Electronic Journal of Statistics
Vol. 14 (2020) 835–856
ISSN: 1935-7524
https://doi.org/10.1214/20-EJS1679

Modal clustering asymptotics with applications to bandwidth selection

Alessandro Casa

Dipartimento di Scienze Statistiche, Università degli Studi di Padova, Padova, Italy
e-mail: casa@stat.unipd.it

and

José E. Chacón

Departamento de Matemáticas, Universidad de Extremadura, Badajoz, Spain
e-mail: jechacon@unex.es

and

Giovanna Menardi

Dipartimento di Scienze Statistiche, Università degli Studi di Padova, Padova, Italy
e-mail: menardi@stat.unipd.it

Abstract: Density-based clustering relies on the idea of linking groups to some specific features of the probability distribution underlying the data. The reference to a true, yet unknown, population structure allows framing the clustering problem in a standard inferential setting, where the concept of ideal population clustering is defined as the partition induced by the true density function. The nonparametric formulation of this approach, known as modal clustering, draws a correspondence between the groups and the domains of attraction of the density modes. Operationally, a nonparametric density estimate is required and a proper selection of the amount of smoothing, governing the shape of the density and hence possibly the modal structure, is crucial to identify the final partition. In this work, we address the issue of density estimation for modal clustering from an asymptotic perspective. A natural and easy to interpret metric to measure the distance between density-based partitions is discussed, its asymptotic approximation explored, and employed to study the problem of bandwidth selection for nonparametric modal clustering.

MSC 2010 subject classifications: Primary 62G20, 62H30; secondary 62G07.
Keywords and phrases: Nonparametric clustering, kernel estimator, mean shift clustering, plug-in bandwidth, gradient bandwidth.

Received January 2019.

1. Introduction

Clustering is commonly referred to as the task of finding groups in a set of data points (see [26], [18] or [23]). While intuitively clear, this task is, in fact, far from being accurately defined. The density-based approach attempts to circumscribe this issue by framing the problem into a statistically rigorous setting where the observed data are assumed to be realizations of a random variable, and the clusters are defined with respect to some characteristic of its underlying probability distribution.

In this sense, a clustering procedure should not be limited to simply produce a partition of the observed data; instead, it must allow obtaining a *whole-space clustering*, that is a partition of the whole sample space [3, 4]. In any case, each methodology is characterized by the way in which the clusters are defined in terms of the true distribution, leading to the concept of *ideal population clustering*. By serving as a reference "ground truth" to aim at, this concept introduces a benchmark to evaluate the performance of data-based partitions.

The ideal population goal in density-based clustering can be defined in terms of two different paradigms: the *model-based* approach, where each cluster is associated with a parametric mixture component, and the *modal* one (see respectively [30] and [32] for some recent reviews). This paper focuses on the latter formulation, whose name stems from the notion of clusters as the "domains of attraction" of the modes of the true density underlying the data [43].

Therefore, in practice density estimation assumes a key role in order to approximate the ideal population goal of modal clustering. While the modal formulation does not preclude using a parametric density estimate as a first step to perform a data-based modal clustering [5, 39], a long-standing practice resorts to nonparametric estimators. Precisely, in this paper the focus lies on those estimators based on kernel smoothing (see e.g. [8] and [47]).

Under- or over-smoothed estimates may lead to deceiving indications about the modal structure of the underlying density function, and this problem is usually quantified through some measure of the discrepancy between the estimate and the target density. In contrast, the aim of this work is to consider nonparametric density estimation as a tool for the final purpose of modal clustering, focusing on an appropriate metric comparing the partitions induced by the true and the estimated distribution.

Our main result provides an asymptotic approximation for the considered metric, which allows introducing new automatic bandwidth selection procedures specifically designed for nonparametric modal clustering. The accuracy of this approximation and the performance of the new methods in practice, with respect to the proposed error criterion, is extensively studied via simulations, and compared with some plausible competitors.

The rest of the paper is structured as follows. Section 2 formally introduces the modal approach to cluster analysis with reference also to algorithmic details. In Section 3 the distance criterion to target density estimation for modal clustering is presented, along with the main asymptotic result and its consequences. Section 4 contains the setup and results of the numerical experiments. A generalization to the multidimensional setting is discussed in Section 5. Finally, some concluding remarks are stated in Section 6.

2. Background

The connection between groups and density features, established by the modal approach to cluster analysis, allows characterizing the concept of ideal population clustering. Informally, a population cluster can be defined as the domain of attraction of a mode of the density [43]. An attempt to formalize this concept has been done in [4] with the aid of Morse Theory, a branch of differential topology focusing on the large scale structure of an object via the analysis of the critical points of a function (see e.g. [29] for an introduction).

Let us consider a continuous d-variate random variable X, with probability density function $f\colon \mathbb{R}^d \to \mathbb{R}$. Assume that f is a Morse function, i.e. a smooth enough function having nondegenerate critical points, and denote by $M_1, \ldots, M_r$ the modes of f (i.e. its local maxima). Indeed, it is common to deal with Morse functions with compact support $\overline{\{x\colon f(x) > 0\}}$, so that all considerations regarding smoothness, critical points etc, are meant with respect to the interior of such a support [29, Section 2.2]. For a given initial value $x \in \mathbb{R}^d$, an *integral curve* of the negative density gradient $-\nabla f$ is defined as the path $\nu_x : \mathbb{R} \to \mathbb{R}^d$ such that

$$\nu_x'(t) = -\nabla f(\nu_x(t)), \qquad \nu_x(0) = x.$$

The set of points whose integral curve starts at a critical point x_0 (as $t \to -\infty$) goes under the name of *unstable manifold* of x_0 and is defined as

$$W_-^u(x_0) = \{x \in \mathbb{R}^d : \lim_{t\to-\infty} \nu_x(t) = x_0\}.$$

It has been showed [44] that the class of the unstable manifolds of every critical point of a Morse function yields a partition of the whole space. With these notions at hand, the ideal population clustering $\mathcal{C} = \{\mathcal{C}_1, \ldots, \mathcal{C}_r\}$ associated to a density function f is then defined as the set of the unstable manifolds $\{W_-^u(M_1), \ldots W_-^u(M_r)\}$ of the modes of f. By borrowing concepts from terrain analysis, the underlying intuition is that, if f is figured as a mountainous landscape where the modes are the peaks, a modal cluster is the region that would be flooded by a fountain emanating from a peak. When $d = 1$, clusters are then unequivocally defined by the locations of the minima points of f, which represent the cluster boundaries.

Equivalently, if the integral curves associated to the positive density gradient are considered, then a modal cluster is defined as the set of points whose integral curves converge (as $t \to +\infty$) at the same mode. The concept of modal clusters as the domains of attraction of the density modes stems naturally from this definition. Operationally, a numerical algorithm is needed to find the eventual destination of an initial point, and most of the contributions in this direction take their steps from the mean-shift algorithm [19], essentially a variant of the gradient ascent algorithm. The algorithm transforms an initial point $x^{(0)}$ recursively, and identifies a sequence $(x^{(0)}, x^{(1)}, x^{(2)}, \ldots)$ according to an updating mechanism defined as

$$x^{(l+1)} = x^{(l)} + A\frac{\nabla f(x^{(l)})}{f(x^{(l)})} \ ,$$

where A is a $d \times d$ positive definite matrix chosen to guarantee the convergence to a local maximum of f. A partition of the data is therefore obtained by simply grouping together the observations climbing to the same density mode, via mean-shift updates.

From a practical point of view the density f is unknown, therefore an estimate is needed. When working in a nonparametric framework a common choice is given by the kernel density estimator. In the following we focus on the univariate case for ease of exposition and mathematical tractability while the multivariate extension will be addressed in Section 5 below. Let $X_1, \dots, X_n$ be a sample of i.i.d. realizations of X. Then, the kernel density estimator is defined by

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) ,$$

where K is the kernel, usually a smooth, non-negative and symmetric function integrating to one, and h is the bandwidth, which controls the smoothness of the density estimate.

