Electronic Journal of Statistics
Vol. 12 (2018) 4288–4312
ISSN: 1935-7524
https://doi.org/10.1214/18-EJS1510

Analysis of a mode clustering diagram*

Isabella Verdinelli

Department of Statistics and Data Science
Pittsburgh, PA USA
e-mail: isabella@stat.cmu.edu
url: www.stat.cmu.edu/~isabella

Larry Wasserman

Department of Statistics and Data Science
Pittsburgh, PA USA
e-mail: larry@stat.cmu.edu
url: www.stat.cmu.edu/~larry

Abstract: Mode-based clustering methods define clusters in terms of the modes of a density estimate. The most common mode-based method is mean shift clustering which defines clusters to be the basins of attraction of the modes. Specifically, the gradient of the density defines a flow which is estimated using a gradient ascent algorithm. Rodriguez and Laio (2014) introduced a new method that is faster and simpler than mean shift clustering. Furthermore, they define a clustering diagram that provides a simple, two-dimensional summary of the clustering information. We study the statistical properties of this diagram and we propose some improvements and extensions. In particular, we show a connection between the diagram and robust linear regression.

MSC 2010 subject classifications: Primary 62H30; secondary 62H86.
Keywords and phrases: Modes, clustering, mean-shift.

Received May 2018.

1. Introduction

Mode-based clustering methods define clusters in terms of the modes of the density function. For example, the mean-shift clustering method (Comaniciu and Meer, 2002; Cheng, 1995) defines the clusters to be the basins of attraction of each mode. Specifically, if we take any point x and follow the path of steepest ascent of the density, then we end up at a mode. This assigns every point to a mode which forms a partition of the space. In practice, the density is estimated using a kernel density estimate. The mean shift algorithm then approximates the steepest ascent paths.

Rodriguez and Laio (2014) introduced a new approach to mode-based clustering that avoids iterative computation of the density estimator. Furthermore, they define a diagram — which we call the *mode clustering diagram* — that provides a useful summary of the clustering information. The diagram is simply

*The authors thanks the reviewers for providing many helpful suggestions.

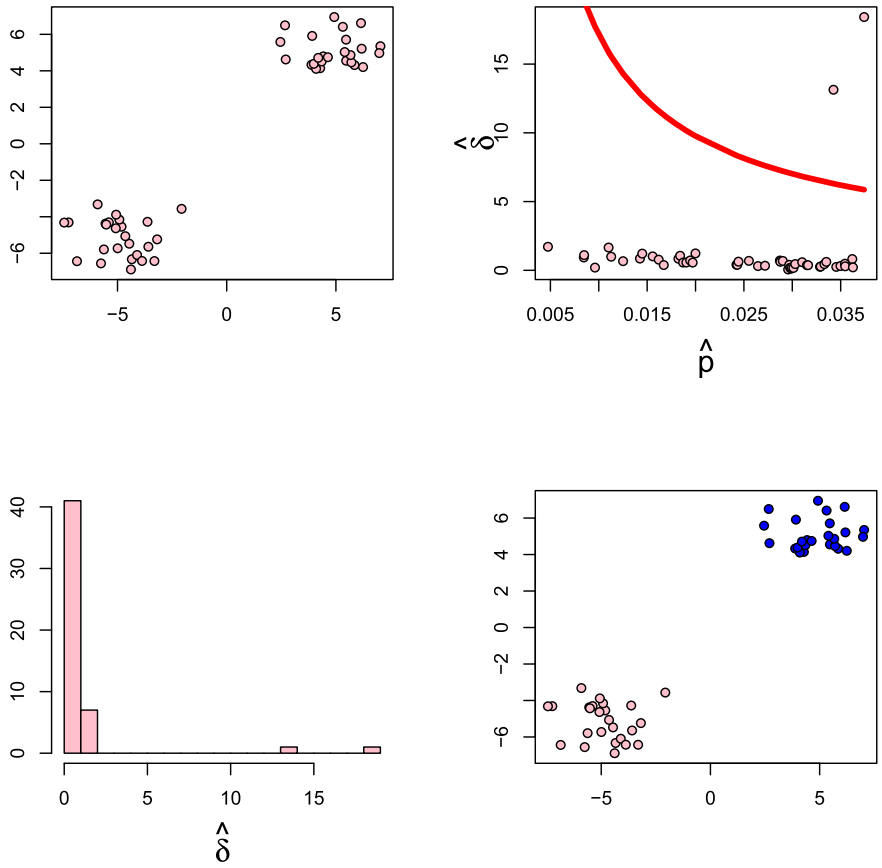

FIG 1. *Top left: data. Top right: the mode plot. The curved line is the threshold function* t_n *corresponding to a robust linear regression of* $\log \widehat{\delta}$ *on* $\log \widehat{p}$. *Points above the function are declared to be modes. Bottom left: Histogram of* $\widehat{\delta}(X_i)$. *Bottom right: the resulting clusters.*

a plot of the pairs $(p(X_i), \delta(X_i))$ where $p(X_i)$ is the density of the i^{th} point and $\delta(X_i)$ is the distance to the nearest neighbor with higher density. Modes tend to appear as isolated points in the top right of the diagram. See Figure 1 for a simple example.

In this paper, we study the properties of this diagram. These properties suggest a heuristic for deciding which points are modes. Specifically, if we perform a robust linear regression of $\log \delta(X_i)$ on $\log p(X_i)$ then modes correspond to large, positive outliers.

1.1. Related work

The most common mode-based clustering method is mean-shift clustering, developed by Cheng (1995) and Comaniciu and Meer (2002). The method has

been developed in the statistics literature by Li et al. (2007); Arias-Castro et al. (2015); Chacón et al. (2015, 2013); Chacón (2012) and Genovese et al. (2016).

The new method — the subject of this paper — is due to Rodriguez and Laio (2014). Extensions, including speedups and methods for dealing with higher dimensional problems include Wang and Xu (2017); Du et al. (2016); Courjault-Radé et al. (2016).

1.2. Paper outline

In Section 2 we establish the notation and the assumptions. We review mode-based clustering in Section 3. We establish the theoretical properties of the population version of the mode diagram in Section 4. We then consider the estimated diagram in Section 5. Based on these results, we suggest a method for thresholding the diagram in Section 6. In Section 7 we illustrate the method with several examples. Section 8 contains some concluding remarks. All proofs are in the appendix.

2. Notation and assumptions

Let $X_1, \ldots, X_n$ be a sample from a distribution P on $\mathbb{R}^d$. We make the following assumptions throughout the paper:

(A1) P is supported on a compact set $\mathcal{C}$ and has bounded, continuous density p. Also, $\inf_{x\in\mathcal{C}} p(x) \geq a > 0$.

(A2) p has bounded and continuous first, second and third derivatives. We let g denote the gradient and we let H denote the Hessian.

Recall that x is a critical point if $||g(x)|| = 0$. A function is *Morse* (Milnor, 2016) if the Hessian is non-degenerate at every critical point.

(A3) p is Morse with finitely many critical points.

The Morse assumption is critical to our proofs. It may be possible to drop this assumption but the proof techniques would have to change considerably. The assumption that $\inf_{x\in\mathcal{X}} p(x) \geq a > 0$ is not critical and could be dropped at the expense of more involved statements and proofs.[1]

A point x is a mode if there exists an $\epsilon > 0$ and a ball $B(x, \epsilon)$ such that $p(x) > p(y)$ for all $y \in B(x, \epsilon)$, $y \neq x$. Let $\mathcal{M} = \{m_1, \ldots, m_k\}$ denote the modes. Because p is Morse, x is a mode if and only if $g(x) = (0, \ldots, 0)^T$ and $\lambda_{\max}(H(x)) < 0$ where $\lambda_{\max}(A)$ denotes the largest eigenvalue of a matrix A.

3. Density mode clustering

In this section, we review mode-based clustering beginning with mean-shift clustering and then we moving on to the approach in Rodriguez and Laio (2014).

[1] More specifically, the proofs require dividing the sample space into two regions: the first where $p(x) \geq n^{-\frac{1}{d+2}}$ and the second where $p(x) < n^{-\frac{1}{d+2}}$. Also, points with $\widehat{p}(x) < n^{-\frac{1}{d+2}}$ should be removed from the cluster diagram.

3.1. Mean-shift clustering

The most common mode-based clustering method is mean-shift clustering (Chacón et al., 2015, 2013; Chacón, 2012; Li et al., 2007; Comaniciu and Meer, 2002; Arias-Castro et al., 2015; Cheng, 1995; Genovese et al., 2016). The idea is to find modes of the density and then define clusters as the basins of attraction of the modes.

Let x be an arbitrary point. If we follow the steepest gradient ascent path starting at x, we will eventually end up at one of the modes. More precisely, the gradient ascent path (or integral curve) starting at x is the function $\pi_x : \mathbb{R} \to \mathbb{R}^d$ defined by the differential equation

$$\pi_x'(t) = \nabla p(\pi_x(t)), \quad \pi_x(0) = x. \tag{1}$$

The *destination* of x is defined by

$$\text{dest}(x) = \lim_{t\to\infty} \pi_x(t). \tag{2}$$

It can be shown that, for almost all x, $\text{dest}(x) \in \mathcal{M}$. (The exceptions, which have measure 0, lead to saddle points.) The path π_x defines the gradient flow from a point x to its corresponding mode.

