A kernel density smoothing method for determining an optimal number of clusters in continuous data

J. Bugrien[1], K. Mwitondi[2] & F. Shuweihdi[3]
[1]*Statistics Department, Benghazi University, Libya*
[2]*Department of Computing, Sheffield Hallam University, UK*
[3]*School of Mathematics, University of Leeds, UK*

Abstract

While data clustering algorithms are becoming increasingly popular across scientific, industrial and social data mining applications, model complexity remains a major challenge. Most clustering algorithms do not incorporate a mechanism for finding an optimal scale parameter that corresponds to an appropriate number of clusters. We propose (BASINS^{-1}), a kernel-density smoothing-based approach to data clustering. Its main ideas derive from two unsupervised clustering approaches – kernel density estimation (KDE) and scale-spacing clustering (SSC). The novel method determines the optimal number of clusters by first finding dense regions in data before separating them based on data-dependent parameter estimates. The optimal number of clusters is determined from different levels of smoothing after the inherent number of arbitrary shape clusters has been detected without a priori information. We demonstrate the applicability of the proposed method under both nested and non-nested hierarchical clustering methodologies. Simulated and real data results are presented to validate the performance of the method, with repeated runs showing high accuracy and reliability.

Keywords: BASINS^{-1}, data clustering, data mining, kernel density estimation, local optimization, scale-space clustering, supervised learning, unsupervised learning.

WIT Transactions on Information and Communication Technologies, Vol 47, © 2014 WIT Press
www.witpress.com, ISSN 1743-3517 (on-line)
doi:10.2495/RISK140151

1 Introduction

Real-world objects are composed of different structures with different levels of detail. An object may therefore appear differently depending on the scale of observations. Lindeberg [1] described the framework which makes it possible to analyse a scene of an object at any scale level or in fact, at all scale levels simultaneously, as *multi-scale representation* of the scene. Such a representation is composed of successive versions of the original dataset or coarser scales. It is assumed then, the bigger the scale, the less information referred to local characteristics of the input data will appear. Also it is imposed that the general information applying to large scales will last through scale (causality). Taking that into account, it is reasonable to think that local and high resolution scale information can be related to general and low resolution information. We present this reasoning in a data clustering context focusing on "interesting" data features typically described by modes (local maxima), anti-modes (local minima) and *bumps*. As in Lindeberg [1], given independently identically distributed random vectors $\{x_1, x_2, \dots, x_n\} \in \mathbb{R}^d$, the scale-space filtering approach to clustering requires that the function $\hat{f}_t(x)$ correspond to a smooth indicator function $I(x; t)$ generated by convolution of the original data with a Gaussian kernel and a fixed isotropic scale parameter $t > 0$ provided that semi group property holds. As the scale increases, $\hat{f}_t(x)$ and/or $I(x; t)$ yield an increasingly coarser structure with fine details disappearing with an increasing scale (causality). The number of maxima in the data tends to decrease due to merging with other critical points – i.e., $t \to 0$ the local maxima tend towards the number of observations, *n,* and as $t \to \infty$ they tend towards a single local (global) maximum. Each local maximum defines a cluster and so for a fixed *t* each mode produces a clustering pattern by which each $\{x_1, x_2, \dots, x_{n-1}, x_n\} \in \mathbb{R}^d$ can be assigned to a cluster $\hat{f}_t(x)$.

We propose two clustering approaches that find dense regions in data based on density estimation and our novel notion of *inverted rainfall basins* (BASINS^{-1}). The main idea of (BASINS^{-1}) is that if we imagine a raindrop falling into a point $\boldsymbol{x} \in \mathbb{R}^{\boldsymbol{d}}$, then we can define an *inverted rainfall basin* of each mode as the region of all points for which raindrops fall under gravity to the corresponding local minimum of $-\hat{f}_t(x)$. The basins can be found by an iterative mode-seeking algorithm such as gradient ascent and so we can associate almost every point $x \in \mathbb{R}^d$ with a unique mode. The paper is organized into four sections – with the methods in Section 2, implementation in 3 and concluding remarks in 4.