While the choice of the function K is known not to have a strong impact in the performance of the estimate [41, Section 3.3.2], choosing h properly turns out to be crucial. A small value of h leads to an undersmoothed density estimate, with the possible appearance of spurious modes, while a too large value results in an oversmoothed density estimate, possibly hiding relevant features.

In order to select the smoothing parameter some measure of the distance between the estimated and the true density is needed. A common choice is the *Integrated Squared Error*, defined as

$$\mathrm{ISE}(h) = \int_{\mathbb{R}} \{\hat{f}_h(x) - f(x)\}^2 dx.$$

Depending on the observed data, the ISE is itself subject to a random variability that could hinder the problem of bandwidth selection (see [21]). Hence, its expected value

$$\mathrm{MISE}(h) = \mathbb{E}\left[\mathrm{ISE}(h)\right] \tag{2.1}$$

is alternatively considered as a non-stochastic error distance. The optimal bandwidth h_{MISE} is then defined as $h_{\mathrm{MISE}} = \mathrm{argmin}_{h>0}\, \mathrm{MISE}(h)$.

Since minimization of the MISE does not lead to closed form solutions for the optimal bandwidth, its asymptotic counterpart – the AMISE – is often considered. Both the MISE and the AMISE depend on the true, unknown density function; for this reason several different approaches to estimate them have been proposed. Examples are the ones based on *least squares cross validation*, *biased cross validation* or *plug-in bandwidth selectors*. A comprehensive review of these methods is beyond the scope of this work and, for a complete exposition, readers can refer to [47] or to the more recent book by [8].

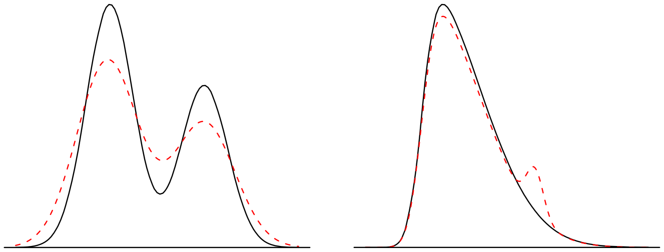


Fig 1. *Left picture: two quite different densities, from an ISE perspective, inducing the same partition of the space. Right picture: two closer densities having different number of clusters.*

3. Density estimation for modal clustering

3.1. Asymptotic bandwidth selection for modal clustering

Bandwidth selectors based on the ISE or akin distances pursue the aim of obtaining an appropriate estimate of the density. However, the goal of modal clustering is markedly different from that of density estimation (see e.g. [13]). In fact, two densities that are close with respect to the ISE may result in quite different clusterings while, on the other hand, densities far away from an ISE point of view could lead to the same partition of the space. A graphical illustration of this idea is provided in Figure 1. The inappropriateness of the ISE, or related distances, depends on its focus on the global characteristics of the density, while modal clustering strongly builds on specific and local features, more closely related to the density gradient or the high-density regions (see also [11]). Therefore, the choice of the amount of smoothing should be tailored specifically for clustering purposes.

So far, the aim of choosing an amount of smoothing for the specific task of highlighting clustering structures has been scarcely pursued in literature. A related idea, although without particular reference to cluster analysis, has been developed by [37], who propose a plug-in type bandwidth selector appropriate for estimation of highest density regions (see also [33] and [15]). Another related work, more focused on the clustering problem, is the one by [17], where the author suggests considering the self-coverage measure as a criterion for bandwidth selection. Alternatively, the potential adequacy of a bandwidth selected to properly estimate the density gradient has been pointed out informally by [6] and explored numerically by [9]. The theoretical motivation of this suggestion lies on the strong dependence of both the population modal clustering and the mean shift updating mechanism on the density gradient. The suggestion in [10] follows the same rationale and the bandwidth is proposed to be selected as a modification of the normal reference rule for density gradient estimation.

To address the problem of bandwidth selection for modal clustering, an appropriate measure of distance should compare the data-based clustering induced by a kernel density estimate with the ideal population one. Stemming from [4], a natural choice is the *distance in measure*, where the considered measure here is the probability $\mathbb{P}$ induced by the density f. Formally, let $\mathcal{C} = \{C_1, \dots, C_r\}$ and

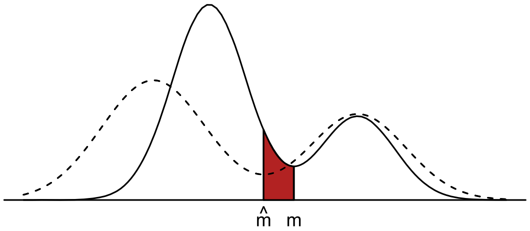


FIG 2. *Graphical interpretation of the distance in measure: the shaded area represents the probability mass that would need to be re-labeled to transform one induced clustering into the other.*

$\mathcal{D} = \{D_1, \dots, D_s\}$ be two partitions with $r \le s$ (i.e. possibly different number of groups). The distance in measure between $\mathcal{C}$ and $\mathcal{D}$ is defined as

$$d(\mathcal{C}, \mathcal{D}) = \frac{1}{2} \min_{\sigma \in \mathcal{P}_s} \left\{ \sum_{i=1}^{r} \mathbb{P}(C_i \Delta D_{\sigma(i)}) + \sum_{i=r+1}^{s} \mathbb{P}(D_{\sigma(i)}) \right\}, \tag{3.1}$$

where $C \Delta D = (C \cap D^c) \cup (C^c \cap D)$ is the symmetric difference between any two sets C and D and $\mathcal{P}_s$ denotes the set of permutations of $\{1, 2, \dots, s\}$. When $r > s$ we can easily define the distance in measure between $\mathcal{C}$ and $\mathcal{D}$ as $d(\mathcal{D}, \mathcal{C})$.

This distance finds an interpretation as the minimal probability mass that would need to be re-labeled to transform one clustering into the other (see Figure 2 for a graphical illustration). In this sense, the second term in (3.1) serves as a penalization for unmatched clusters in one of the clusterings. Practically, this distance conveys the idea that two partitions are similar not when they are physically close, but when the differently-labeled points do not represent a significant portion of the distribution.

It should be noted that the choice of this distance to evaluate the performance of a data-based clustering is not arbitrary. Indeed, many other possibilities are described in [31], but the conclusion of that study is that the distance in measure (called misclassification error there) is “the distance that comes closest to satisfying everyone”. Furthermore, in [45] the distance in measure is considered as “the most convenient choice from a theoretical point of view”.

As with the ISE-MISE duality, the distance in measure is a stochastic error distance, so for the purpose of bandwidth selection it seems more convenient to focus on the *Expected Distance in Measure*

$$\text{EDM}(h) = \mathbb{E}\big[d(\hat{\mathcal{C}}_h, \mathcal{C}_0)\big], \tag{3.2}$$

where $\hat{\mathcal{C}}_h$ is the data-based partition induced by $\hat{f}_h$ and $\mathcal{C}_0$ represents the ideal population clustering. Once the appropriate error distance is defined, the optimal bandwidth h is given by $h_{\text{EDM}} = \text{argmin}_{h>0}\, \text{EDM}(h)$.

As it happened with h_{MISE}, it does not seem possible to find an explicit expression for h_{EDM}. Hence, our goal will be to obtain an asymptotic form for the EDM that allows deriving a simple approximation to h_{EDM}.

To this aim, consider a standard normal random variable Z, and denote by $\psi(\mu,\sigma^2)=\mathbb{E}|\mu+\sigma Z|$ for $\mu\in\mathbb{R}$ and $\sigma>0$. Since $|\mu+\sigma Z|$ has a folded normal distribution [27], it follows that $\psi(\mu,\sigma^2)$ can be explicitly expressed as

$$\begin{aligned}\psi(\mu,\sigma^2) &= (2/\pi)^{1/2}\sigma e^{-\mu^2/(2\sigma^2)}+\mu\big\{1-2\Phi(-\mu/\sigma)\big\} \\ &= (2/\pi)^{1/2}\left\{\sigma e^{-\mu^2/(2\sigma^2)}+|\mu|\int_0^{|\mu|/\sigma}e^{-z^2/2}dz\right\},\end{aligned} \tag{3.3}$$

where Φ denotes the distribution function of Z. This function ψ plays a key role in the asymptotic behavior of the expected distance in measure, as the next result shows (see Appendix A for a proof).

