updated local average, and move to this point, enabling yet another view on its vicinity. It is intuitively clear that by iterating this procedure, the element will move into increasingly dense regions of the population, and eventually be drawn to the center of one of the clusters (which is probably, but not necessarily, the nearest center to its original position).

The fable given above can be seen as a qualitative description of the *mean shift procedure*, which computes an iterated sequence of local means which can be shown to converge to a local mode of its distribution; that is to a local maximum of its density. The mean shift can be traced back to a 1975 paper by Fukanaga and Hostetler [33] (and its conceptual origins even beyond that), gaining significant traction in the computer science literature in the late nineties and early 2000's [18, 22, 23]. It has struggled so far, in our view rather surprisingly, to gain significant foothold in Statistics. Its most immediate fields of application are, clearly, mode detection and clustering, on which we further elaborate in Section 2, though it has also been used for other statistical problems such as regression [7, 16] or the estimation of principal curves and manifolds [30].

As motivated above, the mean shift is closely related to the problem of density estimation. While it has been applied in the context of several density estimators including mixture models [6, 11], we focus in this work on the kernel density estimator (KDE), which we therefore introduce in Subsection 2.2. Section 2 continues with a brief introduction to modal clustering as well as an illustrative data example to which we will return lateron — the Hildago stamp data. Section 3 derives, starting from the the KDE, the mean shift procedure for modal estimation, and discusses some of its properties including the relation to the Expectation-Maximization algorithm. Section 4 focuses on univariate data, discussing modal testing, the estimation of modes and antimodes, as well as aspects of bandwidth selection for modal clustering. Section 5 contributes a simulation study and three real data examples (one of which revisiting the Hidalgo stamp data and the other ones being multivariate), before the paper finishes in Section 6 with a Discussion. Some theoretical considerations, as well as explicit code to reproduce the application studies, in form of R Markdown documents, are provided in supplementary material.

2 Preliminaries

Consider a random vector $X = (X_1, \dots, X_d)^T$, with probability density function $f : \mathbb{R}^d \longrightarrow \mathbb{R}$, and overall mean $\mu = (\mu_1, \dots, \mu_d)^T$. We are given n realizations of X, denoted by $x_i = (x_{i1}, \dots, x_{id})^T$, $i = 1, \dots, n$.

2.1 Introduction to modal clustering

To identify the clusters and determine which observations belong to each cluster, several methods have been proposed in the literature. A comprehensive review of different clustering methods is provided in [39]. Among the various approaches to clustering, there is a wide category of methods based on statistical modelling, where clusters are associated with features of the underlying density function, f, of the data. These methods can be grouped into parametric methods, where cluster centres are linked to distributional parameters, or nonparametric methods, which rely on the modes of X to identify the cluster centers. This latter group of methods is commonly referred to as *modal clustering*.

When performing modal clustering there are three key points. Firstly, some methodology is needed to estimate the modes of X from a given data set; i.e. the (local) maxima of the underlying density f of X. While a wide range of approaches, including model-based ones (see e.g. [9]), can be pursued for this purpose, our interest is on methods which estimate the modes based on kernel density estimation (KDE), which we briefly introduce in Section 2.2. A second point is to decide on the degree of smoothing, in the KDE context expressed by the value of a bandwidth matrix, possibly in conjunction with a method to discard modes that are not significant. This point is highly related to assessing the number of clusters or, equivalently, the number of modes (see Section 4.1). Then, in order to classify observations, the third point would be a principled mechanism to associate observations with clusters (modes).

A traditional approach to this problem are the λ-level sets (see, e.g., [54]), where clusters are defined using the disjoint connected sets of values of x for which the estimated probability density function, $\hat{f}(x)$, is larger or equal than a given value λ. Identifying the connected sets in the multivariate case is not straightforward and solutions depend on the value of λ. Techniques to examine level sets over ranges of levels have been developed; see e.g. [62]'s cluster tree. In the univariate case, such approaches naturally lead to a population-based view where clusters are defined as connected subsets of the real line separated by local minima [10].

A more general approach, which is grounded in Morse theory and applicable in the multivariate case, would use the notion of the domain of attraction (see [10]), where the clusters' regions are the unstable manifolds of the negative gradient flow corresponding to each mode, in order to identify a partitioning of the data space according to the identified cluster centres. The main issue, in that case, is the need for some strong smoothness assumptions over f.

As pointed out in [3], and similarly observed by [11], the two clustering approaches in the previous paragraphs are fundamentally equivalent. In the univariate setting, when the true density is assumed to be bounded and continuous, the problem can be phrased as that of identifying the antimodes as boundaries of the clusters. This leads to the question of how to estimate the antimodes. An obvious approach is to identify the relative minima of the KDE, $\hat{f}$, which we will pursue in Section 4.3.

A more direct strategy is to connect each observation with its "corresponding" mode. In that case, the most natural approach, which will be discussed in Section 3, is to employ a gradient ascent algorithm, the mean shift. This technique will work well in both the univariate and the multivariate setting. Commonly, this approach is only formulated for the observed sample. However, as a referee pointed out, the mean shift paths approximate, by means of the sample, the gradient lines, and these gradient lines do exist, and can be estimated, even for unobserved data points.

2.2 Kernel density estimation

Kernel density estimation works by redistributing the point mass of each observation around its vicinity and then summing over the contributions from all data points. Specifically, Let $H = \text{diag}(h_1^2, \ldots, h_d^2)$ denote a matrix with squares of positive bandwidths h_j on the diagonal, implying positive definiteness of H. For a point $x \in \mathbb{R}^d$, let $||x||$ denote its Euclidean norm; i.e. $||x||^2 = x_1^2 + \ldots + x_d^2$, and $k : \mathbb{R} \longrightarrow \mathbb{R}$ a non-negative function such that $\int_{\mathbb{R}^d} k(||x||^2)dx = 1$. The function $K(x) = k(||x||^2)$ is often referred to as a *kernel* function, and k as its *profile*. For instance, for the multivariate Gaussian kernel, one has

$$k(u) = (2\pi)^{-d/2} e^{-u/2} \tag{1}$$

and so $K(x) = (2\pi)^{-d/2} e^{-||x||^2/2}$.

A *kernel density estimator* (KDE) $\hat{f}$ of f is then given by

$$\hat{f}(x) = \frac{1}{n|H|^{1/2}} \sum_{i=1}^{n} k\left(||H^{-1/2}(x_i - x)||^2\right). \tag{2}$$

If the main interest is the estimation of modes, one could do this, in principle, directly based on (2). Simplistically, one can define some grid spanning the domain of X, and then identify the modes based on a grid search (which is not entirely trivial if $d > 1$). However, this methodology suffers from the obvious drawback that it depends on the specification of the grid, and so modes may be missed if they fall 'between the cracks'. Also, this approach can be computationally costly if d is large. Furthermore, the possibility to obtain an overall density estimate may not always be available; for instance for our 'element' in Section 1, it is not. The approaches discussed in this work aim at identifying the maxima of (2) (and so, estimating the modes of X) without actually requiring the estimation of the density itself.

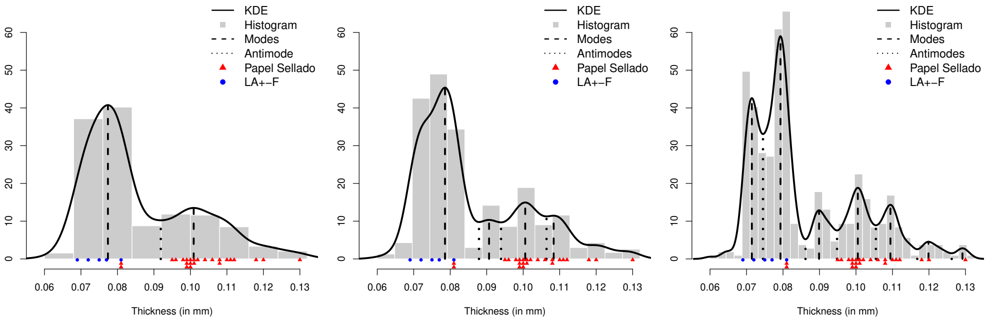

Fig. 1 Histogram and KDE for the 485 stamps data set from the 1872 Hidalgo Issue of Mexico, for different binwidths and bandwidths according to [64] (left), [1] (center) and [42] (right). From the KDE, for each bandwidth value, the estimated modes (dashed) and antimodes (dotted) are identified by vertical lines. The points in the x-axis represent the thickness of stamps watermarked with *LA+-F* (circles) and *Papel sellado* (triangles).

2.3 Hidalgo stamp example

In order to illustrate concepts, we introduce the data known as the *1872 Hidalgo issue of Mexico* (see [42]). This dataset contains the thickness of 485 stamps that were printed in a mixture of paper types in Mexico between 1872 and 1874. The objective here is to determine to which sub-collection each stamp belongs according to its thickness. The problem is that there is not a standard rule in catalogues for classifying this particular stamps issue. The difficulty arises from two sides, first, it is known that the handmade paper presents a lot of variability in the thickness of the paper. Second, since of scarcity of ordinary white wove paper, other types of paper were used (some of them watermarked), such us *Papel Sellado* or *La Croix–Freres* of France (*LA+-F*).

The previous situation creates three questions, related to modal clustering, that we will try to solve in Section 5.2. 1) How many subgroups (clusters/-modes) are there? 2) Where are the modes and antimodes located? 3) What cluster does each stamp belong to? A preliminary hint for solving these questions is presented in Figure 1. There, we provide the histogram and the KDE for different bin- and band-widths for the 1872 Hidalgo issue of Mexico, where we also indicate the 29 stamps where a watermark was present. The three graphs are associated with three historical nonparametric solutions of the previous questions: [64], [1] and [42]. The solutions for some of the previous questions from a parametric perspective were reviewed in [49, Ch. 6].