The *basin of attraction* of the mode m_j is the set

$$\mathcal{C}_j = \Big\{x : \ \text{dest}(x) = m_j\Big\}. \tag{3}$$

In the mean-shift approach to clustering, the population clusters are defined to be the basins of attraction $C_1, \ldots, C_k$. The left plot in Figure 2 shows a bivariate density with four modes. The right plot shows the partition (basins of attractions) induced by the modes.

To estimate the clusters, we find the modes $\widehat{\mathcal{M}} = \{\widehat{m}_1, \ldots, \widehat{m}_r\}$ of a density estimate $\widehat{p}$. A simple iterative algorithm called the *mean shift algorithm* (Cheng, 1995; Comaniciu and Meer, 2002) can be used to find the modes and to find the destination of any point x when $\widehat{p}$ is the kernel density estimator:

$$\widehat{p}(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h^d} K\left(\frac{x - X_i}{h}\right) \tag{4}$$

with kernel K and bandwidth h. For any given x, we define the iteration, $x^{(0)} \equiv x$,

$$x^{(j+1)} = \frac{\sum_i X_i K\left(\frac{||x^{(j)} - X_i||}{h}\right)}{\sum_i K\left(\frac{||x^{(j)} - X_i||}{h}\right)}.$$

See Figure 3. It can be shown that this algorithm is an adaptive gradient ascent method, that approximates the gradient flow defined by (1). The convergence of this algorithm is studied in Arias-Castro et al. (2015).

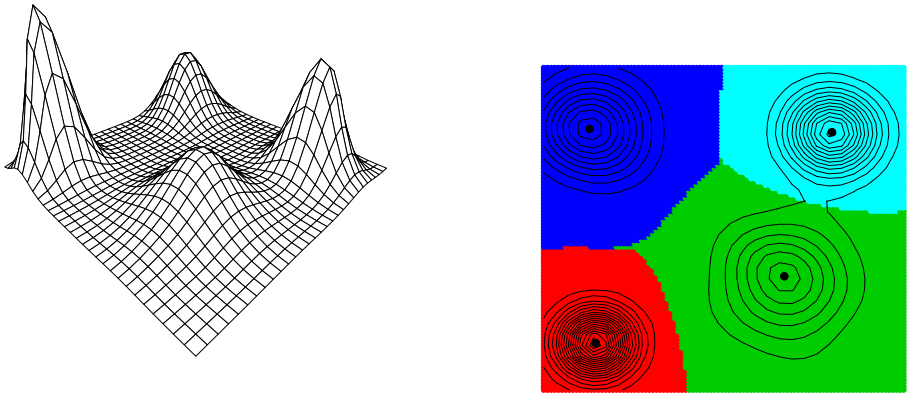

FIG 2. *Left: a density with four modes. Right: the partition (basins of attraction) of the space induced by the modes. These are the population clusters.*

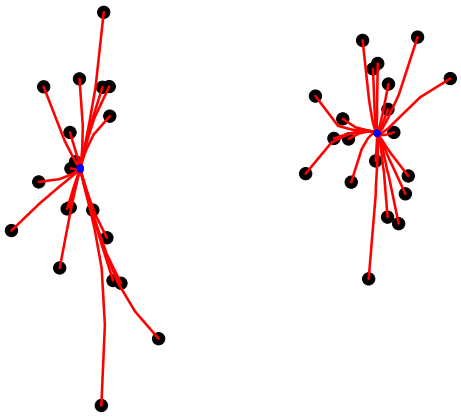

FIG 3. *The mean shift algorithm. The data are represented by the black dots. The modes of the density estimate are the two blue dots. The red curves show the mean shift paths; each data point moves along its path towards a mode as we iterate the algorithm.*

3.2. The mode diagram

Following Rodriguez and Laio (2014), we define

$$\delta(X_i) = \min\Big\{||X_j - X_i|| : \ p(X_j) > p(X_i)\Big\}. \tag{5}$$

That is, $\delta(X_i)$ is the distance of X_i to the closest point with higher density. In the case where there are no points with $p(X_j) > p(X_i)$ (in other words, $p(X_i) = \max_j\{p(X_j) : \ j = 1, \ldots, n\}$) we define $\delta(X_i) = L$ where L is any, arbitrary large constant. The choice of L does not matter. In practice, Rodriguez and Laio (2014) suggest setting $L = \max_{i,j} ||X_j - X_i||$ which is the diameter of

the dataset. In our examples, this is what we shall do. For developing theory, it will be convenient to just keep L as any arbitrary, large constant.

Next we form the *mode plot*, where we plot the pairs $(p(X_i), \delta(X_i))$. The intuition is that $\delta(X_i)$ will be small for most points. But if X_i is close to a local mode, then $\delta(X_i)$ will be large since the nearest point with a higher density will be at another mode. Hence, modes will show up as isolated points in the top right of the mode plot. A simple example is shown in Figure 1. Formally, the mode plot is the collection of pairs

$$\mathcal{D} = \Big\{(p(X_i),\ \delta(X_i)) :\ i = 1, \ldots, n\Big\}. \tag{6}$$

The modes can be identified by inspection of the diagram. In this paper, we suggest a method to separate modes from non-modes using linear regression.

In practice, we need to estimate p. We will use the kernel density estimator defined in (4). Then we define $\widehat{\delta}(X_i) = \min\{||X_j - X_i|| :\ \widehat{p}(X_j) > \widehat{p}(X_i)\}$. We have to decide which points on the diagram correspond to modes. For this purpose, let $t_n : \mathbb{R} \to \mathbb{R}$ be a given function. The points X_i such that

$$\widehat{\delta}(X_i) > t_n(U_i)$$

are the estimated modes, where $U_i = \widehat{p}(X_i)$. Denote these points by $\widehat{\mathcal{M}} = \{\widehat{m}_1, \ldots, \widehat{m}_\ell\}$. We call t_n the *threshold function*.

The main contribution of this paper is to study the properties of the mode diagram. We shall see that the mean of $\log \delta(X_i)$ as a function of $\log p(X_i)$ is approximately linear, for non-modes. But when X_i is close to a mode, $\log \delta(X_i)$ lies far above the line. This suggests the following method for separating modes from non-modes. We show, theoretically, that that it suffices to use a threshold function of the form $t_n(u) = (C \log n/(nu))^{1/d} = C_0 u^{-1/d}$. In practice, we use the more flexible form $t_n(u) = C_0 u^{\beta_1}$ where C_0 and β_1 are esti9mated as follows. We perform a robust linear regression of $\log \widehat{\delta}(X_i)$ on $\log \widehat{p}(X_i)$. That is, we find β_0 and β_1 such that

$$\log \widehat{\delta}(X_i) \approx \beta_0 + \beta_1\ \log \widehat{p}(X_i).$$

Then we look for large positive outliers. These are points for which $\log \widehat{\delta}(X_i) > \beta_0 + \beta_1 \log \widehat{p}(X_i) + Ms$ where s is the estimated residual standard deviation and M is some large constant; we use $M = 3$ for the examples in this paper. These points correspond to modes. This corresponds to taking the threshold function

$$t_n(u) = e^{\beta_0 + Ms} u^{\beta_1}. \tag{7}$$

Thus, X_i is declated to be a mode if $\widehat{\delta}(X_i) > t_n(U_i)$ where $U_i = \widehat{p}(X_i)$ and $t_n(u) = e^{\beta_0 + Ms} u^{\beta_1}$. The reason for this choice of threshold function arises from the theory in Section 4.

To assign points to modes, Rodriguez and Laio (2014) suggest an approach that avoids the iterations of the mean-shift method. Instead, we assign each point to nearest neighbor with higher density. This leads each sample point to a mode without having to recompute the density estimator at any other points.

Input: Data $\{X_1, \ldots, X_n\}$, threshold function t_n.

1. Compute density estimator $\widehat{p}(X_i)$ at each point.
2. Define $\widehat{\delta}(X_i) = \min\{||X_j - X_i|| : \widehat{p}(X_j) > \widehat{p}(X_i)\}$. Take $\widehat{\delta}(X_i) = \max_{i,j} ||X_i - X_j||$ if $\widehat{p}(X_i) > \widehat{p}(X_j)$ for all j.
3. Let $\widehat{\mathcal{M}} = \{X_i : \widehat{\delta}(X_i) > t_n(\widehat{p}(X_i))\}$.
4. Cluster assigment: for each X_i, move to the closest point with higher density. Continue until a mode is reached.
5. Let $\overline{C}_j$ be all points assigned to $m_j \in \widehat{\mathcal{M}}$.

Return: $\overline{C}_1, \ldots, \overline{C}_\ell$.

FIG 4. *The Rodriguez-Laio Algorithm.*

This is essentially a sample-based approximation to the gradient. The steps of the algorithm are summarized in figure 4.

This method has several advantages over mean-shift clustering. We never need to estimate or approximate the gradient of the density. There is no need for any iterative calculation of the density. This makes the method fast. However, our focus is not on the algorithm but on the mode diagram which gives a nice, two-dimensional summary of the clustering information.

4. The oracle diagram

In this section we assume that the density p is known. We then call $\mathcal{D} = \{(p(X_i), \delta(X_i)) : \ i = 1, \ldots, n\}$ the *oracle diagram*. Note that $\mathcal{D}$ is a point process on $\mathbb{R}^2$. The variables $\delta(X_i)$ are not independent since $\delta(X_i)$ depends on the configuration of the other points.