Theorem 1. *Assume that f is a bounded Morse function with support $[\mathcal{A},\mathcal{B}]$, $r\geq 2$ modes and local minima $m_1<\cdots<m_{r-1}$, three-times continuously differentiable on $(\mathcal{A},\mathcal{B})$, with $f^{(1)}(\mathcal{A}+)>0$, $f^{(1)}(\mathcal{B}-)<0$ and $\int_{-\infty}^{\infty}|x|f(x)dx<\infty$, and that the kernel K is supported on $[-1,1]$, has four bounded derivatives and satisfies $\int_{-\infty}^{\infty}K(x)dx=1$, $\int_{-\infty}^{\infty}xK(x)dx=0$ and $\mu_2(K)=\int_{-\infty}^{\infty}x^2K(x)dx<\infty$. Define $R(K^{(1)})=\int_{-\infty}^{\infty}K^{(1)}(x)^2dx$ and suppose also that $h\equiv h_n$ is such that $h\to 0$, $nh^5/\log n\to\infty$ and $(nh^7)^{-1}$ is bounded. Then,* EDM(h) *is asymptotically equivalent to*

$$\mathrm{AEDM}(h)=\sum_{j=1}^{r-1}\frac{f(m_j)}{f^{(2)}(m_j)}\psi\Big(\tfrac{1}{2}\mu_2(K)f^{(3)}(m_j)h^2,R(K^{(1)})f(m_j)(nh^3)^{-1}\Big), \tag{3.4}$$

where $g^{(k)}$ refers to the k-th derivative of a function $g(\cdot)$.

The asymptotically optimal bandwidth h_{AEDM} is then defined as the value of $h>0$ that minimizes AEDM(h). Due to the structure of $\psi(\cdot,1)$, minimization of (3.4) is closely related to the problem of minimizing the L_1 distance in kernel density estimation and, in fact, reasoning as in [22] it is possible to show that h_{AEDM} is of order $n^{-1/7}$. Unfortunately, as it happened with h_{EDM}, it seems that neither h_{AEDM} admits an explicit representation hence, to get further insight into the problem of optimal bandwidth selection for density clustering, it appears necessary to rely on a tight upper bound for AEDM(h).

To find such a bound it is useful to note that many properties of $\psi(u,1)$ are given in [14, Ch. 5], and can be translated to our function of interest by taking into account that $\psi(\mu,\sigma^2)=\sigma\psi(\mu/\sigma,1)$. It follows that $\psi(\mu,\sigma^2)$ is symmetric with respect to μ, nondecreasing for $\mu>0$ and convex, attaining its minimum at $\mu=0$ so that $\psi(\mu,\sigma^2)\geq\psi(0,\sigma^2)=(2/\pi)^{1/2}\sigma$ for all $\mu\in\mathbb{R},\sigma>0$.

By taking into account that $e^{-\mu^2/(2\sigma^2)}$ and $|1-2\Phi(-\mu/\sigma)|$ are both bounded by 1, [14] also noted that

$$\psi(\mu,\sigma^2)\leq(2/\pi)^{1/2}\sigma+|\mu| \tag{3.5}$$

for all $\mu\in\mathbb{R},\sigma>0$. However, a tighter bound for small values of μ is given in the next lemma.

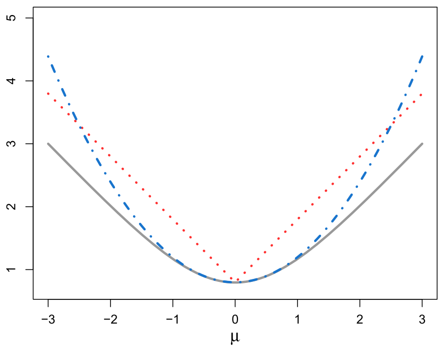


Fig 3. *Graph of $\psi(\mu, 1)$ as a function of μ (grey solid curve), together with the bound (3.5) (red dotted line) and the bound from Lemma 1 (blue dot-dashed curve).*

Lemma 1. *The bound $\psi(\mu,\sigma^2) \le (2/\pi)^{1/2}\sigma + (2\pi)^{-1/2}\mu^2/\sigma$ holds for all $\mu \in \mathbb{R}$ and $\sigma > 0$.*

The bound in Lemma 1 is tighter than (3.5) whenever $|\mu| \le (2\pi)^{1/2}\sigma$, but the situation reverses for bigger values of $|\mu|$, so that none of the two bounds is uniformly better (see Figure 3) hence we should keep track of both of them. They lead to upper bounds for the asymptotic EDM.

Corollary 1. *Under the conditions of Theorem 1, the asymptotic EDM satisfies* $\mathrm{AEDM}(h) \le \min\{\mathrm{AB1}(h), \mathrm{AB2}(h)\}$ *for all $h > 0$, where*

$$\begin{aligned}
\mathrm{AB1}(h) &= (2/\pi)^{1/2} R(K^{(1)})^{1/2} b n^{-1/2} h^{-3/2} + \tfrac{1}{2}\mu_2(K) a_1 h^2,\\
\mathrm{AB2}(h) &= (2/\pi)^{1/2} R(K^{(1)})^{1/2} b n^{-1/2} h^{-3/2} +\\
&\qquad + (32\pi)^{-1/2}\mu_2(K)^2 R(K^{(1)})^{-1/2} a_2 n^{1/2} h^{11/2}.
\end{aligned}$$

Here, $b = \sum_{j=1}^{r-1} b_j$ and $a_\ell = \sum_{j=1}^{r-1} a_{j\ell}$ and for $\ell = 1, 2$, where

$$\begin{aligned}
a_{j1} &= f(m_j)|f^{(3)}(m_j)|/f^{(2)}(m_j), &\qquad b_j &= f(m_j)^{3/2}/f^{(2)}(m_j),\\
a_{j2} &= f(m_j)^{1/2} f^{(3)}(m_j)^2/f^{(2)}(m_j).
\end{aligned}$$

The minimizers of $\mathrm{AB1}(h)$ *and* $\mathrm{AB2}(h)$ *can be computed explicitly, and are given by*

$$h_{\mathrm{AB1}} = \left(\frac{9R(K^{(1)})b^2}{2\pi\mu_2(K)^2 a_1^2}\right)^{1/7} n^{-1/7} \tag{3.6}$$

$$h_{\mathrm{AB2}} = \left(\frac{24R(K^{(1)})b}{11\mu_2(K)^2 a_2}\right)^{1/7} n^{-1/7}\ . \tag{3.7}$$

3.2. Some remarks

In this section we discuss in more depth some of the results derived in Section 3.1. The aim is to provide insights on the behavior of the approximations and bandwidth selectors and to discuss possible competitors.

Remark 1. Theorem 1 provides an asymptotic expression for the EDM that is valid as long as the true density has two or more modes. When the true density is unimodal ($r = 1$), expression (3.4) is not well-defined. However, under the assumptions of the theorem the kernel estimator is also unimodal with probability one for big enough n. Thus, asymptotically the distance in measure would be identically zero, hence the AEDM formula would remain valid under the usual convention setting $\sum_{j=1}^{0} = 0$.

Moreover, for unimodal densities the numerical work in Section 4 suggests that there exists $h_0 > 0$ such that $\mathrm{EDM}(h) = 0$ for all $h \geq h_0$. Hence, in that case it seems sensible to define $h_{\mathrm{EDM}} = \inf\{h > 0 \colon \mathrm{EDM}(h) = 0\}$.

Remark 2. A natural estimator of the density first derivative is the first derivative of the kernel density estimator. For this estimator it is possible to define the MISE as in (2.1), and to consider its minimizer $h_{\mathrm{MISE},1}$ and its asymptotic approximation $h_{\mathrm{AMISE},1}$ (see [42] and [7]). The bandwidths (3.6) and (3.7) share the same order as $h_{\mathrm{AMISE},1}$, whose expression is given by

$$h_{\mathrm{AMISE},1} = \left(\frac{3R(K^{(1)})}{\mu_2(K)^2 R(f^{(3)})} \right)^{1/7} n^{-1/7}, \tag{3.8}$$

with $R(f^{(3)}) = \int_{-\infty}^{\infty} f^{(3)}(x)^2 dx$. This consideration strengthens the intuition, outlined in Section 3.1, that (3.8) could be an adequate bandwidth choice for modal clustering purposes.