3 The mean shift

Rather than building ad-hoc solutions for modal estimation as outlined towards the end of Section 2.2, it appears more natural to go back to first principles and identify density modes through their mathematical properties, being points with vanishing density gradients. This section will demonstrate that this line of reasoning leads us directly to the definition of the mean shift.

3.1 Derivation

A natural estimator for the maxima of f is given by the maxima of $\hat{f}$, which we can find by equating its gradient

$$\nabla \hat{f}(x) = -\frac{2}{n|H|^{1/2}} H^{-1} \sum_{i=1}^{n} k'\left(||H^{-1/2}(x_i - x)||^2\right)(x_i - x) \tag{3}$$

to zero, yielding the equation

$$x = \frac{\sum_{i=1}^{n} k'\left(||H^{-1/2}(x_i - x)||^2\right) x_i}{\sum_{i=1}^{n} k'\left(||H^{-1/2}(x_i - x)||^2\right)} \equiv \mu(x). \tag{4}$$

We see that in (4) x is equated to its local mean, $\mu(x)$, that is, the (weighted) mean of all points in a local neighborhood of x. This equation, which can be considered as a localized form of the self-consistency property [63], does not have an analytical solution, but given some starting point $x^{(0)}$ one can iteratively compute the sequence

$$x^{(\ell)} = \mu(x^{(\ell-1)}), \quad \ell = 1, 2, \ldots \tag{5}$$

which converges to its fixed point $x = \mu(x)$ which is just the solution of (4). This is, in a nutshell, the idea of the *mean shift procedure* [18, 22]. The fixed point solution $x^{\blacktriangle} = \mu(x^{\blacktriangle})$ of (5) corresponds to a local maximum of $\hat{f}$, with $x^{\blacktriangle}$ serving as our estimate[1] of the (local) mode of X. It is worth highlighting that, while conceptually being based on the KDE, the mean shift procedure does not require the explicit computation of the estimated density $\hat{f}$. Also, it is worth noting that if one estimates the density with some parametric model, the idea of using a fixed-point algorithm connects the mean shift with fixed-point cluster analysis (see [38]).

3.2 The mean shift vector

For the Gaussian profile (1), one has $k'(u) = -\frac{1}{2}k(u)$ and so in (4) one can replace k' by k. It is then also clear by dividing (3) by (2) that $\nabla \hat{f}(x)/\hat{f}(x) = H^{-1}(\mu(x) - x)$, or equivalently,

$$s(x) \equiv \mu(x) - x = H \frac{\nabla \hat{f}(x)}{\hat{f}(x)}. \tag{6}$$

[1] Conceptually, one could point out here that the maxima of $\hat{f}$ constitute the *estimates* of the (local) modes, while the iterative solutions entailed by equation (5) act as *approximations* to these estimates. For all practical purposes, $x^{\blacktriangle}$ will still serve as 'the estimate' of a local mode of X; therefore we do not make further efforts to distinguish this conceptual nuance further in this exposition.

The expression $s(x)$ is the *mean shift vector*. For isotropic bandwidth matrices $H = h^2\mathbf{I}_d$, $s(x)$ points exactly into the direction of the gradient of $\hat{f}$, and otherwise, into a bandwidth-rotated version of the latter. In either case, it takes larger steps for larger gradients and smaller densities. Thus, one can reformulate the update step of the mean shift procedure as

$$x^{(\ell)} = x^{(\ell-1)} + s(x^{(\ell-1)}) \tag{7}$$

where the additions $s(x^{(\ell-1)})$ guarantee a hill-climbing process (a proof of this property covering the non-isotropic bandwidth case was provided in [14]). In particular, when initialized far from the mode, the first steps are usually large with quick progress towards the mode [7]. When the latter is reached, at the fixed point $x^{\blacktriangle}$, one has $s(x^{\blacktriangle}) = 0$.

We will restrict to the use of the Gaussian profile (1) for the remainder of this exposition.

3.3 Blurring and non-blurring mean shift

At this point it is worth looking back at the 'element' in our fable. One can distinguish two scenarios: Firstly, if one takes the view that the element itself is part of the data, then by moving along its trajectory into denser regions it is *changing* the data. Assuming all population elements do one step forward simultaneously (jointly forming a 'wave'), one can think of this as a process in which the data condense themselves with each iteration towards the local centers. This concept is known (perhaps rather misleadingly) as *blurring* mean shift in the literature [18]. On the contrary, we can imagine a situation in which the element takes a superior position where, while sailing along its trajectory, it remembers its original position, and sticks with that position as far as all computations are concerned. That is, in such a *non-blurring* mean shift, cluster centers are found without modifying the data. The blurring mean shift is a technique which is mainly relevant from a computer science perspective (see, e.g., [7, 65]). We do not discuss blurring mean shift in this paper; i.e. in all our considerations the data themselves are not moving.

3.4 Mean shift and EM

It is insightful to very slightly reformulate the expression for the local mean $\mu(x)$ defined in (4). In the case of a Gaussian profile, this reads as

$$\mu(x) = \sum_{i=1}^{n} \frac{k\left(||H^{-1/2}(x_i - x)||^2\right)}{\sum_{j=1}^{n} k\left(||H^{-1/2}(x_j - x)||^2\right)} x_i \equiv \sum_{i=1}^{n} w_i x_i, \tag{8}$$

with $w_i = k\left(||H^{-1/2}(x_i - x)||^2\right) / \sum_{j=1}^{n} k\left(||H^{-1/2}(x_j - x)||^2\right)$. Now, assuming for a moment that x represents the realization of a random variable which is a mixture of n multivariate Gaussians centered at the original observations x_i, each with variance matrix H, and that all of these possess a prior probability of $1/n$, it is clear from Bayes' theorem that w_i is just the posterior probability that observation x is generated from the component centered at x_i, and so the computation of w_i mirrors the E-step well known from the Expectation-Maximization (EM) algorithm for finite Gaussian mixtures. A rigorous formulation of the equivalence of procedure (5) (in the case of a Gaussian profile) to an EM algorithm requires a more elaborate argument [6]; we recall this argument, adapted to our notational framework, in the supplementary material (S2). This result is of importance in two ways: Firstly, inferential methodology and theoretical results known for the EM algorithm may be applicable to the mean shift. Secondly, the link between mean shift and EM allows inferring convergence properties of the latter. A discussion of convergence properties of the mean shift algorithm from that perspective is delegated to the supplementary material (S3).

Other EM-style algorithms, albeit not equivalent to the mean shift procedure, have been proposed in the literature for modal estimation. [47] develop the Modal EM (MEM) algorithm which finds hill-climbing paths from each data point to the local density maxima, as well as the ridgelines between clusters, relying on explicit mixture density estimates for their identification. [46] proposed an extension of the EM algorithm, called MM (Minorize/Maximize) algorithm, where a a convex objective function is iteratively minorized by easily-maximizable surrogate functions. [66] applied first the MEM algorithm using Gaussian kernels, and then the MM algorithm allowing for more efficient maximization when using non-Gaussian kernels, for modal linear regression.

4 Mode detection for univariate data

Let us now consider univariate data; i.e. $d = 1$ and hence $H = h^2$ with a positive bandwidth value h. We also still work with the conceptually simpler case of a Gaussian kernel, i.e. we can use k instead of k' in (4) and so the mean shift can be expressed as

$$s(x) = \frac{\sum_{i=1}^{n} k((x_i - x)/h)x_i}{\sum_{i=1}^{n} k((x_i - x)/h)} - x = \frac{\sum_{i=1}^{n} k((x_i - x)/h)(x_i - x)}{\sum_{i=1}^{n} k((x_i - x)/h)}. \tag{9}$$

4.1 Modal testing

In the clustering analysis, there is always a key point which is to decide on the number of clusters. In the modal clustering case, this decision is equivalent to assessing the number of modes. In the multivariate literature, this problem has been initially solved by using automatic bandwidth rules (see, e.g., [27]), from which the modes are estimated. The main problem of automatic rules is that, in general, they do not search for the correct number of modes and spurious modes may appear. Other approaches try to identify the "true" modes, for example by testing if a mode is significant or not [34]. But, these methods still rely on the initial choice of candidates for modes.

At least in the univariate case, an alternative to those methods is testing the number of modes to decide on the number of clusters. The problem of nonparametrically assessing the number of modes has been studied during the last decades, mainly using two approaches, the critical bandwidth [61] and the excess mass [53]. A review of different techniques for testing the number of modes in the univariate setting can be found in [1]. As mentioned before, a well-calibrated test of multimodality, such as the test proposed by [1], could be used before performing the modal clustering to determine the the number of underlying clusters.

Up to the authors' knowledge, for the multivariate case, there is a lack of formal tests for testing the number of modes. Thus, it is not straightforward to extend this idea to the scenario where $d > 1$. An interesting alternative would be to explore the number of modes similarly as done for the SiZer in the univariate case [15]. This can be done by testing for significant modes for different values of the matrix bandwidth H (see [24]). Then, the number of modes (clusters) could be chosen by the practitioner by summarizing the results obtained for a grid of matrices H.

4.2 Mode estimation

We have already established that the iterative execution of (7) defines a trajectory of points converging to the mode. To detect all modes of a random vector X given n realizations $x_1, \dots, x_n$ of X, we execute the following algorithm:

Algorithm 1

```
   Denote $M$ an empty list of modes.
1. Decide on
   (a) a suitable bandwidth $h$ (see Section 4.4);
   (b) a criterion $\mathcal{C}_1$ for the convergence of the mean shift procedure (see
       Remark 1);
   (c) a criterion $\mathcal{C}_2$ for merging modes (see Remark 2).
2. For $i = 1, \dots, n$,
   (i) Set $x^{(0)} = x_i$.
   (ii) Iterate equation (7), using equation (9) for $s(\cdot)$, until $\mathcal{C}_1$ is fulfilled,
        yielding the mode estimate $x^{\blacktriangle}$.
   (iii) Attempt merging this mode with all other modes already present in $M$.
         If the mode can be merged, do nothing. Otherwise, add $x^{\blacktriangle}$ to $M$.
```

Remark 1. Suitable choices of criterion $\mathcal{C}_1$ will indicate convergence at iteration ℓ if $||s(x^{(\ell)})||$, or a normalized version of it, falls below a certain threshold. In the implementation of function `ms` in R package **LPCM** [29], this criterion takes the form

$$\frac{||s(x^{(\ell)})||}{||x^{\ell} - \mu||} < t_1 \ll 1$$

where the positive constant t_1 is by default set to 0.0001; i.e. convergence is taken to be achieved when the mean shift from a point incurs a move of less than $\frac{1}{10000}$ of its Euclidean distance to the overall mean.