We need the following definition from Cuevas et al (1990). A set S is (γ, τ)-standard if there exist $\epsilon_0 > 0$ and $\tau \in (0, 1)$ such that: for all $0 < \epsilon \leq \epsilon_0$ and all $x \in S$,

$$\mu\Big(B(x, \epsilon) \bigcap S\Big) \geq \tau\, \mu(B(x, \epsilon)). \tag{8}$$

A set that is standard does not have sharp protrusions. Our proofs require the assumption that the level sets $\{p > t\}$ are standard. However, this requires some care. Suppose that $x = m_j$ where m_j is a mode of p. Let $S = \{y : \ p(y) \geq p(x)\}$. Then $S \bigcap B(x, \epsilon) = \{x\}$ and so $\mu(S \bigcap B(x, \epsilon)) = 0$ and standardness thus fails for points that are modes. More generally, we cannot lower bound $\mu(\{p(y) \geq p(x)\} \bigcap B(x, \epsilon))$ unless x is at least ϵ far from the modes. We use the following restricted standardness assumption. Let $L_x = \{y : \ p(y) > p(x)\}$.

(A4) There exists $\epsilon_0 > 0$ and $\tau \in (0, 1)$ such that, whenever $\min_j ||x - m_j|| > t$ with $0 < t < \epsilon_0$, we have that

$$\mu\Big(L_x \bigcap B(x, t)\Big) \geq \tau \mu(B(x, t)). \tag{9}$$

In the rest of the section we assume that (A1)-(A4) hold.

Before proceeding, we need a bit more notation. Recall that $\mathcal{M} = \{m_1, \ldots, m_k\}$ is the set of true modes. Let $m_j \in \mathcal{M}$. Let $H(m_j)$ be the Hessian at m_j. Let $J(m_j) = -H(m_j)$ and λ_j be the smallest eigenvalue of $J(m_j)$. Note that

$\lambda_j > 0$. Since the Hessian is a continuous function, there exists $\omega_j > 0$ such that $\lambda_{\min}(J(x)) \geq \lambda_j/2$, for all $x \in B(m_j, \omega_j)$. Let $\Lambda_j = \sup_{x \in B(m_j,\omega)} \lambda_{\max}(J(x))$.

Define

$$\epsilon_n = \left(\frac{r \log n}{n}\right)^{\frac{1}{d}}, \qquad t_n(u) = \left(\frac{C \log n}{n\, u}\right)^{\frac{1}{d}} \tag{10}$$

where

$$C \geq G^d 2^{d/2} r \max_j p(m_j) \left(\frac{\Lambda_j}{\lambda_j}\right)^{d/2}, \quad G \geq \max\left\{\max_j \left[3\left(\frac{\lambda_j}{\Lambda_j}\right)^d \frac{1}{2^{d/2} v_d a \tau}\right]^{\frac{1}{d}}, 1\right\} \tag{11}$$

where v_d denotes the volume of the unit ball in $\mathbb{R}^d$ and $r > 1/(a v_d)$.

Remark. The constants — such as C, r, G and so on — are only used to state the theoretical results. The actual procedure described in Section 6 does not require these constants.

Because p is Morse, the modes are isolated points. It follows that there exists some $c > 0$ such that $B(m_s, c\omega_s) \bigcap B(m_t, c\omega_t) = \emptyset$ for all $1 \leq s < t \leq k$. Without loss of generality, we assume that $c = 1$. Hence, $B(m_s, \omega_s) \bigcap B(m_t, \omega_t) = \emptyset$ for $1 \leq s < t \leq k$.

We assume that (A1)-(A4) hold in the rest of the paper.

4.1. The mode diagram

We first need to define k sample points that can be considered to be sample modes. These are the points that will be in the upper right portion of the mode diagram. (For a unimodal density, this would just be the point X_i that maximizes $p(X_i)$.) We define the sample point $X_j \in B(m_j, \omega_j)$ to be a *sample mode* if $p(X_j) \geq p(X_i)$ for all $X_i \in B(m_j, \omega_j)$. We renumber the points so that $X_1, \ldots, X_k$ denote the k sample modes. Note that these points are not known since they depend on m_j and ω_j. But, as we shall see, we can identify them by using the mode diagram.

The next result shows that X_j is close to m_j and that $\delta(X_j)$ is bounded below by a constant.

Theorem 1. *Let*

$$\psi_{n,j}^2 = \frac{2\, G^2\, \Lambda_j}{\lambda_j}\, \epsilon_n^2 \tag{12}$$

where G was defined in (11). Also, let $\psi_n = \max_j \psi_{n,j}$. If X_j is a sample mode then:

(i) $X_j \in B(m_j, \psi_{n,j})$.

(ii) Any sample point X_i such that $p(X_i) > p(X_j)$ is far from X_j; specifically $\delta(X_i) \geq \omega_j/2$.

Now let $m_j, j = 1, \ldots, k$, denote the modes of $p(x)$, and let $X_j, j = 1, \ldots, k$, be the local modes. Define $\Gamma = \bigcup_{j=1}^{k} B(m_j, \psi_n)$ and divide the dataset into three groups:

$$\mathcal{X}_1 = \Big\{X_1, \ldots, X_k\Big\}, \quad \mathcal{X}_2 = \Big\{X_i : \ X_i \in \Gamma, X_i \notin \mathcal{X}_1\Big\}, \quad \mathcal{X}_3 = \Big\{X_i \in \Gamma^c\Big\}. \tag{13}$$

Note that $\mathcal{X}_1$ is precisely the set of sample modes. Theorem 2, below, shows that $\delta(X_j)$ is bounded away from 0 for the points in $\mathcal{X}_1$ (and hence they lie above the threshold function) while $\delta(X_i)$ lies below the threshold functions for all X_i in $\mathcal{X}_2$ and X_i in $\mathcal{X}_3$.

Remark. If the assumption that $p(x) \geq a > 0$ is dropped, then $\mathcal{X}_3$ needs to be re-defined as $\mathcal{X}_3 = \Big\{X_i \in \Gamma^c\Big\} \cap \Big\{X_i : p(X_i) \geq n^{-1/(d+2)}\Big\}$.

Theorem 2. *(i) Let $X_j \in \mathcal{X}_1$ be the sample mode in $B(m_j, \psi_{n,j})$. Then $p(X_j) = p(m_j) + O_P(\epsilon_n^2)$ and $\delta(X_j)/t_n(p(X_j)) \to \infty$.*

(ii) For all $X_i \in \mathcal{X}_2$, we have $\delta(X_i) \leq t_n(p(X_i))$.

(iii) $P^n\Big(\delta(X_i) \leq t_n(p(X_i))$ for all $X_i \in \mathcal{X}_3\Big) \to 1$.

4.2. The limiting distribution

To get more information about the shape of the mode diagram we show that for any x that is not a mode, the distribution of $n\delta^d(x)$ only depends on $p(x)$ and converges to an exponential random variable with mean $1/(p(x)\,\tau\,v_d)$ where v_d is the volume of the unit ball. This means that $\delta(x) \approx (n\,\tau\,v_d\,p(x))^{-1}E$ where $E \sim \text{Exp}(1)$ and that a plot of $\log \delta(X_i)$ versus $\log p(X_i)$ should look linear for all X_i's not close to a mode. On the other hand if x is a mode, then $n\delta^d(x) \to \infty$.

We will need the following stronger version of (A4). Recall that $L_x = \{y : p(y) > p(x)\}$.

(A4') There exists $\tau(x) \in (0, 1)$ such that, for any $x \notin \mathcal{M}$,

$$\lim_{t \to 0} \frac{\mu\Big(L_x \bigcap B(x, t)\Big)}{\mu(B(x, t))} = \tau(x). \tag{14}$$

Theorem 3. *Suppose that (A1), (A2), (A3) and (A4') hold and that x is not a mode. Then the random variable $n\ \delta^d(x)$ converges in distribution to an exponential random variable, with parameter $p(x)\tau(x)\ v_d$. If x is a mode, then $n\delta^d(x) \to \infty$.*

4.3. The linear heuristic

The results in the previous sections show that, for non-modes, $\log \mathbb{E}[n\,\delta^d(X)]$ should be approximately linear in $\log p(x)$. On the other hand, points closest to

modes will lie far above the thredhold. This suggests the following approach: plot $\log p(X_i)$ versus $\log \delta(X_i)$. Most points will fall below some line. A few points will be above the line. In Section 6, we will fit a robust linear regression to the log-mode plot. The outliers above the line will indicate the modes. We pursue this idea in Section 6.

5. The estimated mode diagram

Since p is not known, we have to estimate the diagram. Let $\widehat{p}$ denote the kernel density estimator and let

$$\widehat{\delta}(X_i) = \min\Big\{||X_j - X_i|| : \ \widehat{p}(X_j) > \widehat{p}(X_i)\Big\}. \tag{15}$$

As before, if there are no points with $\widehat{p}(X_j) > \widehat{p}(X_i)$ we define $\delta(X_i) = L$ where L is any positive constant. The estimated diagram is

$$\widehat{\mathcal{D}} = \Big\{(\widehat{p}(X_i), \widehat{\delta}(X_i)) : \ i = 1, \ldots, n\Big\}. \tag{16}$$

Remark. An alternative approach to defining the estimated diagram is as follows. We draw a sample $X_1^*, \ldots, X_N^*$ from $\widehat{p}$. We then define

$$\widehat{\delta}(X_i^*) = \min\Big\{||X_j^* - X_i^*|| : \ \widehat{p}(X_j^*) > \widehat{p}(X_i^*)\Big\} \tag{17}$$

and

$$\widehat{\mathcal{D}}^* = \Big\{(\widehat{p}(X_i^*), \widehat{\delta}(X_i^*)) : \ i = 1, \ldots, N\Big\}. \tag{18}$$

This approach has the advantage that we can take N to be much larger than n. This gives a more accurate summary $\widehat{p}$. On the other hand, if n is huge, we might even take N smaller than n to reduce computation. At any rate, by sampling from $\widehat{p}$ we have more control.