Remark 3. By explicitly plugging expression (3.3) for ψ into (3.4), it is easily seen that the AEDM can be decomposed into two summands. Studying their behavior, as a function of h, it can be checked that when $h \to 0$ the first term decreases while the second one tends to increase. Vice versa, when h increases, the opposite behavior is witnessed. A similar trade-off occurs with the decomposition of the AMISE into the *Asymptotic Integrated Squared Bias* and the *Asymptotic Integrated Variance*, which are minimized for diverging values of h.

Remark 4. If the true density is locally symmetric around its minima, the considerations in the previous item do not hold anymore. Symmetry around a minimum m implies $f^{(k)}(m) = 0$, for any odd value of k. Therefore the first summand of the AEDM expression, related to the bias, vanishes, leading to a monotonically decreasing behavior of the AEDM itself. This represents a serious issue as in principle it prevents us from using the proposed bandwidth selector. However, such a situation is highly unlikely to occur in practice, as motivated in Remark 5. A similar anomaly was observed in the related problem of mode estimation in [12]: if the true density is symmetric around its mode, then Chernoff's mode estimator is unbiased. Hence, in some special cases symmetry plays a certain role in the performance of these smoothing methodologies.

Remark 5. The derived bandwidths depend on some unknown quantities such as the true density, its local minima and its second and third derivatives. In order to be of practical use we shall resort to plug-in strategies, that is, data-based bandwidth selectors will be proposed in the next section by substituting the aforementioned unknown quantities with pilot estimates. This is the same procedure that is commonly adopted when considering the plug-in bandwidth selector $\hat{h}_{\mathrm{PI},1}$ for density gradient estimation (see [25] and [6]). With reference to Remark 4, note that due to sample variability, resorting to the considered plug-in strategy makes highly unlikely to encounter a situation of perfect symmetry around a minimum in practice.

Remark 6. Theorem 1 assumes that f is a Morse function with compact support. Since the support of a probability distribution is always a closed set, any other assumption (smoothness, critical points, etc.) is intended to be made with respect to the interior of this support. Moreover, in practice any sample takes values in a bounded set, so we may extend the applicability of Theorem 1 to densities with unbounded support, provided that we consider their *significant support* [2], i.e. a subset of the support where most of the probability mass lies. More formally, the significant support of a density f is defined as the density level set $L(c) = \{x \in \mathbb{R} : f(x) > c\}$, where $c = c_\alpha$ is the largest constant such that $\mathbb{P}(L(c_\alpha)) \geq 1-\alpha$, for some small $\alpha > 0$. Note that, by construction, the *significant support* is always bounded hence respecting the theorem's assumptions. Operational details are provided in Section 4.

4. Numerical results

The idea of estimating the density for clustering purposes, via the minimization of the expected distance in measure – or its asymptotic counterpart – is explored in this section via simulations. All the analyses have been performed in the R environment [34] with the aid of the `ks` [16], `meanShiftR` [28], `clue` [24], and `multimode` [1] packages.

A total of $B = 1000$ samples for each of the sizes $n \in \{100, 1000, 10000\}$ are generated from the univariate densities depicted in Figure 4 and whose parameters are reported in Appendix B. The selected densities are designed to illustrate different modal structures to encompass different possible behaviors from a clustering perspective. They are normal mixture densities that have been truncated to their significant support as discussed in Remark 6, with $\alpha = 0.01$, in order to respect the assumptions of Theorem 1. Hence, the samples from the truncated distributions were easily obtained by rejection sampling with respect to the untruncated models.

The first goal of the study was to evaluate the quality of the asymptotic approximation of the EDM and the behavior of the two bounds derived in Corollary 1. Since an explicit expression for the EDM was not available, we obtained a Monte Carlo approximation based on the $B = 1000$ synthetic samples.

The plots displayed in Tables 1 to 5 show the behavior of the asymptotic approximations, with respect to the EDM, as a function of the bandwidth h. As expected, the approximations improve as the sample size increases. The two bounds show a quite different behavior, with characteristics that reflect the theoretical properties pointed out in Section 3.2. The first bound is closer to the AEDM in uniform terms, but despite having a diverging behavior for large

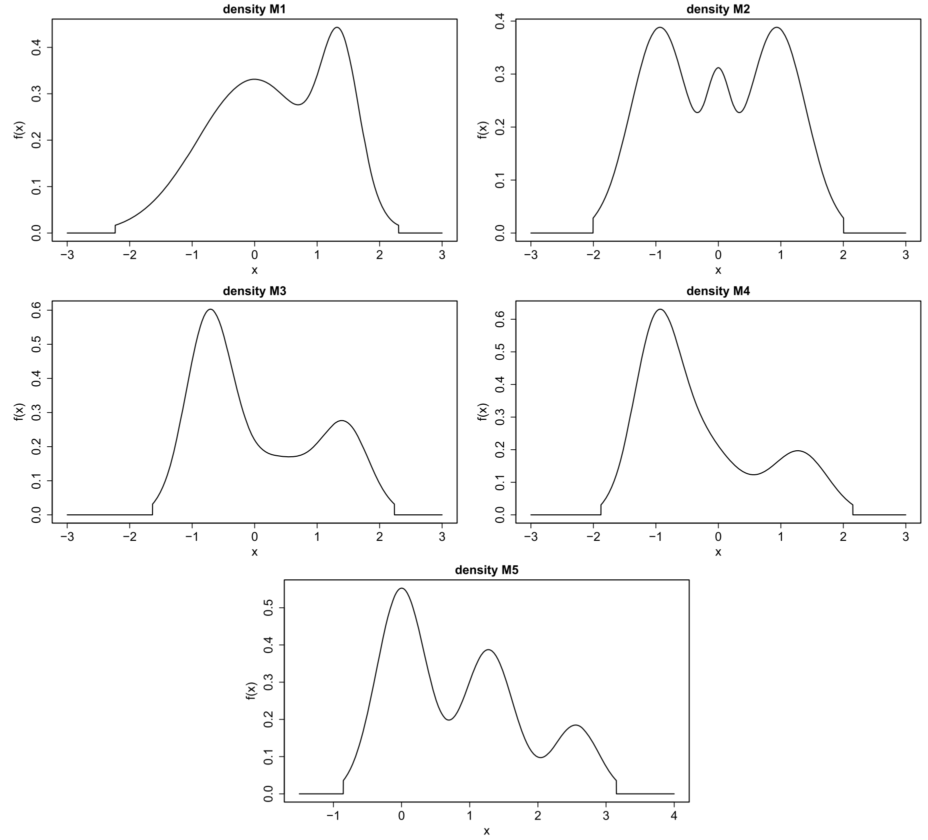


FIG 4. *Univariate density functions selected for simulations.*

h the second bound is usually closer to the AEDM around the location of the minimizer h_{AEDM}.

With regard to the EDM, it presents a nearly flat pattern around its minimizer, thus suggesting a range of plausible bandwidths with very similar performance as the optimal one. This is especially true for densities with a simpler modal structure, captured by the kernel estimate for a wide range of bandwidth values.

To appreciate how much is lost by changing the target from the optimal h_{EDM} to the oracle surrogates h_{AEDM} and $h_{\mathrm{MISE},1}$, the first three lines in each table also present the values for the corresponding EDM, all computed under a full knowledge of the density and its involved features. By construction, $\mathrm{EDM}(h_{\mathrm{EDM}})$ is the lowest of these values and, being derived as an asymptotic approximation, the oracle h_{AEDM} stands close to this optimal value, especially for larger sample sizes. However, it is remarkable that $h_{\mathrm{MISE},1}$, despite being based on a different optimality criterion, also leads to comparable or even improved results over h_{AEDM} in terms of the EDM.