Remark 2. The criterion $\mathcal{C}_2$ for merging modes will check whether, once a new mode m is found, its distance to any existing mode m' is sufficiently small so that the two modes are deemed identical. In the implementation in function `ms` in R package **LPCM** [29], when

$$\frac{||m' - m||}{||m' - \mu||} < t_2 < 1$$

for any $m' \in M$ then the two modes m and m' are merged. The default setting for the positive constant t_2 is $t_2 = \sqrt{t_1}$; i.e. less approximation is required for mode-merging than for establishing convergence.

4.3 Antimode estimation

Antimodes of X are defined as the minima of f; i.e. points where the likelihood of observing a realization of X increases in any direction when moving away from it. This is equivalent to saying that the density f of X attains a local minimum at the antimode, and so a natural strategy is to identify these as the local minima of $\hat{f}$.

We have already presented an efficient way of identifying the maxima of f through a gradient ascent algorithm, the mean shift. Can an 'inverse' version of this algorithm be used to identify density minima, and hence antimodes? To see this, recall firstly that $s(x) = \mu(x) - x = h^2 \nabla \hat{f}(x)/\hat{f}(x)$ which clearly implies

$$-s(x) = -h^2 \nabla \hat{f}(x)/\hat{f}(x) = -h^2 \nabla \log(\hat{f}(x))$$

so that a step from x into the direction $-s(x)$ will lead to gradient descent (Proposition 3 in [67] ensures that such a step downhill is well-defined under a Lipschitz condition on the gradient of $\hat{f}$). This suggests that the iterative, inverse mean shift procedure

$$x^{(\ell)} = x^{(\ell-1)} - s(x^{(\ell-1)}) = 2x^{(\ell-1)} - \mu(x^{(l-1)}) \tag{10}$$

will descend towards an antimode $x^{\blacktriangledown}$, which is attained when $s(x^{\blacktriangledown}) = 0 = \nabla \hat{f}(x^{\blacktriangledown})$. Summarizing, this gives rise to the following algorithm.

Algorithm 2

```
Denote $A$ an empty list of antimodes.

1. Decide on
  (a) a suitable bandwidth $h$ (see Section 4.4);
  (b) a criterion $\mathcal{C}_1$ for stopping the inverse mean shift procedure;
  (c) a criterion $\mathcal{C}_2$ for merging antimodes.
2. For $i = 1, \dots, n$,
  (i) Set $x^{(0)} = x_i$.
```

(ii) Iterate equation (10), using equation (9) for $s(\cdot)$, until $\mathcal{C}_1$ is fulfilled, yielding an estimate of the antimode, $x^{\blacktriangledown}$.

(iii) Attempt merging this antimode with all other antimodes already present in A. If the antimode can be merged, do nothing. Otherwise, add $x^{\blacktriangledown}$ to A.

Remark 3. For the choice of $\mathcal{C}_1$ and $\mathcal{C}_2$, we use the same settings as for modal estimation in Algorithm 1, but with constant $t_1 = 10^{-6}$.

Remark 4. Convergence of Algorithm 2 is not mathematically guaranteed, and may be violated if the antimode is associated with a very small value $\hat{f}(x^{\blacktriangledown})$. For, instance if one considers the estimation of the antimode between two standard Gaussian distributions, then such convergence problems (typically expressed as oscillating behavior of $x^{(\ell)}$ around the true antimode) begin to arise when the means of these distributions are more than six standard deviations apart. While such problems can be addressed algorithmically in principle, the application studies provided in this manuscript do not require, and hence do not make, any such adjustments. It is furthermore clear that, if the starting point of the procedure lies below the smallest or beyond the largest mode, then Algorithm 2 cannot converge, but will instead diverge to infinity. We will see such behavior in the application studies. The occurrence of such an event is still meaningful, as it indicates that there is no antimode located between that starting point and the respective boundary. In practice, the recommended way of assessing this question is to look at the whole trajectory of points $x^{(\ell)}$ (does it converge, diverge, or oscillate). The function `ms.rep.min` in R package **LPCM** [29] provides this full trajectory as an output component.

4.4 Clustering through modes and antimodes

Two approaches for clustering derive naturally from the methodologies given in Sections 4.2 and 4.3. Firstly, fully in the spirit of the mean shift procedure, one can allocate all observations to the cluster center to which their trajectory has converged. In the univariate setting, a second, and perhaps more attractive route, is to employ the antimodes as boundaries of the clusters' regions, effectively attaining a *partition* [10, 12] of the domain of X into a non-overlapping sequence of intervals. While the two approaches are equivalent as far as the cluster assignment of the actual data points are concerned, the second approach is more general as it also implies a clustering of non-observed samples, giving rise to a population- rather than sample-based view of the problem [10].

A crucial aspect in practice is then the estimation of the true antimodes from the data. In Section 4.3, we provided an efficient algorithm for doing that estimation, but results still depend on the value of the bandwidth h. After determining the number of clusters (see Section 4.1), from a theoretical perspective, an appealing choice would be to select as h the critical bandwidth for that specific number of modes. Assume that the modes and antimodes lie in a known closed interval I, then the critical bandwidth of [36] for k modes[2] is defined as,

$$h_k = \inf\{h : \hat{f}_h \text{ has at most } k \text{ modes in } I\}. \tag{11}$$

The critical bandwidth of [61] is defined analogously replacing I by the complete real line. When knowing the closed interval I, the theoretical justification for employing the critical bandwidth of [36] to estimate the antimodes can be found in [1]. In that paper, under some smoothness assumptions over f, it is proven that the modes and antimodes of $\hat{f}_{h_k}$ almost surely converge to the true modes and antimodes of f, as $n \to \infty$.

In practice, the main issue of employing the critical bandwidth as h is that it always provides an almost-flat density region (see, e.g., the central panel in Figure 1). Since, in that region, the estimated density derivative is almost zero, the mean shift algorithm needs many iterations to find the modes. Thus, for computational reasons, is better to choose another bandwidth still guaranteeing that it has the desired number of modes. Although one can think in several choices satisfying that requirement, when $k > 1$, one alternative would be to employ the geometric mean of h_k and h_{k-1}, i.e., to use the bandwidth $h'_k = (h_k h_{k-1})^{1/2}$. By using that bandwidth, which will be larger than h_k (but closer to h_k than to h_{k-1}), it is still guaranteed that the kernel density estimation has k modes and the mean shift algorithm will converge faster to the estimated modes and antimodes.

Collecting the results of this section our proposed modal clustering procedure for the univariate case can be summarized in the following algorithm.

Algorithm 3

```
1. If the number of clusters is unknown, determine the number of modes for
   a significance level $\alpha$ by using the test of [1].
2. Denote with $k$ the number of modes.
   (a) If $k = 1$, all the data points belong to the same cluster.
   (b) If $k > 1$, compute $h'_k$, the geometric mean of the critical bandwidths $h_k$
       and $h_{k-1}$.
3. If $k > 1$, estimate the location of the antimodes of $\hat{f}_{h'_k}$ (see Section 4.3).
   Denote with $x_j^{\blacktriangledown}$, $j \in \{1, \ldots, k-1\}$ the location of the estimated antimodes.
4. If $k > 1$, the data points can be associated with the clusters, which regions
   are $(-\infty, x_1^{\blacktriangledown})$, $[x_1^{\blacktriangledown}, x_2^{\blacktriangledown})$, ..., $[x_{k-1}^{\blacktriangledown}, \infty)$.
```

[2] We have so far consistently spoken of the *modes* as a property of a random variable (or vector) X, corresponding to the (local) *maxima* of the associated density f of X. In this subsection, for ease of presentation and consistency with the source literature, we allow us to be a bit more lenient and speak of modes and antimodes of a density f (rather than X), with the obvious meaning.

5 Application studies

5.1 Simulation study

We first start with a brief simulation study to show the potential of the proposed Algorithm 3 for clustering (see Section 4.4). We fixed the scenario where the number of clusters is known as the correct behaviour of the test of [1] for analysing the number of modes was already studied in their paper. Thus, we performed steps 2-4 of Algorithm 3 by imposing the correct number of modes (k) in 6 mixture models. Those models consist in: simple mixtures, where the components are normal, $\mathrm{N}(\mu, \sigma^2)$, and their number coincides with the number of modes (M1 and M4), Gaussian mixtures where the number of components is greater than k (M2), and more complex mixtures where some components are Beta(θ, ϕ) or Gamma(α, β) (M3, M5, and M6; the same notation as in [43] is employed). These models, based on those employed in [1], are:

Bimodal models ($k = 2$)*:*

- M1: $0.5 \cdot N(0.3, 0.0197) + 0.5 \cdot N(0.7, 0.0197)$.
- M2: $0.6 \cdot N(0.384, 0.01202) + 0.2 \cdot N(0.2, 0.05) + 0.2 \cdot N(0.9, 0.00272)$.
- M3: $0.3 \cdot N(0.13, 0.1) + 0.3 \cdot N(0.81, 0.1) + 0.2 \cdot \mathrm{Gamma}(3, 9) + 0.2 \cdot \mathrm{Beta}(7, 2)$.