The rest of the section is devoted to showing that $\widehat{\mathcal{D}}$ has the same behavior as the oracle diagram in Theorem 2. First, we recall some facts about $\widehat{p}$. Let $\mathcal{M} = \{m_1, \ldots, m_k\}$ denote the modes of p and let $\mathcal{C} = \{c_1, \ldots, c_r\}$ denote the remaining critical points of $\widehat{p}$. Let $\widehat{g}$ be the gradient of $\widehat{p}$ and let $\widehat{H}$ be the Hessian. Then, with high probability, for all large n, $\widehat{p}$ is Morse the the same number of critical points as p. This is summarized in the next result.

Lemma 4. *Assume (A1)-(A3). Take the bandwidth to be* $h_n \asymp n^{-1/(d+6)}$. *Let*

$$r_n = a_1(\log n/n)^{2/(4+d)}, \quad s_n = a_2(1/n)^{2/(d+6)} \tag{19}$$

where a_1 *and* a_2 *are positive constants. There exists a sequence of events* $\mathcal{A}_n$ *such that* $P^n(\mathcal{A}_n) \to 1$ *and such that, on* $\mathcal{A}_n$*:*
(i) $\sup_{x \in \mathcal{X}} ||\widehat{p}(x) - p(x)|| \leq r_n$.
(ii) $\widehat{p}$ *is Morse and has exactly* k *modes* $\widehat{m}_1, \ldots, \widehat{m}_k$ *with* $\max_j ||\widehat{m}_j - m_j|| \leq s_n$.
(iii) The remaining critical points $\widehat{\mathcal{C}} = \{\widehat{c}_1, \ldots, \widehat{c}_r\}$ *of* $\widehat{p}$ *have the same cardinality*

as the number of critical points of p *and also satisfy* $\max_j ||\widehat{c}_j - c_j|| \leq s_n$.
(iv) $\sup_x ||\widehat{g}(x) - g(x)||_\infty = o_P(1)$ *and* $\sup_x \max_{j,k} ||\widehat{H}_{jk}(x) - H_{jk}(x)|| = o_P(1)$, *and the supremum of the third derivative is* $o_P(1)$.

For proofs of these facts, see Genovese et al. (2016) and Chazal et al. (2017). In what follows, we assume that the event $\mathcal{A}_n$ holds. In particular, $\widehat{p}$ is Morse. In what follows, we refer to positive constants c_1, c_2 which come from Lemma 8.

As in the previous section, we may find constants ω_j such that the balls $B_j = B(\widehat{m}_j, \omega_j)$ are disjoint and each contains at least one data point. For $j = 1, \ldots, k$ let $X_j = \text{argmin}_{X_i \in B_j} \widehat{p}(X_i)$. Let $\mathcal{X} = \{X_1, \ldots, X_n\}$. Define

$$\mathcal{X}_1 = \{X_1, \ldots, X_k\}, \;\; \mathcal{X}_2 = \left\{ X_i : \; X_i \in \bigcup_{j=1}^{k} B(\widehat{m}_j, c_1 \epsilon_n) \right\}, \;\; \mathcal{X}_3 = \mathcal{X} - (\mathcal{X}_1 \bigcup \mathcal{X}_2).$$

As before define $t_n(\widehat{p}(x)) = (C \log n/(n\widehat{p}(x)))^{1/d}$. In what follows, we sometimes write $t_n(x)$ as short for $t_n(\widehat{p}(x))$. The behavior of the diagram in this case is essentially the same as the oracle diagram as the next result shows. The proof is much more complicated since $\widehat{p}$ is a random function and is obviously correlated with the data.

Theorem 5. *Let* t_n *be defined as above. Then:*
(i) There exists $c > 0$ *such that, with probabality tending one,* $\widehat{\delta}(X_i) \geq c$ *for all* $X_i \in \mathcal{X}_1$. *Hence,* $\widehat{\delta}(X_i)/t_n(\widehat{p}(X_i)) \to \infty$.
(ii) $\widehat{\delta}(X_i) \leq c_2 t_n(\widehat{p}(X_i))$ *for all* $X_i \in \mathcal{X}_2$.
(iii) For $\mathcal{X}_3$ *we have that*

$$P^n\Big(\widehat{\delta}(X_i) \leq c_2 t_n(\widehat{p}(X_i)) \text{ for all } X_i \in \mathcal{X}_3\Big) \to 1.$$

6. Choosing the threshold using robust regression

We now know that both $\mathcal{D}$ and $\widehat{\mathcal{D}}$ have the following behavior. There are k points $X_1, \ldots, X_k$, corresponding to the k modes, such that $\widehat{p}(X_j)$ and $\widehat{\delta}(X_j)$ are large. For the remaining points, $\widehat{\delta}(X_i)$ is small. Specifically, $\widehat{\delta}(X_i) < t_n(\widehat{p}(X_i)) = (C \log n/(n\widehat{p}(X_i)))^{1/d}$. In other words, for non-modes, the points $(\log \widehat{\delta}(X_i), \log \widehat{p}(X_i))$ should fall on or below a line.

The modes could be selected visually by examining the log-log mode plot. Alternatively, if we perform a robust linear regression of $\log \widehat{\delta}(X_i)$ on $\log \widehat{p}(X_i)$, we expect the modes to show up as outliers. Let β_0 and β_1 be the estimated intercept and slope from the regression. Thus,

$$\log \widehat{\delta}(X_i) \approx \beta_0 + \beta_1 \log p(X_i).$$

Now we look for large positive outliers. These are points for which $\log \delta(X_i) > \beta_0 + \beta_1 \log p(X_i) + Ms$ where s is the estimated residual standard deviation and

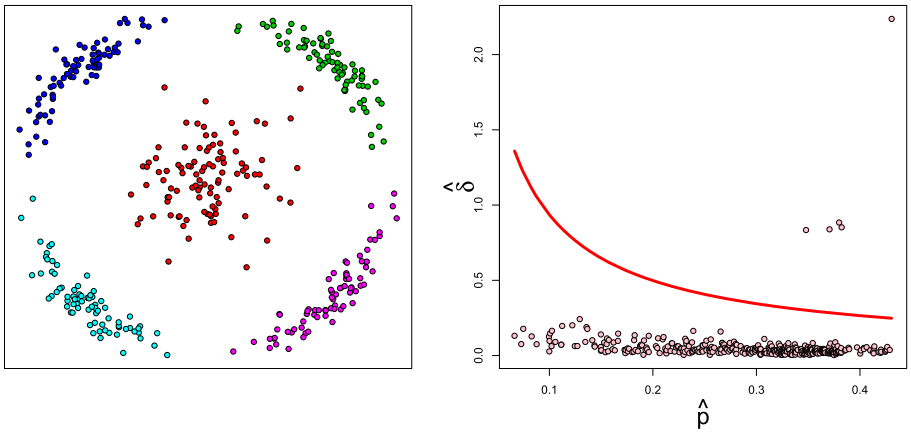

FIG 5. *Broken Circle data. Left: Data set. Right: Estimated Mode diagram* $\widehat{\mathcal{D}}$*, with threshold function.*

M is some large constant; we use $M = 5$ in the examples in this paper. These points are the estimated modes. This corresponds to taking

$$t_n(u) = e^{\beta_0 + Ms} u^{\beta_1}. \tag{20}$$

Remark. It is possible to define some post-processing diagnostics to make sure that the claimed modes are, in fact, modes. For example, if X_j is declared a mode, are $\mathcal{N}$ is a set of neighbors of X_j, then we can verify that $\widehat{p}(X_j) - \widehat{p}(X_i)$ $X_i \in \mathcal{N}$.

7. Examples

A first example illustrating the theory in the paper was presented in Figure 1. That picture shows a simple two-dimensional data set, with two well separated clusters. In Figure 1 the threshold function $t_n(\widehat{p}(X_i))$ was obtained as described in Section 6. A histogram of $\widehat{\delta}(X_i)$ was also added for completeness.

The examples of this section consist of data with four or five clusters in two dimensions, and two three-dimensional data-set with four clusters. The data sets can be enriched with added random noise. A two-dimensional data set is Figure 5. The left panel shows five separate clusters, with shapes that are parts of a *broken circle*. The data consist of 500 points. The density p is estimated with the kernel function $\widehat{p}$ in (4).

The estimated mode plot $\widehat{\mathcal{D}}$ from (15) and (16), in the right panel of Figure 5, displays values of $\widehat{p}(X_i)$ and $\widehat{\delta}(X_i)$ for each data point X_i with the threshold function $t_n(\widehat{p}(X))$, obtained from the robust regression line. The diagram shows five outliers at the top right corner, above $t_n(\widehat{p}(X))$. They correspond to points with large values of $\widehat{p}$ and $\widehat{\delta}$ that estimate the modes of the density. The robust regression line for $\log(\widehat{p})$ versus $\log(\widehat{\delta})$ is in the right panel of Figure 6.