Table 1

Top panel: the EDM (solid line), the AEDM (dashed grey line), and the bounds AB1 (dotted line) and AB2 (dot-dashed line) versus h, for $n = 100, 1000, 10000$. All the expressions are evaluated by assuming f and all the involved quantities known. The minimum EDM is reported below the plots, together with the EDM for the oracle bandwidths h_{AEDM} and $h_{\mathrm{MISE},1}$. Middle panel: average distances in measure (and their standard error) for the proposed bandwidth selectors and the plug-in bandwidth for density gradient estimation. Bottom panel: percentages of times when the estimated number of cluster $\hat{r}$ matches the true one r. Results refer to density M1.

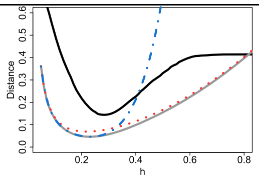
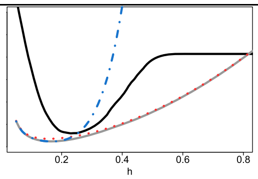
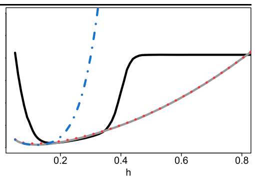
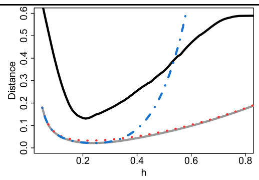
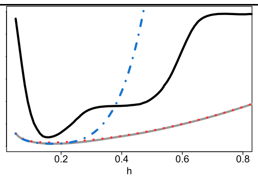
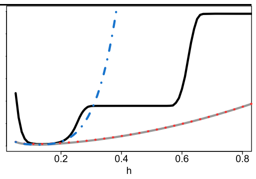
	$n = 100$	$n = 1000$	$n = 10000$
			
h_{EDM}	0.144	0.060	0.020
h_{AEDM}	0.164	0.103	0.050
$h_{\mathrm{MISE},1}$	0.146	0.081	0.044
$\hat{h}_{\mathrm{AEDM}}$	0.267 (0.173)	0.103 (0.130)	0.045 (0.075)
$\hat{h}_{\mathrm{AB1}}$	0.256 (0.174)	0.105 (0.127)	0.056 (0.084)
$\hat{h}_{\mathrm{AB2}}$	0.265 (0.173)	0.102 (0.129)	0.048 (0.079)
$\hat{h}_{\mathrm{PI},1}$	0.221 (0.176)	0.063 (0.084)	0.029 (0.052)
$\% \hat{r} = r$	54.5	91.7	92.6

Table 2

Cf. Table 1. Results refer to density M2.

	$n = 100$	$n = 1000$	$n = 10000$
			
h_{EDM}	0.131	0.040	0.008
h_{AEDM}	0.143	0.047	0.008
$h_{\mathrm{MISE},1}$	0.165	0.041	0.011
$\hat{h}_{\mathrm{AEDM}}$	0.324 (0.200)	0.061 (0.070)	0.010 (0.016)
$\hat{h}_{\mathrm{AB1}}$	0.301 (0.195)	0.053 (0.066)	0.011 (0.017)
$\hat{h}_{\mathrm{AB2}}$	0.318 (0.199)	0.058 (0.069)	0.010 (0.016)
$\hat{h}_{\mathrm{PI},1}$	0.256 (0.159)	0.092 (0.076)	0.008 (0.005)
$\% \hat{r} = r$	2.8	58.0	100.0

As a second goal, we propose new data-based bandwidth selectors specifically designed for modal clustering purposes. The first step consists in estimating the number of local minima, and their location. This is achieved by numerically finding the roots of a pilot estimate of $f^{(1)}$, constructed as the derivative of the kernel density estimator using the plug-in gradient bandwidth $\hat{h}_{\mathrm{PI},1}$. Then, similarly, we obtain pilot estimates of f, $f^{(2)}$ and $f^{(3)}$ at the estimated local minima using kernel estimates with the same bandwidth $\hat{h}_{\mathrm{PI},1}$. These quanti-

TABLE 3
Cf. Table 1. Results refer to density M3.

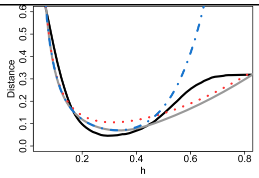
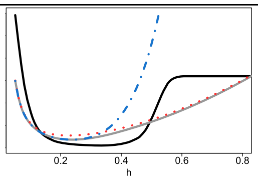
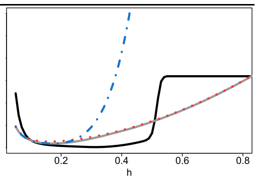
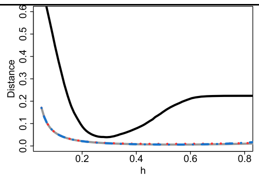
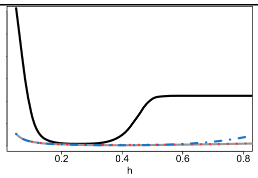
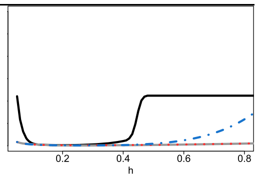
	$n = 100$	$n = 1000$	$n = 10000$
			
h_{EDM}	0.045	0.010	0.003
h_{AEDM}	0.051	0.016	0.011
$h_{\mathrm{MISE},1}$	0.054	0.034	0.022
$\hat{h}_{\mathrm{AEDM}}$	0.090 (0.110)	0.039 (0.057)	0.026 (0.036)
$\hat{h}_{\mathrm{AB1}}$	0.087 (0.104)	0.042 (0.058)	0.028 (0.035)
$\hat{h}_{\mathrm{AB2}}$	0.091 (0.109)	0.040 (0.058)	0.026 (0.036)
$\hat{h}_{\mathrm{PI},1}$	0.050 (0.072)	0.024 (0.025)	0.019 (0.017)
$\% \hat{r} = r$	91.0	91.6	88.1

TABLE 4
Cf. Table 1. Results refer to density M4.

	$n = 100$	$n = 1000$	$n = 10000$
			
h_{EDM}	0.039	0.009	0.003
h_{AEDM}	0.187	0.040	0.005
$h_{\mathrm{MISE},1}$	0.040	0.015	0.005
$\hat{h}_{\mathrm{AEDM}}$	0.077 (0.088)	0.030 (0.057)	0.007 (0.016)
$\hat{h}_{\mathrm{AB1}}$	0.074 (0.086)	0.029 (0.053)	0.009 (0.021)
$\hat{h}_{\mathrm{AB2}}$	0.076 (0.089)	0.030 (0.057)	0.007 (0.017)
$\hat{h}_{\mathrm{PI},1}$	0.051 (0.069)	0.011 (0.014)	0.005 (0.005)
$\% \hat{r} = r$	85.4	97.2	99.8

ties are subsequently plugged-in in the formulas of the AEDM, AB1 and AB2, and the minimizers of the resulting estimated criteria are found; in the case of the estimated AEDM by numerical minimization, and according to expressions (3.6) and (3.7) for AB1 and AB2 respectively. The data-based bandwidths thus obtained are denoted $\hat{h}_{\mathrm{AEDM}}$, $\hat{h}_{\mathrm{AB1}}$ and $\hat{h}_{\mathrm{AB2}}$, respectively.

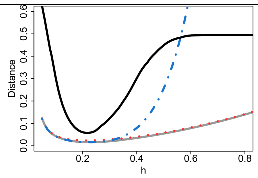
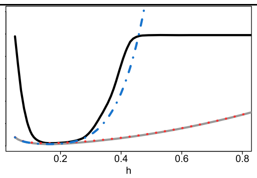
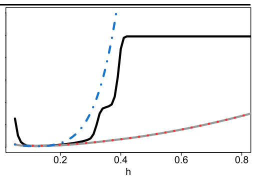
Occasionally (although rarely) the first step in the procedure above yielded a single mode, and then the AEDM was undefined. In those cases, and according to the rationale exposed in Remark 1, a sensible choice for h is the *critical bandwdith* proposed by [40],

$$\hat{h}_{\mathrm{crit}} = \inf\{h > 0 : \hat{f}_h(\cdot) \text{ has exactly one mode}\},$$

so in that case we set $\hat{h}_{\mathrm{AEDM}} = \hat{h}_{\mathrm{AB1}} = \hat{h}_{\mathrm{AB2}} = \hat{h}_{\mathrm{crit}}$.