Trimodal models ($k = 3$)*:*

- M4: $0.45 \cdot N(0.25, 0.015) + 0.45 \cdot N(0.6, 0.015) + 0.1 \cdot N(0.95222, 0.00049)$.
- M5: $0.3 \cdot N(0.2, 0.005) + 0.3 \cdot N(0.5, 0.005) + 0.4 \cdot \mathrm{Beta}(5, 2)$.
- M6: $0.3 \cdot N(0.2, 0.01) + 0.3 \cdot N(0.5, 0.01) + 0.2 \cdot \mathrm{Gamma}(3, 9) + 0.2 \cdot \mathrm{Beta}(7, 2)$.

The "true" classification of the randomly generated data is assigned according to the data point values following the ideas of [10] in its "population goal of modal clustering". The probability density functions and the cluster borders of Models M1–M6 are represented in Figure 2. Note that, for each model, the cluster borders are always the same for all simulation runs. The "true" classification of the data is done according to the interval (cluster) to which it belongs. For each model, this is shown for one random sample of size $n = 100$ on the rug plots in Figure 2.

The objective is not only to show the correct performance of the proposed algorithm but also to compare it with other density-based clustering methods. More specifically, we will compare with:

- Clustering based on Gaussian mixture models. We have employed the `mclust` [59] R package to fit the mixture model. The variance of each component may be different. We imposed two scenarios: (i) the number of components is equal to the known number of modes ("MC (k)"), and (ii) the number of components is chosen between 1 and 9 according to BIC, default option in `mclust` ("MC"). The classification is done following the mixture model modal clustering proposed by [11].
- Nonparametric clustering based on level-set methods. We have employed the `pdfCluster` [4] R package. The output of the clustering depends on the

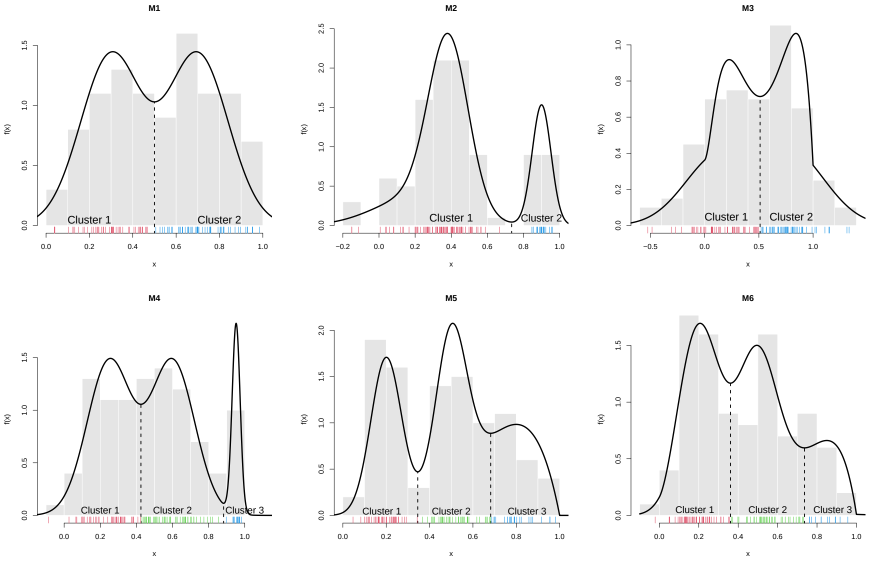

Fig. 2 Density functions of models M1–M6, dashed lines divide the different true clusters. A rug plot (in colours along the x-axis) and the histogram (in grey) of 100 random data generated from each model are also displayed. Each colour in the rug plot identifies one true cluster.

bandwidth. We imposed two scenarios: (i) bandwidth by default (“PC”), and (ii) h'_k (“PC (hg)”), the geometric mean of the critical bandwidths.

- Density based clustering DBSCAN (“DB”) introduced by Ester *et al.* [31]. We have employed the `fpc` [37] R package. Here the number of neighbours (`MinPts`) within the radius ϵ (`eps`) need to be imposed. In [57] some ideas are given of how to set these parameters. First, it is suggested to consider `MinPts`$= 2 \cdot$dim. In our experimental results this choice created more clusters than expected. For that reason, we considered the minimum value of `MinPts`, between 2 and 10, from which the number of clusters is lower or equal to the known number of modes (k); or `MinPts`=10, if the number of clusters is always greater than k. Regarding `eps`, for one sample, it is suggested to look at the l-distance plot and searching for a knee (see [31] or [57]). Since we will have several samples, following [48], we will automatize the process of searching for a knee in the `MinPts`-distance plot. This is done using the “Kneedle” algorithm of [56].

We have generated 1000 samples of sizes $n = 50$, $n = 100$, and $n = 200$ from each model. For each sample, we have computed the dissimilarity between the true cluster and the one obtained for each clustering method with the Misclassification Error Distance (MED), see [50]. This index is computed from the proportion of unmatched clusters. More precisely, MED$=$ $1 - 1/n \max_{\pi} \sum_{i=1}^{k} n_{k,\pi(k)}$, where $n_{k,j}$ are the elements of the confusion matrix

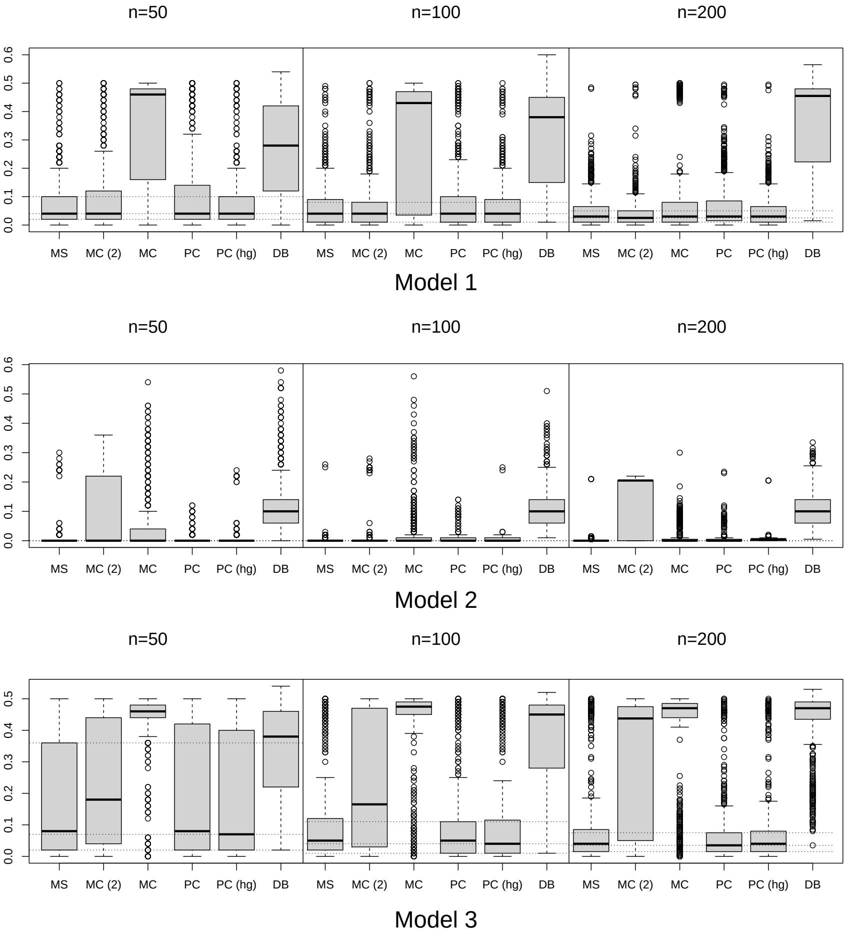

Fig. 3 Boxplots associated with the MED of 1000 samples of size n, for the bimodal models M1–M3. For each sample size, the horizontal dotted lines indicate the value of the lowest quartile values among the analysed clustering methods.

obtained from the k and k' classes of the two clustering methods, π is a mapping of $\{1, \ldots, k\}$ into $\{1, \ldots, k'\}$, and $k \leq k'$. The MED is between 0 and 1, and values close to zero indicate that the output of the clustering method is close to the true cluster partition. For each model and each sample size, the boxplots associated with the MED of the 1000 samples are represented in Figures 3 and 4.

By looking at the results of applying Algorithm 3 ("MS") to data generated by the 6 models, we can see that this algorithm performs well in the sense that: (i) the median value of MED decreases with the sample size, being close to zero for $n = 200$ (the median is between 0 and 0.065), the dispersion (interquartile range) decreases with the sample size. When comparing with