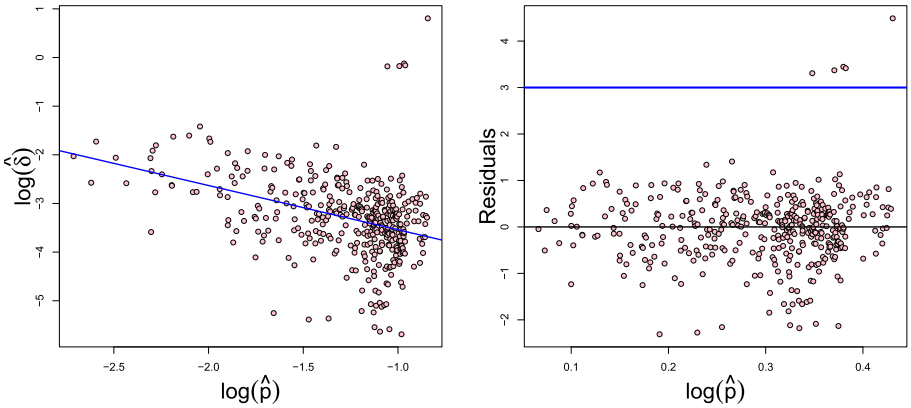

FIG 6. *Broken Circle data. Left: Robust regression line for* $log(\widehat{p})$ *vs.* $\log(\widehat{\delta})$. *Right: Residuals from robust regression.*

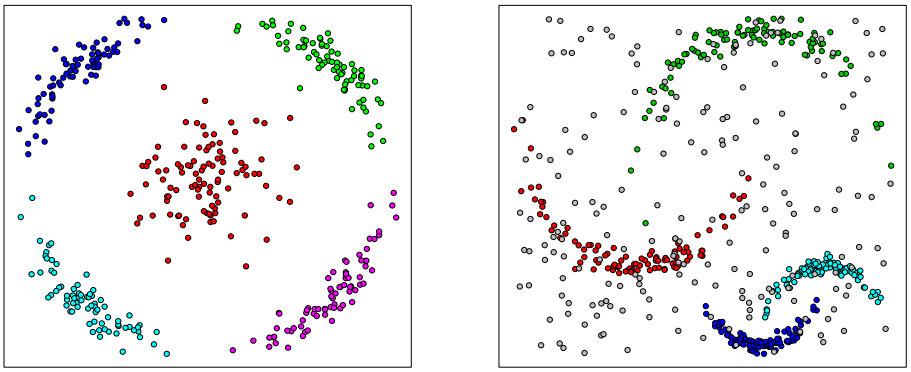

FIG 7. *Left: Broken Circle data, estimated clusters. Right: New dataset. Four crescent clusters, with added uniform noise.*

Residuals from the robust regression are in the left panel of Figure 6. The five outliers are even more noticeable, there. Finally, the estimated clusters for the *Broken circle* data are in the left panel of Figure 7.

The dataset in the right panel of Figure 7, consists of 400 points clustered in the shape of four crescent, augmented with 200 points of uniform noise. Despite the added noise, the mode diagram $\widehat{\mathcal{D}}$, in the left panel of Figure 8, correctly identified four modes. The right panel in the same Figure shows the identified clusters. Part of the random noise has been, correctly, assigned to some of the four main clusters.

The left panel in Figure 9 contains an example of a three dimensional data set, consisting of 400 data points, forming four clusters, and 400 points of uniform noise. The procedure is unchanged when data are in more than two dimensions.

This is clearly shown in the right panel of Figure 9 where the mode diagram

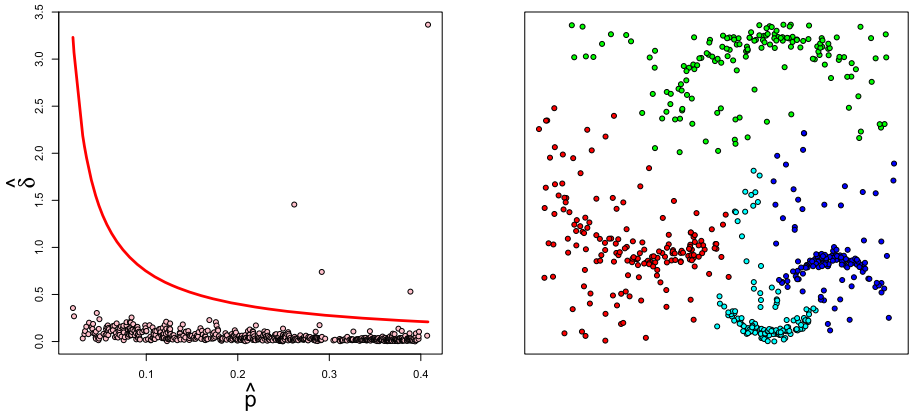

FIG 8. *Four crescent clusters. Left: Estimated Mode diagram* $\widehat{\mathcal{D}}$ *with threshold function. Right: Estimated Clusters.*

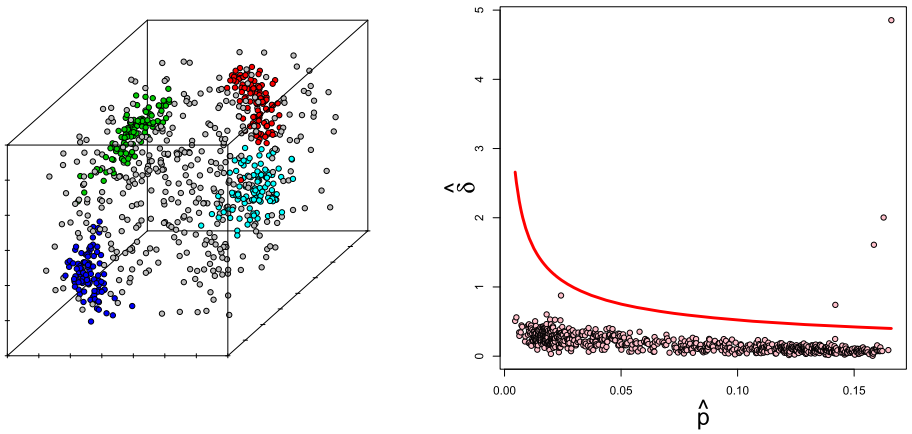

FIG 9. *Three dimensional data set. Left: The data. Right: Mode diagram* $\widehat{\mathcal{D}}$ *with threshold function.*

$\widehat{\mathcal{D}}$ identifies four points above the threshold function. The left panel in Figure 10 shows the four estimated clusters where, as before, points of random noise are assigned to the main clusters, according to their closeness to the modes.

The mode plot above includes one point with small values of both $\widehat{\delta}$ and $\widehat{p}$, just below the threshold function. In a different run of the code, points in the lower left section of the mode diagram might be slightly above the threshold function.

This would signal the existence of an extra cluster. Data for such an example are in the right panel of Figure 10. This new data-set presents, in the left panel of Figure 11, a mode plot where one point in the lower left section of the diagram is slightly above the threshold function.

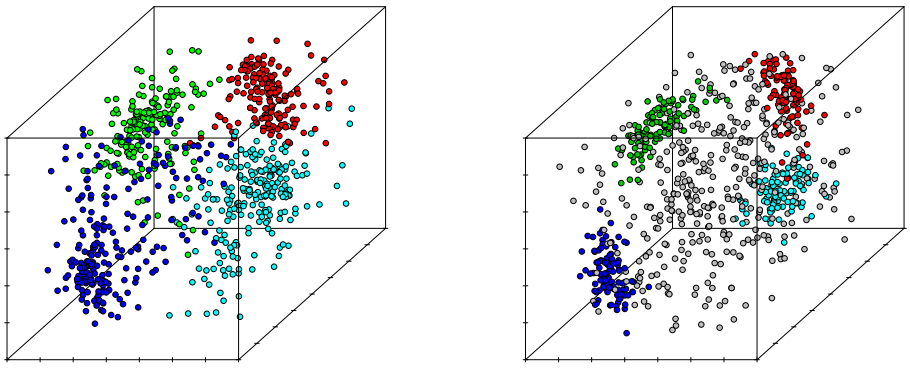

FIG 10. *Three dimensional noisy data set. Left: Estimated clusters* $\widehat{\mathcal{D}}$ *with threshold function. Right: Three dimensional data for an alternative code run.*

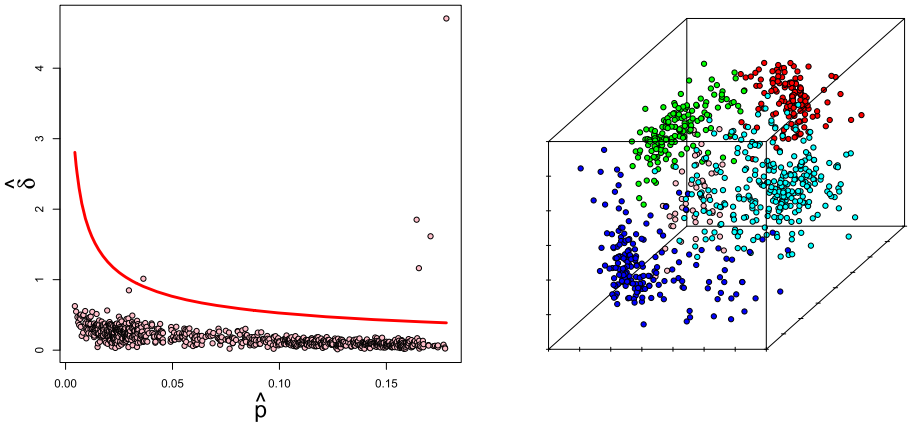

FIG 11. *New set of Three Dimensional Noisy data. Left: Estimated Mode diagram* $\widehat{\mathcal{D}}$*, and threshold function, showing five modes. Right: Five estimated clusters.*

But for a sample point to be a mode, its values of both $\widehat{\delta}$ and $\widehat{p}$ need to be large. Thus, these type of points are not relevant for the final result This can be seen in the right panel of Figure 11 where the fifth cluster detected from $\widehat{\mathcal{D}}$, consists of just a few points, in pink, that do not affect the qualitative result of clustering. This observation will be useful for inspecting mode diagrams when clustering data sets in more than three dimensions.