Table 5

Cf. Table 1. Results refer to density M5.

	$n = 100$	$n = 1000$	$n = 10000$
			
h_{EDM}	0.058	0.012	0.005
h_{AEDM}	0.059	0.012	0.005
$h_{\text{MISE},1}$	0.058	0.013	0.005
$\hat{h}_{\text{AEDM}}$	0.160 (0.175)	0.017 (0.034)	0.006 (0.007)
$\hat{h}_{\text{AB1}}$	0.144 (0.169)	0.017 (0.030)	0.006 (0.007)
$\hat{h}_{\text{AB2}}$	0.157 (0.174)	0.017 (0.034)	0.006 (0.007)
$\hat{h}_{\text{PI},1}$	0.179 (0.158)	0.013 (0.009)	0.005 (0.003)
$\% \hat{r} = r$	42.7	99.7	100.0

Tables 1 to 5 also contain the Monte Carlo averages and standard deviations of the distances in measure obtained when performing modal clustering using the bandwidth selectors $\hat{h}_{\text{AEDM}}$, $\hat{h}_{\text{AB1}}$ and $\hat{h}_{\text{AB2}}$. For completeness, their performance is also compared to that of $\hat{h}_{\text{PI},1}$, which so far probably represents their most sensible competitor in the clustering framework (see [9]).

In general, $\hat{h}_{\text{AB1}}$ and $\hat{h}_{\text{AB2}}$ led to more accurate clusterings than $\hat{h}_{\text{AEDM}}$, with a slight preference for $\hat{h}_{\text{AB1}}$. The gradient-based bandwidth $\hat{h}_{\text{PI},1}$, in turn, not only produces competitive results, but its Monte Carlo average distance in measure appears lower than the one produced by the asymptotic EDM minimizers. In fact, a deeper insight into the standard errors of the obtained distances shows that $\hat{h}_{\text{AEDM}}$, as well as $\hat{h}_{\text{AB1}}$ and $\hat{h}_{\text{AB2}}$, produce more variable results. The higher variability seems to be due to the sensitivity of the minimizers to the plugged in pilot estimates, which strongly depend on local features of the density. Some further investigations, not fully reported here, suggest that the main responsible for this behavior is not the pilot estimate of the local minima but the pilot density derivatives estimates at the minimum points. Also, due to the use of different pilot bandwidths to estimate the unknown m_j, $f^{(2)}$, and $f^{(3)}$, it may occur, indeed, that $\hat{f}^{(2)}(\hat{m}_j)$ assumes even negative values. On the other hand, while relying as well on some plug-in estimates, the gradient-based bandwidth $\hat{h}_{\text{PI},1}$ produces more robust clusterings, as the quantities to be estimated refer conversely to global features of the density. As expected, this diverging behavior tends to vanish with increasing sample size since the asymptotic approximations improve. As a confirmation, with $n = 10000$, all the considered bandwidths perform comparably.

5. Multidimensional generalization

The concepts discussed so far refer to the one-dimensional setting where a mathematically rigorous treatment is feasible. The multidimensional generalization poses some difficulties since obtaining an asymptotic approximation of the EDM appears far from trivial. Hence, in order to gain some insight into the problem of selecting the amount of smoothing for nonparametric clustering in more than one dimension, some numerical comparisons are performed assuming the true density as known.

Denote by $f : \mathbb{R}^d \to \mathbb{R}$ the true density function and by

$$\hat{f}_{\mathbf{H}}(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} |\mathbf{H}|^{-1/2} K \left(\mathbf{H}^{-1/2}(\mathbf{x} - \mathbf{X}_i) \right) \ , \tag{5.1}$$

its kernel estimate based on a sample $\mathbf{X}_1, \dots, \mathbf{X}_n$ and indexed by a symmetric positive definite $d \times d$ bandwidth matrix $\mathbf{H}$. The problem of bandwidth selection is considered by studying the EDM between the clustering induced by the kernel estimate $\hat{\mathcal{C}}_{\mathbf{H}}$ and the ideal population clustering $\mathcal{C}_0$. These clusterings are not so easily identifiable as in the unidimensional setting, due to the arbitrary forms that the cluster boundaries may adopt, however an approximation of the distance in measure $d(\hat{\mathcal{C}}_{\mathbf{H}}, \mathcal{C}_0)$ can be computed by resorting to a discretization scheme as follows (see [9] for further details):

1. Take a grid over the sample space and rule the grid by considering hyper-rectangles centered at each grid point.
2. Assign a cluster membership to each grid point by running a population version of the mean-shift algorithm i.e. using the true density. This produces a discretized version of $\mathcal{C}_0$.
3. Similarly, obtain the data-based partition $\hat{\mathcal{C}}_{\mathbf{H}}$ induced by $\hat{f}_{\mathbf{H}}$.
4. Compute the probability mass of each single hyper-rectangle in $\mathcal{C}_0$.
5. Compute the distance in measure as in (3.2) where the involved probabilities are evaluated based on the previous step.

For the multidimensional simulation study, a total of $B = 1000$ samples for each of the sizes $n \in \{100, 1000\}$ were generated from the bivariate densities whose contour plots are shown in Figure 5 and described in Appendix B. The densities have been chosen to generalize the settings M1 and M5 included in the univariate study.

Three different parametrizations for the bandwidth matrix were considered: a scalar bandwidth $\mathbf{H} = h^2\mathbf{I}$, with $\mathbf{I}$ the identity matrix, a diagonal bandwidth $\mathbf{H} = \operatorname{diag}(h_1^2, h_2^2)$, and a full, unconstrained bandwidth matrix $\mathbf{H}$. For density and density derivative estimation, [46] and [7] showed that the use of the simplest scalar bandwidth can be quite detrimental in practice, a diagonal bandwidth may suffice in some scenarios, but in general it is advantageous to employ unconstrained bandwidth matrices (see also [8]). However, such results have never been obtained in a modal clustering framework; thus one of the goals of this simulation study is to examine how the bandwidth matrix parametrization affects the performances of the procedures.

Using the synthetic samples from each density in the study, it was possible to obtain a Monte Carlo estimate of the (discretized version of the) EDM, which was then minimized over the class of scalar, diagonal and unconstrained band-

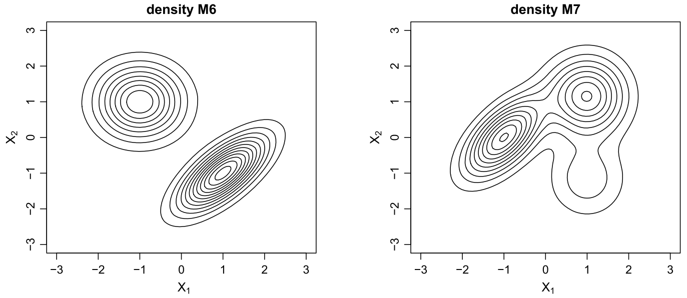


FIG 5. *Bivariate density functions selected for simulations.*

TABLE 6

Minimum EDM associated with a density estimate with bandwidth matrix $\mathbf{H}$ *selected to minimize the EDM (*$\mathbf{H}_{\text{EDM}}$*) and the MISE for gradient estimation (*$\mathbf{H}_{\text{MISE},1}$*). Different parametrizations for* $\mathbf{H}$ *are considered. In both cases, the true density as well as all the involved quantities are assumed to be known. Results refer to density M6.*

	$\mathbf{H}_{\text{EDM}}$		$\mathbf{H}_{\text{MISE},1}$	
	$n = 100$	$n = 1000$	$n = 100$	$n = 1000$
$\begin{pmatrix} h^2 & 0 \\ 0 & h^2 \end{pmatrix}$	0.006	0.004	0.064	0.040
$\begin{pmatrix} h_1^2 & 0 \\ 0 & h_2^2 \end{pmatrix}$	0.006	0.004	0.064	0.040
$\begin{pmatrix} h_1^2 & h_{12} \\ h_{12} & h_2^2 \end{pmatrix}$	0.005	0.003	0.042	0.024

TABLE 7

Cf. Table 6. Results refer to density M7.