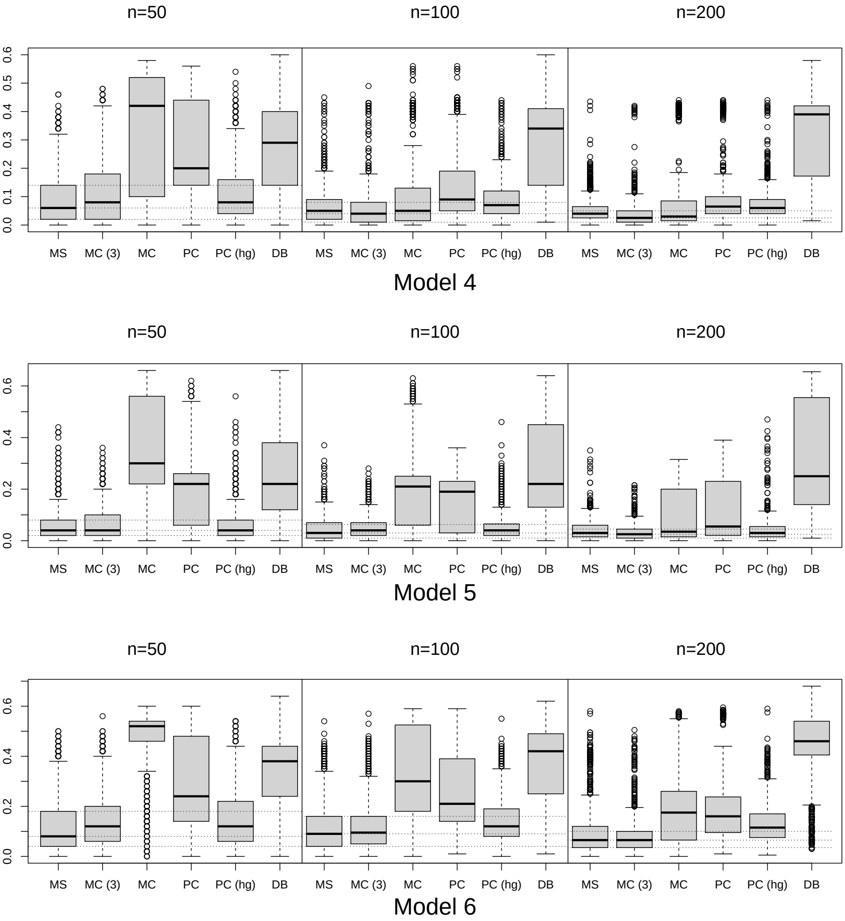

Fig. 4 Boxplots associated with the MED of 1000 samples of size n, for the trimodal models M4–M6. For each sample size, the horizontal dotted lines indicate the value of the lowest quartile values among the analysed clustering methods.

other methods available in the literature, we can see that the proposed procedure is (almost) always the one providing the lowest (quartile) MED values. The only exceptions are the models well approximated by a mixture of k normal densities (Model 1 and 4), where the clustering based on the Gaussian mixture model performs slightly better. This was expected as we are imposing the true model in the parametric fitting. But, as we can see in Model 3, the clustering based on Gaussian mixture model may be completely erratic if the data is generated from a different distribution. `pdfCluster` also provides a competitive clustering method, although, it is generally outperformed by "MS". The similar obtained results when using the same bandwidth, of "PC (hg)" and "MS", were as expected, given the earlier stated insight (Section

k	1	2	3	4	5	6	7
h_k	0.00673	0.00323	0.00301	0.00283	0.00263	0.00242	0.00149
h'_k		0.00467	0.00312	0.00292	0.00273	0.00252	0.00190

Table 1 Critical bandwidth (h_k) and geometric mean of h_k and h_{k-1} (h'_k) for the Hidalgo stamp data. From left to right, the number of modes k is between 1 and 7.

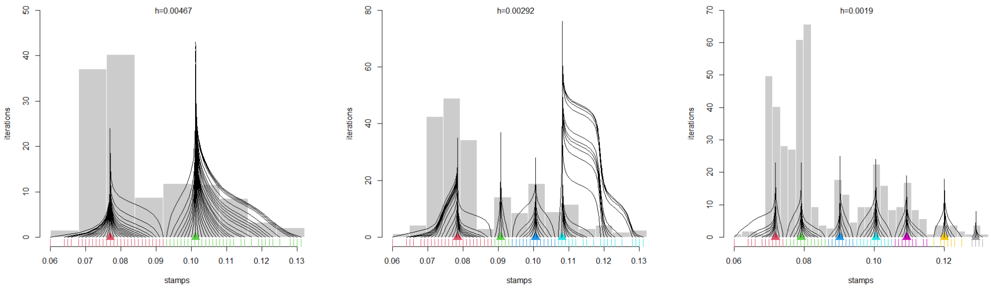

Fig. 5 Histograms for the 485 stamps data set from the 1872 Hidalgo Issue of Mexico. The vertical dimension is meaningless for the histograms in the displayed graphs, but of course the histograms integrate to 1. The black rising lines shows the mean shift trajectories (horizontal) as a function of iteration number (vertical). The detected modes are highlighted through triangles on the horizontal axis, and the rugs below the axis indicate the induced clustering via mean shift.

2.1) that the clustering approaches based on density level-sets and domains of attraction are essentially equivalent.

In Figures 3 and 4, we show that our proposal of using the geometric mean of the critical bandwidths can also help to improve the clustering results achieved by `pdfCluster`. Thus, if `pdfCluster` is employed, when the number of clusters (modes) is known, we also recommend employing our Step 2 in Algorithm 3 to select the bandwidth (“PC (hg)”) instead of the normal rule bandwidth employed by default (“PC”).

5.2 Hidalgo stamp data

We apply the techniques introduced in Section 4 on the Hidalgo stamp data introduced in Subsection 2.3. There, in Figure 1, we represented the KDE with the critical bandwidths for 2, 4, and 7 modes (h_2, h_4, and h_7, see Table 1). In Figure 5, we illustrate the application of Algorithm 1 onto this data set. Each plot shows how the mean shift trajectories proceed as a function of the iteration number (with the latter being displayed on the vertical axis). The three panels in Figure 5 correspond to the scenarios with the same number of modes as those illustrated in Figure 1. For computational reasons (see Section 4.4), in this section, instead of using the critical bandwidths, the bandwidths h'_2, h'_4, and h'_7 (see Table 1) were employed. We see that convergence in these scenarios is attained within 30 to 80 iterations. The precise values of the modes after convergence are provided in Table 2.

h	Modes (*Antimodes*)						
0.00467	0.0770	0.1012					
	(0.0926)						
0.00292	0.0785	0.0907	0.1006	0.1081			
	(0.0882)	*(0.0938)*	*(0.1067)*				
0.00190	0.0718	0.0792	0.0902	0.1005	0.1093	0.1201	0.1291
	(0.0745)	*(0.0864)*	*(0.0946)*	*(0.1055)*	*(0.1168)*	*(0.1261)*	

Table 2 Modes (*Antimodes*) for the Hidalgo stamp data under different bandwidths. The bandwidths given in the first column correspond, from top to bottom, to h'_2, h'_4 and h'_7.

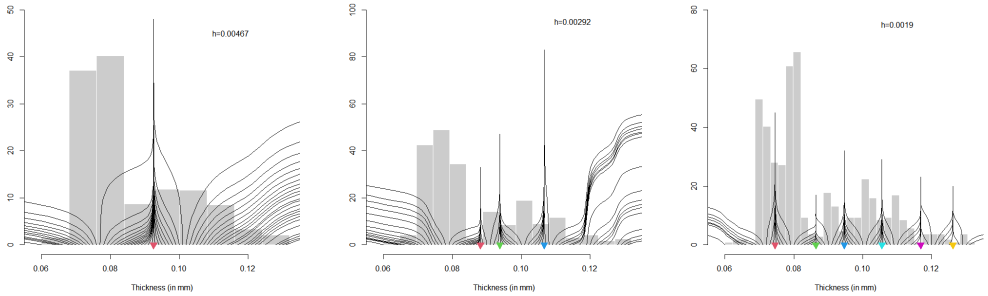

Fig. 6 Histograms for the 485 stamps data set from the 1872 Hidalgo Issue of Mexico. The vertical dimension is meaningless for the histograms in the displayed graphs, but of course the histograms integrate to 1. The black rising lines shows the inverse mean shift trajectories (horizontal) as a function of iteration number (vertical), for the computation of antimodes. The positions of the antimodes are indicated by downfacing triangles.

For the estimation of antimodes via Algorithm 2, results are summarized in Figures 6 and Table 2 (italic). We see that there are two types of trajectories: Trajectories from starting points situated between two modes always converge to an antimode between those modes. Trajectories which start to the left of the left-most mode, or to the right of the right-most mode, will not converge (rightly so; because there are no antimodes to detect here). In summary, we find that 1, 3 and 6 antimodes have been identified in the respective scenarios, consistent with the number of modes.

For determining the number of modes, we have applied the well-calibrated multimodality test of [1], available in the R package `multimode` [2]. For a significance level of $\alpha = 0.05$, the null hypothesis of having k modes is rejected until $k = 3$, and it is not for $k \geq 4$ (see [2]). Thus, it can be assumed that the number of clusters/modes is equal to 4. Following the Algorithm 3, for our clustering method, we will employ as h the geometric mean of h_4 and h_3, i.e., h'_4 (see Table 1). In Table 2, we provide the antimodes estimation for h'_4. As mentioned in Section 4.4, the clusterings of observations obtained from the partitioning of the real axis through the antimodes, and the one induced by the mean shift trajectories, are equivalent. The clustering result is displayed, for the bandwidth h'_4, in form of a colored rug plot in the central panel of Fig.

5. According to this outcome and in the context of the historical explanations of [42], we would obtain the clusters described below.

Cluster 1: the stamps with a thickness lower than 0.0882 (64.5% of the stamps in the sample). This first cluster would correspond to stamps produced with a medium paper, mainly printed in sheets of *LA+-F*. Cluster 2: the stamps with a thickness between 0.0882 and 0.0938 (6.8% of the stamps in the sample). Cluster 2 would correspond with the block of stamps produced with a thicker paper of *Papel Sellado*. Cluster 3: the stamps with a thickness between 0.0938 and 0.1067 (15.5% of the stamps in the sample). Due to the lack of regular paper in 1872, it is known that a temporary supply of much thicker *Papel Sellado* paper was employed only in 1872. Cluster 3 would correspond to the medium paper of that second block of stamps. Cluster 4: the stamps with a thickness greater than 0.1067 (13.2% of the stamps in the sample). Cluster 4 would correspond to the thick paper of the second block of *Papel Sellado* stamps produced in 1872.