We conclude this section by mentioning the differences between the method of Rodriguez and Laio (2014) versus the more traditional mean shift algorithm. The two main differences are: (i) the new method only requires computing the estimated density once, at the beginning of the algorithm, while the mean shift requires recomputing it at each iteration and (ii) the new method provides a mode clustering diagram. Also, the new algorithm is much faster.

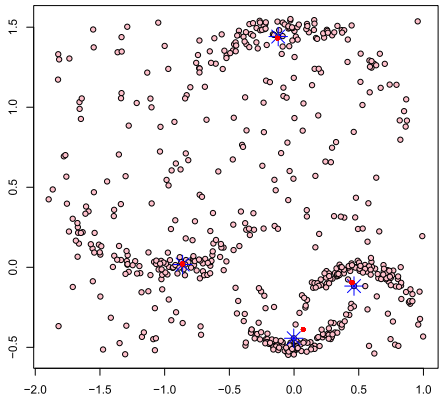

FIG 12. *The blue stars show the modes found using the method of Rodriguez and Laio (2014). The red dots show the modes found using the mean shift algorithm. The estimates are very close but the mean shift algorithm is much slower.*

As an example, Figure 12 shows the dataset of four crescent, with the addition of 200 points of uniform noise. In this case, we find that the new algorithm is typically about 30 to 40 times faster. Of course, they will not find exactly the same modes since the new method must estimate the modes to be one of the sample points. But typically, they are very close. In Figure 12, the blue stars show the modes found from Rodriguez and Laio (2014) algorithm, and the red dots show the result of 200 interations of the mean shift algorithm. The blue stars can be masked by the red dots, when the mean shift algorithm identifies one of the sample points to be a mode. Although the mode estimates are not identical they are vety similar. In this paper, we have mainly focused on point (ii). The mode clustering diagram provides a simple visualization tool for mode based clustering. It should be pointed out that these two approaches can be combined. We could find the modes with mean shift clustering if we are willing to do the extra computation, and then display the mode clustering diagram. This would be more expensive but it would eliminate the need to create a thresholding rule since we would know what the modes are. In summary, the new approach is much faster and provides a useful visual summary of the clustering while the mean shift clustering provides a more exact estimate of the modes.

8. Conclusion

We have studied the properties of the mode diagram introduced by Rodriguez and Laio (2014). We have seen that, for non-modes, $\log \widehat{\delta}(X_i)$ falls on or below a linear function of $\log \widehat{p}(X_i)$. Based on this observation, we suggested a robust regression method for classifying points on the mode diagram as modes or non-modes. We would like to emphasize that we think that the mode plot is a useful visualization method for mode-based clustering regardless of how one separates modes from non-modes.

Our analysis depended on a number of assumptions. In particular, we assumed that the density p is Morse. This assumption is made — explicitly or implicitly — in most density based clustering methods. Loosely speaking, this means that p has no "flat regions." (A notable exception is Jiang and Kpotufe (2016) who specifically allow for such flat region.) We conjecture that the mode plot still provides useful information when p is not Morse but proving this will require new tools.

It would be interesting to develop a similar diagnostic plot for other clustering methods such as k-means clustering. Currently, we are not aware of any such diagnostic plots.

Appendix: Proofs

Proof of Theorem 1. (i) Let $X_{(j)}$ be the closest point to m_j. From Lemma 6 below, $||X_{(j)} - m_j|| \leq \epsilon_n$. For any point $x \in B(m_j, \omega_j)$ we have

$$p(x) = p(m_j) - \frac{1}{2}\,(x - m_j)^T \left[\int_0^1 J(u\,m_j + (1-u\,)x)\; du \right] (x - m_j).$$

Thus for X_j

$$\begin{aligned} p(X_j) &= p(\,m_j) - \frac{1}{2}(X_j - m_j)^T \left[\int_0^1 J(u, m_j + (1-u)\,X_j)\,du\right](X_j - m_j) \\ &\leq p(m_j) - \frac{\lambda_j}{4}||X_j - m_j||^2 \end{aligned} \tag{21}$$

and for $X_{(j)}$

$$\begin{aligned} p(X_{(j)}) &= p(\,m_j) - \frac{1}{2}\,(X_{(j)} - m_j)^T \left[\int_0^1 J(u\,m_j + (1-u)\,X_{(j)})du\right](X_{(j)} - m_j) \\ &\geq\; p(m_j) - \frac{\Lambda_j}{2}\;||X_{(j)} - m_j||^2. \end{aligned}$$

Then

$$p(\,m_j) - \frac{\lambda_j}{4}\;||X_j - m_j||^2 \geq p(X_j) \geq p(X_{(j)}) \geq p(\,m_j) - \frac{\Lambda_j}{2}\;||X_{(j)} - m_j||^2$$

which implies that

$$||X_j - m_j||^2 \leq \frac{2\Lambda_j}{\lambda_j}\;||X_{(j)} - m_j||^2 \leq \frac{2\Lambda_j}{\lambda_j}\;\epsilon_n^2 \leq \psi_{n,j}^2. \tag{22}$$

This proves *(i)*.

(ii) Let X_i be any point with $p(X_i) > p(X_j)$. By definition, $X_i \notin B(m_j, \omega_j)$. By the triangle inequality,

$$\omega_j \le ||X_i - m_j|| \le ||X_i - X_j|| + ||X_j - m_j|| \le ||X_i - X_j|| + \sqrt{\frac{2\Lambda_j}{\lambda_j}}\,\epsilon_n$$

Thus, since $\epsilon_n \to 0$

$$||X_i - X_j|| \ge \omega_j - \sqrt{\frac{2\Lambda_j}{\lambda_j}}\,\epsilon_n \ge \omega_j/2,$$

and $\delta(X_j) \ge \omega_j/2$. □

Lemma 6. *Let $X_{(j)}$ be the closest point to m_j for $j = 1, \ldots, k$. Then*

$$P^n\Big(\max_{1\le j\le k} ||X_{(j)} - m_j|| > \epsilon_n\Big) \to 0. \tag{23}$$

Hence, with probability tending to 1, each ball $B(m_j, \epsilon_n)$ contains at least one point.

Proof. Let v_d be the volume of the unit ball, let $B = B(m_j, \epsilon_n)$. For a sequence of points $y_{n,j} \in B$ converging to m_j as $n \to \infty$

$$P(B) = p(y_{n,j})\mu(B) = p(y_{n,j})\epsilon_n^d\, v_d$$

Since $y_{n,j} \to m_j$ as $n \to \infty$, when n is large $p(y_{n,j}) > p(m_j)/2$, so

$$\begin{aligned} P^n(||X_{(j)} - m_j|| > \epsilon_n) &= P^n(||X_i - m_j|| > \epsilon_n \text{ for all } i) \\ &= [1 - P(B)]^n \le [1 - p(y_{n,j})\epsilon_n^d\, v_d]^n \\ &\le \exp\left\{-np(y_{n,j})\epsilon_n^d\, v_d\right\} \\ &= \left(\frac{1}{n}\right)^{v_d\, p(y_{n,j})} \le \left(\frac{1}{n}\right)^{v_d\, p(m_j)/2}. \end{aligned}$$

Hence,

$$P^n\Big(\max_{1\le j\le k} ||X_{(j)} - m_j|| > \epsilon_n\Big) \le \sum_{j=1}^{k} \left(\frac{1}{n}\right)^{v_d\, p(m_j)/2} \to 0.$$ □

Proof of Theorem 2. Let $t_n(u)$ be defined in (10).

(i) Let $X_j \in \mathcal{X}_1$ be a local mode. From (21) and (22) it is immediate that $p(X_j) = p(m_j) + O_P(\psi_n^2)$. From Theorem 1, $\delta(X_j) \ge \omega_j/2 > 0$, then from the definition of $t_n(p(X_j))$ we have $\frac{\delta(X_j)}{t_n(p(X_j))} \to \infty$.

(ii) Let $X_i \in \mathcal{X}_2$. Since $X_i \notin \mathcal{X}_1$ it is not a local mode. So there exists a local mode $X_i \in B(m_j, \psi_{n,j})$ with $p(X_i) < p(X_j) \le p(m_j)$. Hence, $\delta(X_i) \le \psi_{n,j}$. So, from (11),

$$\delta^2(X_i) \le \psi_{n,j}^2 = \frac{2G^2\Lambda_j\epsilon_n^2}{\lambda_j} \le \left(\frac{C\log n}{np(m_j)}\right)^{2/d} \le \left(\frac{C\log n}{np(X_i)}\right)^{2/d} = t_n^2(p(X_i)) \tag{24}$$

as required.