	$\mathbf{H}_{\text{EDM}}$		$\mathbf{H}_{\text{MISE},1}$	
	$n = 100$	$n = 1000$	$n = 100$	$n = 1000$
$\begin{pmatrix} h^2 & 0 \\ 0 & h^2 \end{pmatrix}$	0.114	0.044	0.116	0.054
$\begin{pmatrix} h_1^2 & 0 \\ 0 & h_2^2 \end{pmatrix}$	0.114	0.042	0.115	0.055
$\begin{pmatrix} h_1^2 & h_{12} \\ h_{12} & h_2^2 \end{pmatrix}$	0.110	0.040	0.121	0.054

width matrices. The EDM was computed also for the MISE-optimal bandwidth for density gradient estimation over the same matrix classes. In both cases, the true density as well as all the involved quantities were assumed to be known. The EDM minimizers were determined numerically, by running the procedure over a grid of sensible values of the entries, while the optimal matrices for gradient estimation were determined as in [7].

The results are reported in Tables 6 and 7. Clustering based on the optimal bandwidth according to the EDM is very accurate in both of the considered examples, and improves considerably for increasing sample size. The use of more complex bandwidth parametrizations does not seem worth for modal clustering since results obtained with a full, unconstrained bandwidth matrix are comparable with those obtained with a scalar bandwidth, while the latter requires a substantially smaller computational effort.

In the multidimensional setting, the gradient bandwidth is quite competitive in terms of EDM, as in the univariate case. Again the comparable performance of unconstrained bandwidth matrices does not seem to justify the use of more complex parametrizations.

6. Conclusions

The modal clustering methodology provides a framework to perform cluster analysis with a clear and explicit population goal. It allows clusters of arbitrary shape and size, which can be captured by means of a nonparametric density estimator. In this context, the distance in measure represents a natural and easily interpretable error criterion. Therefore, in this paper we have presented an asymptotic study of this criterion for the case where density estimates of kernel type are employed to obtain a whole-space clustering via the mean shift algorithm.

Our asymptotic approximations are useful to gain insight into the fundamental problem of bandwidth selection for modal clustering and, at the same time, serve as the basis to propose practical data-based bandwidth choices specifically designed for clustering purposes.

The finite-sample performance of the new proposals was investigated in a thorough simulation study, and compared to the oracle bandwidths i.e. the optimal choices when the true population is fully known. The gradient bandwidth, designed for the closely related problem of density gradient estimation, was also included as a natural competitor in the study.

The results of this simulation study have suggested that all the methods perform quite satisfactorily, and exhibit a very similar behavior for large sample sizes. For smaller samples, the performance of the gradient bandwidth was rather remarkable, since it obtained comparable or even better results than the new proposals, even without being specifically conceived for modal clustering.

This phenomenon resembles the conclusions obtained in [36] regarding the related problem of level set estimation. There, it was shown that the traditional bandwidth selectors for density estimation often outperformed more sophisticated methods designed for level set estimation purposes. The common pattern in both situations is that the optimal choices for the specific problems (level set estimation and modal clustering, respectively) depend on very subtle local features of the unknown density function, which are difficult to estimate, so that choices based on a more global, yet somehow related, perspective represent a sensible alternative.

Appendix A: Proofs

Proof of Theorem 1. From Theorem 4.1 in [4] it follows that, with probability one, there exists $n_0 \in \mathbb{N}$ such that the kernel density estimator $\widehat{f}_h$ has the same number of local minima as f for all $n \geq n_0$. Let us denote by $\widehat{m}_{h,1} < \cdots <$

$\widehat{m}_{h,r-1}$ the local minima of $\widehat{f}_h$. Then, the expected distance in measure between the data-based clustering $\widehat{\mathcal{C}}_h$ and the population clustering $\mathcal{C}_0$ can be written as

$$\mathrm{EDM}(h) = \sum_{j=1}^{r-1} \mathbb{E}|F(\widehat{m}_{h,j}) - F(m_j)|. \tag{A.1}$$

Write, generically, $\widehat{m}$ and m for any of the estimated and true local minima. A Taylor expansion with integral remainder allows writing

$$F(\widehat{m}) - F(m) = (\widehat{m} - m)\int_0^1 f\big(m + t(\widehat{m} - m)\big)dt.$$

The assumptions imply that $\widehat{m} \to m$ almost surely [see, for instance, 35] and, since f is bounded and continuous, this readily yields $\int_0^1 f\big(m + t(\widehat{m} - m)\big)dt \to f(m)$ almost surely, which entails that $\mathbb{E}|F(\widehat{m}) - F(m)| \sim f(m)\mathbb{E}|\widehat{m} - m|$. The result then follows from Equation (2.6) in [20], where the asymptotic form of $\mathbb{E}|\widehat{m} - m|$ is given. □

Proof of Lemma 1. From $\psi(\mu, \sigma^2) = \sigma\psi(\mu/\sigma, 1)$, it suffices to show that $\psi(u, 1) \leq (2/\pi)^{1/2} + (2\pi)^{-1/2}u^2$ for $u \geq 0$. From the definition of ψ, this is equivalent to proving that $\alpha(u) \leq 1$, where $\alpha(u) = e^{-u^2/2} + u\int_0^u e^{-z^2/2}dz - u^2/2$. Since $\alpha(0) = 1$, it is enough to show that α is nonincreasing, but this immediately follows from the fact that $\alpha'(u) = \int_0^u e^{-z^2/2}dz - u$. □

Appendix B: Parameter settings

In the following the parameter settings of the densities selected for the simulations are presented. Since all the densities are mixtures of Gaussian models, we adopt the usual notation where, for a given k component, π_k represents the *k-th* mixture weight, μ_k and σ_k^2 (Σ_k for the bivariate models) the mean and variance (covariance matrix).

B.1. Unidimensional parameter settings

B.1.1. Density M1

Components	π_k	μ_k	σ_k^2
1	0.75	0.00	0.83
2	0.25	1.37	0.09

B.1.2. Density M2

Components	π_k	μ_k	σ_k^2
1	0.45	-0.93	0.22
2	0.45	0.93	0.22
3	0.1	0.00	0.04

B.1.3. Density M3

Components	π_k	μ_k	σ_k^2
1	0.5	-0.74	0.14
2	0.3	0.37	0.55
3	0.2	1.47	0.14

B.1.4. Density M4

Components	π_k	μ_k	σ_k^2
1	0.15	0.00	0.44
2	0.15	-0.33	0.19
3	0.5	-0.99	0.14
4	0.2	1.32	0.19

B.1.5. Density M5

Components	π_k	μ_k	σ_k^2
1	0.5	0.00	0.14
2	0.35	1.28	0.14
3	0.15	2.56	0.11

B.2. Bidimensional settings

B.2.1. Asymmetric bimodal

Components	π_k	μ_k	Σ_k
1	0.5	$\begin{pmatrix} 1 \\ -1 \end{pmatrix}$	$\begin{pmatrix} 0.44 & 0.31 \\ 0.31 & 0.44 \end{pmatrix}$
2	0.5	$\begin{pmatrix} -1 \\ 1 \end{pmatrix}$	$\begin{pmatrix} 0.44 & 0 \\ 0 & 0.44 \end{pmatrix}$

B.2.2. Trimodal

Components	π_k	μ_k	Σ_k
1	0.43	$\begin{pmatrix} -1 \\ 0 \end{pmatrix}$	$\begin{pmatrix} 0.36 & 0.25 \\ 0.25 & 0.49 \end{pmatrix}$
2	0.43	$\begin{pmatrix} 1 \\ 1.15 \end{pmatrix}$	$\begin{pmatrix} 0.36 & 0 \\ 0 & 0.49 \end{pmatrix}$
3	0.14	$\begin{pmatrix} 1 \\ -1.15 \end{pmatrix}$	$\begin{pmatrix} 0.36 & 0 \\ 0 & 0.49 \end{pmatrix}$

References

[1] Ameijeiras-Alonso, J. and Crujeiras, R. M. and Rodriguez-Casal, A. (2018). Multimode: An R Package for Mode Assessment. *arXiv preprint arXiv:***1803.00472**.