5.3 Aircraft data

In the following example, we consider a dataset containing different characteristics of aircraft technology throughout the years 1914–1984. This data example was considered among others by [5], and it contains 709 models of aircraft and different measures such as total engine power, wingspan, length, maximum take-off weight, maximum speed, and range. From the original measures, [5] applied principal components analysis, based on the correlation matrix, to reduce the complexity of the data. The two first components account for 92% of the variation in the data. In Figure 7 (right), we reproduce the scatterplot of the first two component scores. Dividing the data into three periods (1914–1935, 1936–1955, and 1956–1984) and using the KDE (2), [5] concluded that the first two periods were bimodal, while the third had three modes.

In this section our objective is, using those principal component scores, to classify the different aircraft models. To do that, first, we need to establish the number of clusters. Here, due to the lack of formal tests for testing the number of modes in the multivariate case, and according to what is suggested at the end of Section 4.1, we searched for significant features of the KDE by applying the tools available in the R package **feature** [25]. For different matrix bandwidths, we first assessed for which locations the gradient of the smoothed curve is different from zero for a significance level $\alpha = 0.05$. When the norm of the first partial derivatives is significantly different from zero for a given location, in Figure 7 (left), we represent with an arrow the direction of the derivative. Second, the black pixels represent the locations where both the second derivative norm is significantly different from zero and the first derivative is not. We repeated those plots using different diagonal matrix bandwidths of the form $H = h^2 H_{\text{RT}}$, where H_{RT} is the normal reference rule matrix bandwidth [58, Section 6.3.1]. From the grid of considered values of h ranging from $h = 0.1$ to $h = 2$, we represented in Figure 7 (left), four interesting cases: $h = 1$, $h = 1.35$, $h = 1.45$ and $h = 1.55$. For all the considered matrix bandwidths, the largest

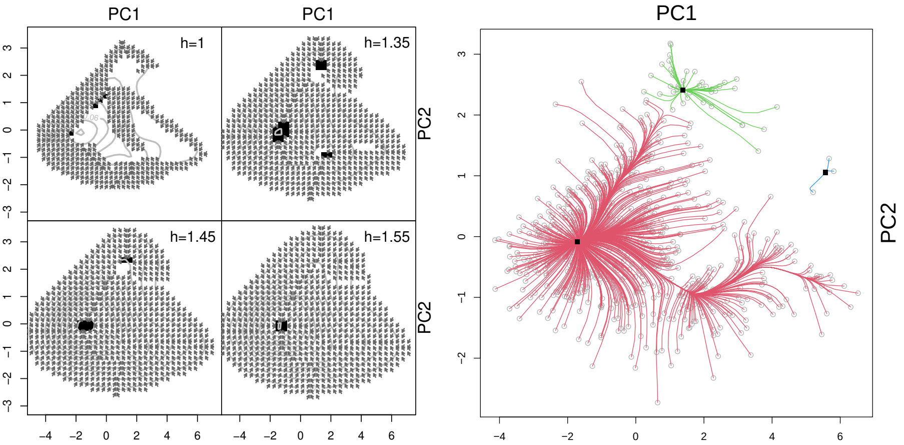

Fig. 7 First two principal component scores of the aircraft data. Left: significant features of the KDE, with Gaussian kernel and matrix bandwidth $H = h^2 H_{\mathrm{RT}}$. For each location, the arrows indicate the direction of the significant derivative and the pixels the locations where the second derivative norm is significant and the first derivative is not. Right: scatterplot of the data. After applying the mean shift algorithm for $h = 1.001$, the different colors indicate the mean shift trajectory and, therefore, which cluster each data-point belongs to. The black squares indicate the mode locations.

amount of modes that we found was equal to 3 (see the panel with $h = 1.35$). That number coincides with the findings of [5] of having at most 3 modes.

Since the critical bandwidth cannot be easily computed for the multivariate case [58, Section 9.2.4.2], we will employ the smallest value of h we found for which a trimodal estimation is obtained. That value coincides with $h = 1.001$. Thus, using the Gaussian kernel and the matrix bandwidth $H = 1.001^2 H_{\mathrm{RT}}$, we applied the mean shift algorithm described in Section 3 for clustering the data.

The three obtained clusters are represented with different colors in Figure 7 (right). As we can see in Figure 7 (right), most of the aircraft models are part of the same cluster (93.9%), which we could classify as belonging to the standard category. The other two clusters represent, respectively, 5.6% and 0.4% of the aircraft models. These last two clusters are composed of models produced in the period 1954–1982. All except two aircraft models have a maximum speed greater than 1450 km/h (the fastest aircraft can reach 3219 km/h), being this quantity for the other two aircraft models belonging to these two clusters 1038 and 1186 km/h. While only two aircraft models belonging to the first cluster reach that maximum speed, being their values 1509 and 1529 km/h. Thus, it seems that maximum speed is the main characteristic distinguishing the standard aircraft models from the other two modal clusters. Now, the three aircraft models belonging to Cluster 3 (blue color) are “bigger”, in the sense of presenting larger numbers in the remaining aircraft measures, than the models of Cluster 2 (green color). In particular, two of these three aircraft models are the *Concorde* and its Soviet equivalent the *Tupolev TU144*, which according to [5] are "unusual in combining high speed with large size".

5.4 Image segmentation

Figure 8 (left) displays a bordered grey scale photography (256 × 256 pixels) of the Cliffs of Moher, County Clare, Republic of Ireland. The photography can be described by a few simple features: There is rock, sea, sky, with a few contaminations of those (such as white spray in the sea), plus the black border. It is a relevant task in Computer Vision, for instance for automated driving and piloting systems [26], to segment an image into its key features.

It is probably fair to say that the first applied field where the mean shift has really made an impact, and caught significant attention, is that of image segmentation. This is mainly due to a sequence of papers by Dorin Comaniciu and co-workers around the change of the millenium [21–23], with the second of the listed papers exceeding 15000 citations on Google Scholar. The basic idea to use clustering for image segmentation is considerably older, dating back to the year 1979 [20]. That paper by Colemean and Andrews proposed, essentially, to segment an image into homogeneous components by (i) identifying a suitable space of features representing the image (these may include grey or 3-d color scales, the original spatial coordinates, or 'engineered' variables representing aspects such as texture or brightness); (ii) a 'feature decorrelation' step, using principal component analysis; (iii) a cluster analysis in the decorrelated space, for which [20] uses k-means; (iv) the actual segmentation step, where each original pixel is replaced by the feature vector of the center of the cluster to which it belongs. Comaniciu and Meer [22] recognized that this method may perform poorly if the clusters have non-trivial shapes, where mean shift clustering has a clear advantage compared to k-means or Gaussian mixtures. Furthermore, they reduced the space of feature variables to spatial coordinates and the color space, hence avoiding the need for step (ii).

We apply this technique now in the simplest possible form to the Cliffs of Moher image. For ease of presentation, we work with a bivariate feature space, which only contains the grey scale information (on a continuous scale between 0=black to 1=white), and the vertical coordinate information. This feature space is displayed in Figure 8 (right). In the feature space representation, the direction of the vertical axis is inverted in relation to the image representation, that is the bottom part in the feature space corresponds to the sky region, where we only have 'black' information (borders) and 'white' (sky), with the latter being less pure than the former, and so the right-bottom stretch in Figure 8 (right) being thicker than the former. It is further to be noted that we have rescaled the coordinate information to the range $[1/256, 1]$, hence allowing us to use an isotropic kernel with bandwidth matrix $H = h^2\mathbf{I}_2$, rather than a product kernel with different bandwidths for the color and the spatial component as in [22].

Having the feature space available, the task is now conceptually simple: Run Algorithm 1 on the scatterplot depicted in Figure 8 (right). Then replace the grey scale of each point by that one of the mode to which its mean shift trajectory has converged. Finally, reassemble and display the image. However, this procedure has a bottleneck: For, say, a 256 × 256 pixel image, 65536 mean shift trajectories have to be run until convergence! This is computationally infeasible, and would take several days of computation on a standard laptop. Therefore, we propose a simple but effective modification which solves this problem. This modification is based on the trivial but important insight that mean shift trajectories are 'not simply anywhere': They are, by construction, close to where the bulk of the data are. Hence, a mean shift trajectory will, on its journey towards a mode, pass close to (or, in the one-dimensional case, through) many other observations, and it is intuitively clear that most of these observations will not have much choice than joining the same trajectory, once it is their turn. Hence, one can argue that there is no actual reason to run step 2(ii) in Algorithm 1 for such observations; one can simply associate all such 'close' observations with the mode of that trajectory. In the 'mop-up' mean shift algorithm, suggested in a Tutorial at CMStatistics 2019 [28], this notion of closeness is defined by all points lying within a distance $c \times h$ around any $x^{(\ell)}$ (forming an approximate 'tube' of radius ch around the trajectory), where c is a positive constant. In a somewhat similar scheme discussed in [7], the image is discretized into cells, and all points within a cell are allocated to the mode to which the trajectory, which visited that cell first, converged. Yet another approach of relevance in this context could be the *medoid shift* [60], where – returning to our introductory fable – the element would directly move to the densest element in its vicinity (rather than a localized mean of elements), hence "ticking it off" on its way through.

In Figure 9 we apply the mop-up technique on our feature space using $h = 0.1$ and $c = 2$ (so, a quite crude allocation). The top left panel of this figure displays *all* trajectories which need to be run until all points have been allocated to a cluster, which are now only 19 instead of 65536! This computation took 110 seconds on a Intel (R) Core (TM) 2.40GHz 8GB laptop. The induced clustering, with four clusters, is shown in the top right panel, with the segmented image in the bottom left panel. We see that the segmentation is reasonable, with however some impurities in all clusters (essentially representing impurities of the raw image, as mentioned above), and some parts incorrectly classified, such as the remote cliffs being incomplete, or water being mistaken for sky. The impurities can be partially addressed through a mode filter [20]; see the bottom right panel for the outcome using the simplest possible such filter which replaces each pixel by a majority vote of its immediate four neighbors plus itself [28]. Overall, we see that a reasonable segmentation can already be achieved using these very basic means. The accuracy of the segmentation can be improved by reducing c, and the number of segments can be increased by reducing h, in both cases at the expense of an increase in computational burden.