(iii) Let $x \in \Gamma^c$. Recall that $p(x) \geq a > 0$ for all x. From (24), we see that $t_n(x) \geq \psi_n = \max \psi_{n,j}$. So

$$\begin{aligned} P(\delta(x) > t_n(x)) &\leq P(\delta(x) > \psi_n(x)) = \prod_i P(X_i \notin B(x, \psi_n(x)) \bigcap L_x) \\ &= \left[1 - P(X_i \in B(x, \psi_n(x)) \bigcap L_x)\right]^n \\ &\leq \left[1 - a\mu(B(x, \psi_n(x)) \bigcap L_x)\right]^n \\ &\leq [1 - a\tau\mu(B(x, \psi_n(x))]^n \\ &\leq \exp\left(-na\tau v_d \psi_n^d\right) = \exp\left(-na\tau v_d 2^{d/2} G^d \left(\frac{\Lambda}{\lambda}\right)^d \frac{\log n}{n}\right) \\ &\leq \exp\left(-3\log n\right) = \left(\frac{1}{n}\right)^3 \end{aligned}$$

where we used the fact that $\mu(B(x, \psi_n(x)) \bigcap L_x) \geq \tau\mu(B(x, \psi_n(x)))$ due to (A4). So,

$$P\Big(\delta(X_i) > t_n(X_i) \text{ for some } X_i \in \Gamma^c\Big) \leq n\left(\frac{1}{n}\right)^3 \to 0. \qquad \square$$

Proof of Lemma 3. Fix $s > 0$ and let $B_n = B\left(x, \left(\frac{s}{n}\right)^{1/d}\right)$. Define $A_x = B_n \bigcap \Big\{y : \ p(y) > p(x)\Big\}$. There exists a sequence $y_n \to x$ such that $P(A_x) = p(y_n)\mu(A_x)$. From (A4'),

$$\begin{aligned} P(A_x) &= p(y_n)\mu(A_x) = p(y_n)\frac{\mu(A_x)}{\mu(B_n)}\mu(B_n) \\ &= (p(x) + o(1))(\tau(x) + o(1))\left(\frac{s}{n}\right) = \frac{sp(x)\tau(x)}{n} + o\left(\frac{1}{n}\right). \end{aligned}$$

Then

$$\begin{aligned} P\left(n\,\delta(x)^d \leq s\right) &= P\left[\delta(x) \leq (s/n)^{1/d}\right] = 1 - P\left[\delta(x) > (s/n)^{1/d}\right] \\ &= 1 - \prod_{i=1}^n P\left[X_i \notin A_x \text{ for all } X_i\right] \\ &= 1 - [P(X \notin A_x)]^n = 1 - [1 - P(X \in A_x)]^n \\ &= 1 - \exp\left\{n \log\left[1 - P(X \in A_x)\right]\right\} \\ &= 1 - e^{-s\tau(x)p(x)} e^{o(1)} \to 1 - e^{-s\tau(x)p(x)}. \end{aligned}$$

The final statement, about modes, follows since $\delta(x)$ is strictly positive. $\square$

Remark. If we had not assume that p is bounded from below, then one needs to work with the truncated region of the density $p(x) \geq a_n = n^{-1/(d+2)}$ and then $P^n(F_n^c) \leq N[1 - a\, v_d\, \epsilon_n^d]^n = \left[\frac{C_2}{\epsilon_n}\right]^d \left[1 - n^{-\frac{1}{d+2}}\, v_d \epsilon_n\right]^n \to 0$.

Proof of Theorem 5. The proofs of (i) and (ii) mimic the proof if Theorem 2, with $\widehat{p}$ replacing p. We focus on (iii).

In Lemma 7, we show that there exists balls $B(c_1, \epsilon_n), \ldots, B(c_N, \epsilon_n)$ such that the support of p is contained in $\bigcup_{s=1}^N B(c_j, \epsilon_n)$ and such that, $P^n(F_n) \to 1$ where F_n is the event that each ball contains at least one data point.

Let

$$\Gamma = \bigcup_{j=1}^k B(\widehat{m}_j, c_1 \epsilon_n).$$

In Lemma 8, we show that the following is true. For every $x \in \Gamma^c$, there exists a ball B such that (i) the radius of B is $2\epsilon_n$, (ii) $x \notin B$ but (iii) $\min_{z \in B} ||z - x|| \leq c_2 t_n(x)$ for some $c_2 > 0$ not depending on x. Since this holds for all $x \in \Gamma^c$ it also holds for all $X_i \in \mathcal{X}_3$. So there is a ball B_i such that (i) the radius of B_i is $2\epsilon_n$, (ii) $X_i \notin B_i$ but (iii) $\min_{z \in B} ||z - X_i|| \leq c_2 t_n(X_i)$. Now B must contain at least one of the covering balls $B(c_j, \epsilon_n)$. On F_n, this ball contains at least one point X_j, which is distinct from X_i. It follows that $\widehat{\delta}(X_i) \leq c_2 t_n(X_i)$. As this holds simultaneously for all $X_i \in \mathcal{X}_3$, the result follows. □

Lemma 7. *There exists a set $\mathcal{B} = \{B_1, \ldots, B_N\}$ where each B_j is a ball of radius ϵ_n, $N = [\xi/\epsilon_n]^d$ for some $\xi > 0$, $\mathcal{X} \subset \bigcup_j B_j$. Let F_n denote the event that each ball contains at least one data point. Then $P^n(F_n) \to 1$.*

Proof of Lemma 7. Since $\mathcal{X}$ is a compact subset of $\mathbb{R}^d$, there exists a covering $\mathcal{B} = \{B_1, \ldots, B_N\}$ of the sample space with balls of size ϵ_n where $N = [\xi/\epsilon_n]^d$ for some $\xi > 0$. Let x_j denote the center of B_j. Note that $P(X \in B_j) \geq a\, \epsilon_n^d v_d$ where v_d is the volume of the unit ball. Then

$$\begin{aligned}
P^n(F_n^c) &= P^n(\text{some } B_j \text{ is empty}) \leq \sum_{j=1}^N P^n(B_j \text{ is empty}) \\
&= \sum_{j=1}^N \prod_{i=1}^n P(X_i \notin B_j) = \sum_{j=1}^N \prod_{i=1}^n [1 - P(X_i \in B_j)] \\
&\leq \sum_{j=1}^N \prod_{i=1}^n [1 - a\, v_d\, \epsilon_n^d] = \sum_{j=1}^N [1 - a\, v_d\, \epsilon_n^d]^n = N[1 - a\, v_d\, \epsilon_n^d]^n \\
&\leq N e^{-n\, a\, v_d \epsilon_n^d} = \frac{\xi^d\, n}{\log n} \left(\frac{1}{n}\right)^{a\, v_d\, r} \to 0
\end{aligned}$$

since $r > 1/(av_d)$. Hence $P^n(F_n) \to 1$. □

Lemma 8. *Let p be a Morse function with finitely many critical points and modes $\mathcal{M} = \{m_1, \ldots, m_k\}$. There exists positive constants c_1 and c_2 such that*

the following is true. For every $x \in \left(\bigcup_{j=1}^{k} B(m_j, c_1\epsilon)\right)^c$, *there exists a ball* B *of radius* $2\epsilon_n$ *such that:*

1. $\max_{z\in B} ||z - x|| \leq c_2 t_n(x)$ *and*
2. $x \notin B$,
3. $p(z) > p(z)$ *for all* $z \in B$.

Proof of Lemma 8. Let $S = \{s_1, \ldots, s_N\}$ be the set of critical points that are not modes. Let $u_n = c_1\epsilon_n/2$.

Let A_s denote the supremum of all the s^{th}-order partial derivatives of p for $s = 0, 1, 2, 3$. (Hence, $A_0 = \sup_x p(x)$.)

Case 1: Suppose that $||x - s_j|| \leq u_n$ for some $s_j \in S$. Note that s_j cannot be a mode since $u_n < c_1\epsilon_n$. Let λ_j be the largest eigenvalue of $H(s_j)$ and note that $\lambda_j > 0$ since s_j is not a mode. Let v be the corresponding eigenvector and define $B \equiv B(y, 2\epsilon_n)$ where

$$y = x + c_3\epsilon_n v$$

where $c_3 \equiv c_3(x)$ is such that

$$\max\left\{2, 4c_1\sqrt{2}q_j/\lambda_j,\ \sqrt{\frac{16}{\lambda_j}[\sqrt{2}q_j + 4A_2]},\ \frac{32A_2}{\lambda_j}\right\} < c_3 < c_2\left(\frac{C}{A_0}\right)^{1/d} - 2. \quad (25)$$

Here, c_1 and c_2 are any positive constants such that the above interval is nonempty. Let q_j^2 be the largest eigenvalue of $H^2(s_j)/2$. Because p is Morse, $q_j^2 > 0$.

1. Let $z \in B$. Then

$$||z - x|| \leq ||z - y|| + ||y - x|| \leq 2\epsilon_n + c_3\epsilon_n \leq c_2 t_n(x)$$

where we used the fact (from (25)) that $c_3 \leq c_2(C/A_0)^{1/d} - 2$ which implies that $c_3 \leq c_2(p(x)/A_0)^{1/d} - 2$ and hence $(2 + c_3)\epsilon_n \leq c_2 t_n(x)$.

2. Next, note that

$$\min_{z\in B} ||z - x|| = ||x - y|| - 2\epsilon_n = c_3\epsilon_n - 2\epsilon_n > 0$$

since $c_3 > 2$. Hence, $x \notin B$.