[2] Baillo, A., Cuesta-Albertos, J. A. and Cuevas, A. (2001). Convergence rates in nonparametric estimation of level sets. *Statistics & probability letters*. **53(1)** 27–35. MR1843338

[3] Ben-David, S., von Luxburg, U. and Pál, D. (2006). A sober look at clustering stability. In *Proceedings of the 19th Annual Conference on Learning Theory (G. Lugosi and H.-U. Simon, eds.)*, pp. 5–19. Springer. MR2277915

[4] Chacón, J. E. (2015). A population background for nonparametric density-based clustering. *Statistical Science*. **30(4)** 518–532. MR3432839

[5] Chacón, J. E. (2019). Mixture model modal clustering. *Advances in Data Analysis and Classification*. **13(2)** 379–404. MR3954514

[6] Chacón, J. E. and Duong, T. (2013). Data-driven density derivative estimation, with applications to nonparametric clustering and bump hunting. *Electronic Journal of Statistics*. **7** 499–532. MR3035264

[7] Chacón, J. E. and Duong, T. and Wand, M. P. (2011). Asymptotics for general multivariate kernel density derivative estimators. *Statistica Sinica*. **21** 807–840. MR2829857

[8] Chacón, J. E. and Duong, T. (2018). *Multivariate Kernel Smoothing and Its Applications*. Chapman & Hall. MR3822372

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

[10] Chen, Y.-C., Genovese, C. R. and Wasserman, L. (2016). A comprehensive approach to mode clustering. *Electronic Journal of Statistics*. **10(1)** 210–241. MR3466181

[11] Chen, Y.-C., Genovese, C. R. and Wasserman, L. (2017). Statistical inference using the Morse-Smale complex. *Electronic Journal of Statistics*. **11(1)** 1390–1433. MR3635917

[12] Chernoff, H. (1964). Estimation of the mode. *Annals of the Institute of Statistical Mathematics*. **16** 31–41. MR0172382

[13] Cuevas, A., Febrero, M. and Fraiman, R. (2001). Cluster analysis: a further approach based on density estimation. *Computational Statistics & Data Analysis*. **36(4)** 441–459. MR1855727

[14] Devroye, L. and Györfi, L. (1985). *Nonparametric Density Estimation: the L_1 View Wiley, New York*. MR0780746

[15] Doss, C. R. and Weng, G. (2018). Bandwidth selection for kernel density estimators of multivariate level sets and highest density regions. *Electronic Journal of Statistics*. **12(2)** 4313–4376. MR3892342

[16] Duong, T. (2018). *ks: Kernel Smoothing* URL `https://CRAN.R-project.org/package=ks` *R package version 1.11.3*.

[17] Einbeck, J. (2011). Bandwidth selection for mean-shift based unsupervised learning techniques: a unified approach via self-coverage. *Journal of pattern recognition research*. **6(2)** 175–192.

[18] Everitt, B. S., Landau, S., Leese, M. and Sthal, D. (2011). *Cluster Analysis. (5th Ed.)*. John Wiley & Sons, Inc. MR3155074

[19] Fukunaga, K. and Hostetler, L. (1975). The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Transactions on information theory.* **21(1)** 32–40. MR0388638

[20] Grund, B. and Hall, P. (1995). On the minimisation of the L^p error in mode estimation. *Annals of Statistics* **23** 2265–2284. MR1389874

[21] Hall, P. and Marron, J. S. (1991). Lower bounds for bandwidth selection in density estimation. *Probability Theory and Related Fields* **90** 149–173. MR1128068

[22] Hall, P. and Wand, M. P. (1988). On the minimization of absolute distance in kernel density estimation. *Statistics and Probability Letters* **6** 311–314. MR0933288

[23] Hennig, C., Meila, M., Murtagh, F. and Rocci, R. (2016). *Handbook of Cluster Analysis*. Chapman & Hall. MR3644705

[24] Hornik, K. (2018). *Clue: Cluster ensembles. URL https://CRAN.R-project.org/package=clue R package version 0.3-55.*

[25] Jones, M. C. (1992). Potential for automatic bandwidth choice in variations on kernel density estimation. *Statistics and Probability Letters* **13** 351–356.

[26] Kaufman, L. and Rousseeuw, P. J. (2005). *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley & Sons, Inc. MR1044997

[27] Leone, F. C., Nelson, L. S. and Nottingham, R. B. (1961). The folded normal distribution. *Technometrics* **3** 543–550. MR0130737

[28] Lisic, J. (2018). *MeanShiftR: A Computationally Efficient Mean Shift Implementation. URL https://CRAN.R-project.org/package=meanShiftR. R package version 0.52.*

[29] Matsumoto, Y. (2002). *An introduction to Morse Theory. American Mathematical Society*. MR1873233

[30] McNicholas, P. D. (2016). Model-based clustering. *Journal of Classification.* **33(3)** 331–373. MR3575621

[31] Meilă, M. (2016). Criteria for comparing clusterings. In C. Hennig, M. Meilă, F. Murtagh and R. Rocci (Eds.), *Handbook of Cluster Analysis* 619–635. CRC Press. MR3644730

[32] Menardi, G. (2016). A review on modal clustering. *International Statistical Review* **84(3)** 413–433. MR3580423

[33] Qiao, W. (2020). Asymptotics and optimal bandwidth selection for nonparametric estimation of density level sets. *Electronic Journal of Statistics* **14(1)** 302–344. MR4048601

[34] R Core Team (2018) *R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.*

[35] Romano, J. P. (1988). On weak convergence and optimality of kernel density estimates of the mode. *Annals of Statistics* **16** 629–647. MR0947566

[36] Saavedra-Nieves, P., González-Manteiga, W. and Rodríguez-Casal, A. (2014). Level set estimation. *In Topics in Nonparametric Statistics (M. G. Akritas, S. N. Lahiri and D. N. Politis, eds.). Springer Proceedings in Mathematics & Statistics* **74** 299–307. MR3333356

[37] Samworth, R. J. and Wand, M. P. (2010). Asymptotics and optimal bandwidth selection for highest density region estimation. *Annals of Statistics* **38(3)** 1767–1792. MR2662359

[38] Scott, D. W. (2015). Multivariate density estimation: theory, practice and visualization. John Wiley & Sons. MR3329609

[39] Scrucca, L. (2016). Identifying connected components in Gaussian finite mixture models for clustering. *Computational Statistics & Data Analysis* **93** 5–17. MR3406192

[40] Silverman, B. W. (1981). Using kernel density estimates to investigate multimodality. *Journal of the Royal Statistical Society. Series B* **43** 97–99. MR0610384

[41] Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman & Hall. MR0848134

[42] Singh, R. S. (1987). MISE of kernel estimates of a density and its derivatives. *Statistics and Probability Letters*. **5** 153–159.

[43] Stuetzle, W. (2003). Estimating the cluster tree of a density by analyzing the minimal spanning tree of a sample. *Journal of Classification*. **20(1)** 25–47. MR1983120

[44] Thom, R. (1949). Sur une partition en cellules associée à une fonction sur une variété. *Comptes Rendus Hebdomadaires des Séances de l'Académie des Sciences*, **228** 973–975. MR0029160

[45] von Luxburg, U. (2010). Clustering stability: an overview. *Foundations and Trends in Machine Learning*, **2** 235–274.

[46] Wand, M. P. and Jones, M. C. (1993). Comparison of smoothing parameterizations in bivariate kernel density estimation. *Journal of the American Statistical Association* **88(422)** 520–528. MR1224377

[47] Wand, M. P. and Jones, M. C. (1995). *Kernel smoothing*. Chapman & Hall. MR1319818