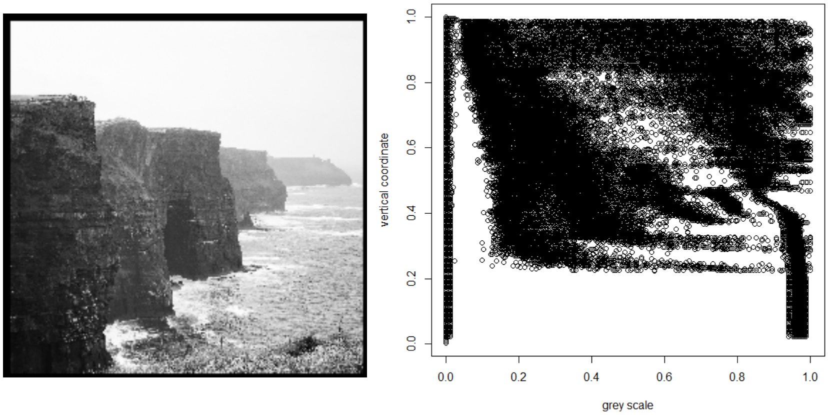

Fig. 8 Left: Cliffs of Moher; right: feature space consistent of grey scales and the vertical coordinate.

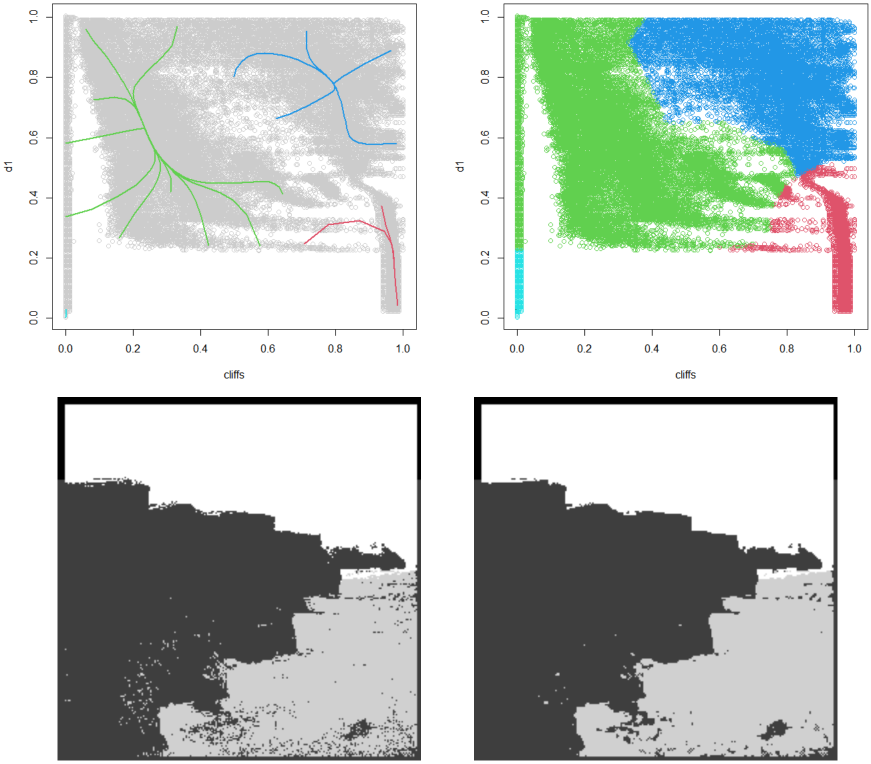

Fig. 9 Top left: Mean shift trajectories using $h = 0.1$; top right: complete clustering using mop-up; bottom left: segmented image; bottom right: result after mode filtering.

6 Discussion

While this paper is not to be understood as a full review of mean-shift based modal clustering, we have revisited the mean shift in terms of its connections with the statistical literature, in this manner hoping to contribute to an invigoration of work by Statisticians on the topic. We have also demonstrated, by means of three elaborated real data examples, how modal clustering relates in a meaningful and direct way to physical properties of the system being considered; in this respect we see a clear advantage of the mean-shift approach to other clustering techniques such as k-means or Gaussian mixtures.

To some extent, a similar line of thought appears to have driven the contribution by [12], albeit touching the term 'mean shift' only in passing. The paper by Carreira-Perpiñán [7] also contains several results which are important from a statistical point of view, including a generalized version of the discussed EM analogy, and, in their equation (5), an explicit relationship between the Hessian of f, the Jacobian of μ, and the local covariance matrix at the fixed point. Such results could be of relevance to develop a formal framework for uncertainty quantification for mean-shift based mode estimates.

Beside the mentioned reference [7], another noteworthy review-type contribution concerning modal clustering is the work by Menardi [51]. For the mean shift, [51] provides alternative references discussing some algorithmic aspects, software in languages other than R, and bandwidth selection methods.

Some other works have focused on the problem of bandwidth selection. [13] argued (in line with earlier statements in the nonparametric smoothing literature [44]), that the problem of modal estimation is more closely related to density *derivative* estimation than to density estimation itself, and that hence corresponding bandwidth selection devices should be used. [8] and [17] use kernel asymptotics, applied on the estimated density gradient, to derive optimal bandwidths for modal clustering, with the latter authors suggesting a normal-reference-type rule of of thumb based on such concepts. [14] compared the performance of different kernel density bandwidth selection routines for mean-shift based modal clustering.

The significance of density derivatives and density-derivative ratios for the mode estimation problem has also been discussed in the machine learning literature [55]. [65] discussed mean shift-based clustering with Gaussian, Cauchy, and Epanechnikov kernels, and suggested bandwidth selection based on the concept of 'decomposition stability'; i.e. the bandwidth is selected as the center of the maximum range over which no change in the modality is observed.

This is similar in spirit to the concept considered here: We employ the critical bandwidths to identify when a change in modality would occur, and then use the geometric mean of two neighboring critical bandwidths to inform a (slightly downwardly nudged) "center" of this interval. For finding the critical bandwidth a key point is to previously determine the number of modes. When the number of clusters is unknown, in the univariate setting a test for assessing that number could be employed [1]. In the multivariate case, the lack of formal procedures for testing the number of modes makes this approach less suitable.

Using a univariate testing procedure, multiple testing correction techniques could be employed (see, e.g., [40]), but depending on the kind of dependence among the variables, different algorithms would be recommended. A second alternative, employed in Section 5.3, would be to explore the number of modes with techniques similar to the SiZer [24, 35]. Another approach, employed among others by [34] and [19], would be to test for significant modes for a given matrix bandwidth. Once the number of clusters is chosen, exploratory tools such as the mode tree [45, 52], would be useful to find a "small" bandwidth for which the desired number of modes is observed. See also [58, Section 9.2.4.2], for a review on this last topic.

A different line of work, not considered at all in this exposition, has targeted the density estimate itself, rather than its bandwidth. In [41]'s contribution, a mode-flattening procedure is proposed on the back of a mixture-based density estimate to smooth out spurious and minor modes, hence allowing for more accurate estimates in regions 'where it matters'. Such methods could be of interest in conjunction with the techniques proposed in here.

Finally, we have provided some suggestions and directions for novel methodologies, including an 'inverse mean shift' technique to identify antimodes, employing them as the clusters' boundary; a concept for employing the critical bandwidth for determining the number of modes in modal clustering; and an accelerated 'mop-up' version of the mean shift algorithm. Notably, the latter contribution is immediately applicable in the multivariate case (and is arguably most useful then). Furthermore, the combination of the first two methodologies, as outlined in Algorithm 3, presents a promising approach for modal clustering in the univariate case.

In terms of limitations of the methodology, one could cite here the mentioned convergence difficulties of Algorithm 2 in the case of antimodes associated with extremely small densities; however one can argue that this will only occur in situations where the clusters are so far away from each other that there is little point in the clustering problem to start with. Hence the primary limitation of the ideas discussed in Section 4 lies in their extension to the multivariate case. As previously mentioned, formal tests for determining the number of modes in this context are lacking, and even when the number of modes is known, there are no theoretical assurances that the estimated modes will converge to the true modes with the proposed matrix bandwidth. Additionally, in the multivariate setting, the 'inverse mean shift' becomes less relevant since cluster boundaries in this context are not points but rather manifolds of co-dimension one.

Lastly, it is worth mentioning that all ingredients required to execute Algorithms 1-3 are implemented in R packages **LPCM** or **multimode**, with explicit code to reproduce the application studies in Sections 5.2 to 5.4 provided in the supplementary material. While successful in our analyses, further scrutiny of such methods from a theoretical point of view is encouraged in future research.

Supplementary information. This article is accompanied by supplementary material (S1): Replicable R Code for real data examples; (S2): Algorithm 1 is an EM algorithm, and (S3): Additional notes on convergence and EM.

References

[1] J. Ameijeiras-Alonso, R. M. Crujeiras, A. Rodríguez-Casal, Mode testing, critical bandwidth and excess mass, Test 28 (3) (2019) 900–919.

[2] J. Ameijeiras-Alonso, R. M. Crujeiras, A. Rodríguez-Casal, multimode: An R package for mode assessment, Journal of Statistical Software 97 (9) (2021), 1–32.

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

[4] A. Azzalini, G. Menardi, Clustering via Nonparametric Density Estimation: The R Package pdfCluster. Journal of Statistical Software 57(11) (2014) 1-26.

[5] A. Bowman, P. Foster, Density based exploration of bivariate data, Statistics and Computing 3 (4) (1993) 171–177.

[6] M. Carreira-Perpiñán, Gaussian mean-shift is an EM algorithm, IEEE Transactions on Pattern Analysis and Machine Intelligence 29 (2007) 767–776.