3. For all $0 \leq r \leq 1$, $v^T H(ry + (1-r)x)v \geq v^T H(s_j)v - O(\epsilon_n) \geq \lambda_j/2$ for all large n. So

$$\begin{aligned} p(y) &= p(x) + c_3\epsilon_n v^T g(x) + \frac{c_3^2\epsilon_n^2}{2} v^T \int_0^1 H(sy + (1-s)x)ds\ v \\ &\geq p(x) + c_3\epsilon_n v^T g(x) + \frac{c_3^2\epsilon_n^2\lambda_j}{4} \geq p(x) + \frac{c_3^2\epsilon_n^2\lambda_j}{4} - c_3\epsilon_n|v^T g(x)|. \end{aligned} \quad (26)$$

Now

$$g(x) = g(s_j) + H(s_j)(x - s_j) + R_n = H(s_j)(x - s_j) + R_n$$

where the norm of the remainder R_n is bounded by $\sqrt{d}A_3u_n^2$. So, for all large n,

$$g(x)^T g(x) = (x - s_j)^T H^2(s_j)(x - s_j) + O(u_n^3) \leq u_n^2 q_j^2 + O(u_n^3) \leq 2u_n^2 q_j^2.$$

So $||g(x)|| \leq \sqrt{2}q_j u_n$ and so $c_3\epsilon_n |v^T g(x)| \leq c_1 c_3 \sqrt{2} q_j \epsilon_n^2/2$ and from (26) we conclude that

$$p(y) \geq p(x) - \frac{c_1 c_3 \sqrt{2} q_j \epsilon_n^2}{2} + \frac{c_3^2 \epsilon_n^2 \lambda_j}{4} \geq p(x) + \frac{c_3^2 \epsilon_n^2 \lambda_j}{8} \tag{27}$$

since $c_3 \geq 8c_1\sqrt{2}q_j/(2\lambda_j)$.

Now consider any $z \in B$. Then

$$p(z) - p(x) = p(z) - p(y) + p(y) - p(x) \geq p(z) - p(y) + \frac{c_3^2 \epsilon_n^2 \lambda_j}{8}.$$

We have

$$\begin{aligned} p(z) &= p(y) + (z - y)^T g(y) + R_n > p(x) + \frac{c_3^2 \epsilon_n^2 \lambda_j}{8} + (z - y)^T g(y) + R_n \\ &= p(x) + \frac{c_3^2 \epsilon_n^2 \lambda_j}{8} + (z - y)^T [g(x) + \widetilde{R}_n] + R_n \\ &= p(x) + \frac{c_3^2 \epsilon_n^2 \lambda_j}{8} + (z - y)^T g(x) + (z - y)\widetilde{R}_n + R_n \end{aligned}$$

where

$$|(z - y)^T g(x)| \leq ||z - y||\ ||g(x)|| \leq 2\epsilon_n \sqrt{2} q u_n = \sqrt{2}\epsilon_n^2 q,$$
$$|R_n| \leq ||z - y||^2 A_2 \leq 4\epsilon_n^2 A_2$$

and

$$||(z - y)\widetilde{R}_n|| \leq ||z - y||\ ||y - x|| A_2 \leq (2\epsilon_n)(c_3\epsilon_n)A_2 = 2c_3\epsilon_n^2 A_2$$

so that

$$p(z) > p(x) + \frac{c_3^2 \epsilon_n^2 \lambda_j}{8} - \sqrt{2}\epsilon_n^2 q - 4\epsilon_n^2 A_2 - 2c_3\epsilon_n^2 A_2 > p(x)$$

since

$$c_3 > \max\left\{ \sqrt{\frac{16}{\lambda_j}[\sqrt{2}q_j + 4A_2]},\ \frac{32A_2}{\lambda_j} \right\}.$$

Case 2. Suppose that $||x - s_j|| > u_n$ for all s_j. First, we will need to lower bound $||g(x)||$. By definition,

$$x \in \left[\left(\bigcup_r B(s_j, u_n) \right) \bigcup \left(\bigcup B(m_j, c_1\epsilon_n) \right) \right]^c.$$

For all large n, the minimum of $||g(x)||$ over this set occurs at the boundary of one of these balls. That is

$$||g(x)|| \geq \min\left\{ \min_{s_j \in S} \inf_{w \in \partial B(s_j, u_n)} ||g(w)||, \quad \min_{m_j \in \mathcal{M}} \inf_{w \in \partial B(m_j, c_1\epsilon_n)} ||g(w)|| \right\}.$$

Using a Taylor expansion of $g(w)$ as in Case 1, we then have that

$$||g(x)|| > \min_j c_1 \epsilon_n q_j/4 = c_1 \epsilon_n q/4. \tag{28}$$

Choose $c_3 \equiv c_3(x)$ such that

$$8\left(\frac{A_0}{C}\right)^{1/d} < c_3 < \min\left\{ c_2 - 2\left(\frac{A_0}{C}\right)^{1/d},\ c_1\left(\frac{a}{C}\right)^{1/d}\frac{q}{4A_2} \right\}. \tag{29}$$

Here, c_1 and c_2 are any positive constants such that the interval is nonempty. Let $B = B(y, 2\epsilon_n)$ where

$$y = x + \frac{c_3 t_n(x) g(x)}{||g(x)||}.$$

So $||y - x|| = c_3 t_n(x)$.

1. Let $z \in B$. Then

$$||z - x|| \leq ||z - y|| + ||y - x|| \leq 2\epsilon_n + c_3 t_n(x) \leq c_2 t(x)$$

since $c_3 \geq c_2 - 2(A_0/C)^{1/d}$.

2. Next

$$\min_{z \in B} ||z - x|| = ||x - y|| - 2\epsilon_n = c_3 t_n(x) - 2\epsilon_n > 0$$

since $c_3 > 2(A_0/C)^{1/d}$. So $x \notin B$.

3. First we note that

$$p(y) = p(x) + (y - x)^T g(x) + R_n = p(x) + c_3 t_n(x)||g(x)|| + R_n$$

where

$$|R_n| \leq ||y - x||^2 A_2/2 = c_3^2 t_n^2(x) A_2/2 < \frac{c_3 t_n(x)||g(x)||}{2}$$

since $||g(x)|| > c_1\epsilon_n q/4$ and $c_3 < (a/C)^{1/d} c_1 q/(4A_2)$. So

$$p(y) > p(x) + \frac{c_3 t_n(x)||g(x)||}{2}.$$

Now consider any $z \in B$. Then

$$p(z) - p(x) = p(z) - p(y) + p(y) - p(x) > p(z) - p(y) + \frac{c_3 t_n(x)||g(x)||}{2}.$$

Now

$$p(z) - p(y) = (z-y)^T g(y) + O(\epsilon_n^2)$$

and so

$$|p(z) - p(y)| \leq 2\epsilon_n ||g(y)|| + O(\epsilon_n^2) \leq 2\epsilon_n ||g(x)|| + O(\epsilon^2)$$

and this $p(z) > p(x)$ since

$$\frac{c_3 t_n(x) ||g(x)||}{2} > 2\epsilon_n ||g(x)||$$

since $c_3 > 8\left(\frac{A_0}{C}\right)^{1/d}$. □

References

Ery Arias-Castro, David Mason, and Bruno Pelletier. On the estimation of the gradient lines of a density and the consistency of the mean-shift algorithm. *Journal of Machine Learning Research*, 2015. MR3491137

Chacón. Clusters and water flows: a novel approach to modal clustering through morse theory. *arXiv preprint arXiv:1212.1384*, 2012.

José E Chacón, Tarn Duong, et al. Data-driven density derivative estimation, with applications to nonparametric clustering and bump hunting. *Electronic Journal of Statistics*, 7:499–532, 2013. MR3035264

José E Chacón et al. A population background for nonparametric density-based clustering. *Statistical Science*, 30(4):518–532, 2015. MR3432839

Frédéric Chazal, Brittany T Fasy, Fabrizio Lecci, Bertrand Michel, Alessandro Rinaldo, and Larry Wasserman. Robust topological inference: Distance to a measure and kernel distance. *To Appear: Journal of Machine Learning Research*, 2017. MR3813808

Yizong Cheng. Mean shift, mode seeking, and clustering. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 17(8):790–799, 1995.

Dorin Comaniciu and Peter Meer. Mean shift: A robust approach toward feature space analysis. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(5):603–619, 2002.

Vincent Courjault-Radé, Ludovic D'Estampes, and Stéphane Puechmorel. Improved density peak clustering for large datasets. 2016.

Mingjing Du, Shifei Ding, and Hongjie Jia. Study on density peaks clustering based on k-nearest neighbors and principal component analysis. *Knowledge-Based Systems*, 99:135–145, 2016.

Christopher R Genovese, Marco Perone-Pacifico, Isabella Verdinelli, and Larry Wasserman. Non-parametric inference for density modes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(1):99–126, 2016. MR3453648

Heinrich Jiang and Samory Kpotufe. Modal-set estimation with an application to clustering. *arXiv preprint arXiv:1606.04166*, 2016.

Jia Li, Surajit Ray, and Bruce G Lindsay. A nonparametric statistical approach to clustering via mode identification. *Journal of Machine Learning Research*, 8(Aug):1687–1723, 2007. MR2332445

John Milnor. *Morse Theory.(AM-51)*, volume 51. Princeton university press, 2016. MR0163331

Alex Rodriguez and Alessandro Laio. Clustering by fast search and find of density peaks. *Science*, 344(6191):1492–1496, 2014.

Xiao-Feng Wang and Yifan Xu. Fast clustering using adaptive density peak detection. *Statistical methods in medical research*, pages 2800–2811, 2017. MR3738283