[7] M. Carreira-Perpiñán, Clustering methods based on kernel density estimators: Mean-shift algorithms, in: R. Rocci, F. Murtagh, M. Meila, C. Hennig (Eds.), Handbook of Cluster Analysis, CRC, New York, 2015.

[8] A. Casa, J. E. Chacón, G. Menardi, Modal clustering asymptotics with applications to bandwidth selection, Electronic Journal of Statistics 14 (1) (2020) 835–856.

[9] A. Casa, L. Scrucca, G. Menardi, Better than the best? Answers via model ensemble in density-based clustering. Advances in Data Analyis and Classification 15 (2021), 599–623. https://doi.org/10.1007/s11634-020-00423-6

[10] J. E. Chacón, A population background for nonparametric density-based clustering, Statistical Science 30 (4) (2015) 518–532.

[11] J. E. Chacón, Mixture model modal clustering, Advances in Data Analysis and Classification 13 (2019) 379–404.

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

[13] J. E. Chacón, T. Duong, Data-driven density derivative estimation, with applications to nonparametric clustering and bump hunting 7 (2013), 499–532.

[14] J. Chacón, P. Monfort, A comparison of bandwidth selectors for mean shift clustering, in: C. Skiadas (Ed.), Theoretical and applied issues in Statistics and Demography, ISAST, Athens, 2013.

[15] P. Chaudhuri, J. S. Marron, Sizer for exploration of structures in curves, Journal of the American Statistical Association 94 (447) (1999) 807–823.

[16] Y.-C. Chen, Modal regression using kernel density estimation: A review. WIREs Computational Statistics, 10 (2018), e1431.

[17] Y.-C. Chen, C.R. Genovese, L. Wasserman. A Comprehensive Approach to Mode Clustering, Electronic Journal of Statistics 10 (1) (2016), 210–241.

[18] Y. Cheng, Mean shift, mode seeking, and clustering, IEEE Transactions on Pattern Analysis and Machine Intelligence 17 (8) (1995) 790–799.

[19] Y. Cheng, S. Ray, Multivariate modality inference using gaussian kernel, Open Journal of Statistics 4 (5) (2014) 419–434.

[20] G. B. Coleman, H. C. Andrews, Image segmentation by clustering, Proceedings of the IEEE 67 (5) (1979) 773–785.

[21] D. Comaniciu, An algorithm for data-driven bandwidth selection, IEEE Transactions on Pattern Analysis and Machine Intelligence 25 (2) (2003) 281–288.

[22] D. Comaniciu, P. Meer, Mean shift: a robust approach toward feature space analysis, IEEE Transactions on Pattern Analysis and Machine Intelligence 24 (5) (2002) 603–619.

[23] D. Comaniciu, V. Ramesh, P. Meer, The variable bandwidth mean shift and data-driven scale selection, in: Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001, Vol. 1, 2001, pp. 438–445.

[24] T. Duong, A. Cowling, I. Koch, M. P. Wand, Feature significance for multivariate kernel density estimation, Computational Statistics & Data Analysis 52 (9) (2008) 4225–4242.

[25] T. Duong, M. Wand, feature: Local Inferential Feature Significance for Multivariate Kernel Density Estimation, R package version 1.2.13 (2015). https://CRAN.R-project.org/package=feature

[26] W. Eaton, W. Chen, Image segmentation for automated taxiing of unmanned aircraft, in: 2015 International Conference on Unmanned Aircraft Systems (ICUAS) (2015), pp. 1–8.

[27] J. Einbeck, Bandwidth selection for based unsupervised learning techniques: a unified approach via self-coverage, Journal of Pattern Recognition Research. 6 (2) (2011) 175–192.

[28] J. Einbeck, R programming and mixture models, with application to image analysis. Tutorial at CMStatistics 2019, London (2019).

[29] J. Einbeck, L. Evers, LPCM: Local Principal Curve Methods, R package version 0.46-7 (2020).
https://CRAN.R-project.org/package=LPCM

[30] J. Einbeck, L. Evers, B. Powell, Data compression and regression through local principal curves and surfaces, International Journal of Neural Systems 20 (03) (2010) 177–192.

[31] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. Institute for Computer Science, University of Munich. Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96) (1996).

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

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

[34] C. R. Genovese, M. Perone-Pacifico, I. Verdinelli, L. Wasserman, Nonparametric inference for density modes, Journal of the Royal Statistical Society. Series B (Methodological) 78 (1) (2016) 99–126.

[35] F. Godtliebsen, J. Marron, P. Chaudhuri, Significance in scale space for bivariate density estimation, Journal of Computational and Graphical Statistics 11 (1) (2002) 1–21.

[36] P. Hall, M. York, On the calibration of Silverman's test for multimodality, Statistica Sinica 11 (2) (2001) 515–536.

[37] C. Hennig, fpc: Flexible Procedures for Clustering. R package version 2.2-9 (2020).

[38] C. Hennig, N. Christlieb, Validating visual clusters in large datasets: fixed point clusters of spectral features, Computational Statistics & Data Analysis 40 (4) (2002) 723-739.

[39] C. Hennig, M. Meila, F. Murtagh, R. Rocci, Handbook of cluster analysis, CRC press, Boca Raton, 2015.

[40] Y. Hochberg, A sharper Bonferroni procedure for multiple tests of significance, Biometrika 75 (4) (1988) 800–802.

[41] S. Hu, Y. Wang, Modal Clustering Using Semiparametric Mixtures and Mode Flattening. Statistics and Computing 31, 5 (2021). https://doi.org/10.1007/s11222-020-09985-z

[42] A. J. Izenman, C. J. Sommer, Philatelic mixtures and multimodal densities, Journal of the American Statistical Association 83 (404) (1988) 941–953.

[43] N. L. Johnson, S. Kotz, N. Balakrishnan, Continuous Univariate Distributions, Volume 1 and 2, New York: Wiley Series in Probability and Statistics (1995).

[44] M. C. Jones, Rough-and-ready assessment of the degree and importance of smoothing in functional estimation, Statistica Neerlandica 54 (1) (2000) 37–46.

[45] J. Klemelä, Mode trees for multivariate data, Journal of Computational and Graphical Statistics 17 (4) (2008) 860–869.

[46] K. Lange, The MM Algorithm. In: Optimization. Springer Texts in Statistics, vol 95. Springer, New York, NY.

[47] J. Li, S. Ray, B. G. Lindsay, A nonparametric statistical approach to clustering via mode identification. Journal of Machine Learning 8 (2007), 1687–1723.

[48] P. Liu, D. Zhou, N. Wu. VDBSCAN: varied density based spatial clustering of applications with noise. In 2007 International conference on service systems and service management. IEEE (2007), pp. 1-4.

[49] G. J. McLachlan, D. Peel, Finite mixture models, John Wiley & Sons, New York, 2000.

[50] M. Meila. Criteria for comparing clusterings. In Handbook of cluster analysis (pp. 640-657). CRC Press, Boca Raton (2015).

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

[52] M. C. Minnotte, D. W. Scott, The mode tree: A tool for visualization of nonparametric density features, Journal of Computational and Graphical Statistics 2 (1) (1993) 51–68.

[53] D. W. Müller, G. Sawitzki, Excess mass estimates and tests for multimodality, Journal of the American Statistical Association 86 (415) (1991) 738–746.

[54] A. Rinaldo, L. Wasserman, Generalized density clustering, The Annals of Statistics 38 (5) (2010) 2678–2722.

[55] H. Sasaki, T. Kanamori, A. Hyvärinen, G. Niu, M. Sugiyama, Mode-seeking clustering and density ridge estimation via direct estimation of density-derivative-ratios, Journal of Machine Learning Research 18 (180) (2018) 1–47.

[56] V. Satopaa, J. Albrecht, D. Irwin B. Raghavan. Finding a “kneedle” in a haystack: Detecting knee points in system behavior. In 2011 31st international conference on distributed computing systems workshops. IEEE (2011), pp. 166-171.

[57] E. Schubert, J. Sander, M. Ester, H.-P. Kriegel, X. Xu. DBSCAN revisited, revisited: why and how you should (still) use DBSCAN. ACM Transactions on Database Systems (TODS) 42(3) (2017) 1-21.

[58] D. W. Scott, Multivariate density estimation: theory, practice, and visualization, John Wiley & Sons, New Jersey (2015).

[59] L. Scrucca, M. Fop, T. B. Murphy, A. E. Raftery, mclust 5: clustering, classification and density estimation using Gaussian finite mixture models The R Journal 8/1 (2016) 289-317.

[60] Y.A. Sheikh, E.A. Khan and T. Kanade, Mode-seeking by Medoidshifts, 2007 IEEE 11th International Conference on Computer Vision (2007), pp. 1-8.

[61] B. W. Silverman, Using kernel density estimates to investigate multimodality, Journal of the Royal Statistical Society. Series B (Methodological) 43 (1) (1981) 97–99.

[62] W. Stuetzle, R. Nugent, R., A Generalized Single Linkage Method for Estimating the Cluster Tree of a Density. Journal of Computational and Graphical Statistics, 19(2), (2010) 397–418. http://www.jstor.org/stable/25703575

[63] T. Tarpey, B. Flury, Self-consistency: a fundamental concept in statistics, Statistical Science 11 (3) (1996) 229–243.

[64] I. Wilson, Add a new dimension to your philately, The American Philatelist 97 (1983) 342–349.

[65] K.-L. Wu, M.-S. Yang, Mean shift-based clustering, Pattern Recognition 40 (11) (2007) 3035 – 3052.

[66] R. Yamasaki, T. Tanaka, Kernel Selection for Modal Linear Regression: Optimal Kernel and IRLS Algorithm, 2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA), Boca Raton, FL, USA, 2019, pp. 595-601.

[67] R. Yamasaki, T. Tanaka, Properties of mean shift, IEEE Transactions on Pattern Analysis and Machine Intelligence 42 (9) (2020) 2273–2286.