2 Methods

The (BASINS^{-1}) approach to data clustering has statistical links to two conventional methods to data clustering – the Kernel Density Estimation (KDE) (Silverman [2]) and the space-scale theory (Witkin [3]) described below.

2.1 Kernel density estimation (KDE)

KDE is a well-documented method (Bowman and Azzalini [4], Wand and Jones [5] and Scott [6]). Given a sample $x \in \mathbb{R}^d$ from some unknown density, the general form of a multivariate kernel density estimate at x is computed as

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

where T is a symmetric positive defined *d* by *d* bandwidth matrix $PDS_{(d)}$ and K is a *d*-variate function (the kernel), usually assumed to be a *pdf* where

$$K_T(y) = |T|^{-1/2} K\left(T^{-1/2} y\right) \tag{2}$$

The kernel is generally taken to be an even, $K(x) = K(-x)$ bounded function, centered and scaled to satisfy the relationships $\int_{\mathbb{R}^d} xK(x)dx = 0$ and $\int_{\mathbb{R}^d} xx^T(x)dx = c_k I$ where c_k is a constant. The global bandwidth kernel estimator (1) is equivalent to a mixture density with the function *K(*)* equally weighted and centered at each x_i. If $K(*)$ is assumed to be a density function, then $\hat{f}_T(x)$ is nonnegative and integrates to unity. Using a fully parameterized *T* increases the complexity of the estimation and, in practice, the bandwidth matrix *T* is chosen either as a diagonal, $T = \text{diag}[t_1, t_2, \ldots, t_{n-1}, t_n]$ or proportional to the identity matrix *T=tI*.

The kernel estimator is a sum of the height of *bumps* placed at the observations. That is, the kernel *K* determines the shape of the *bumps*, while the bandwidth *t* determines their width. Employing only one bandwidth parameter, the kernel density estimator (1) becomes the well-known expression (3) in Silverman [2].

$$\hat{f}_t(x) = \frac{1}{nt^{d/2}} \sum_{i=1}^{n} K_t\left(\frac{x - x_i}{t^{1/2}}\right) \tag{3}$$

The bandwidth *t*, also called the smoothing parameter is a rescaling factor which determines the extent of the region over which the probability mass for a point x_i is spread and also controls the degree of smoothing. A small value of *t* lead to under-smoothed estimates (spurious peaks), while a large value leads to over-smoothed estimates (masking effect). Thus, the quality of a kernel density estimator depends on a choice of the smoothing parameter *t*. One way to estimate an optimum value of *t* is by measuring the mean of the squared error between the density and its estimate integrated over the domain of definition (MISE), i.e.,

$$\int_{\mathbb{R}^d} E\left(\hat{f}_t(x) - f(x)\right)^2 dx \tag{4}$$

However, only an asymptotic approximation of this measure (AMISE) can be estimated. An objective or data-driven choices of t can be made, for which a wide range of methods have been proposed and described in detail as in Wand

and Jones [5]. The usual approach in constructing the KDE is to fix the kernel $K(x)$ in $\hat{f}_t(x)$ and then assess the bandwidth t given data. Silverman [2] notes that although the shape of the resulting KDE does not depend on a choice of origin and is relatively insensitive to the exact form of the kernel, the choice of the bandwidth t is by far more important than the shape of the KDE. Our proposed approach based on the multivariate Gaussian kernel density

$$K_G(x) = (2\pi)^{-d/2}\exp\left\{-\frac{\|x\|^2}{2}\right\} \tag{5}$$

We define local modes or clusters of the data at scale t, whereby, analogous to the identity $\hat{f}_t(x) \equiv clusters$, we define a novel identity based on "rainfall basins". Each rainfall basin forms a cluster which can be found by an iterative mode-seeking algorithm with almost every point $\boldsymbol{x} \in \mathbb{R}^{\boldsymbol{d}}$ (except for saddle points and points converging to them) being associated with a unique mode.

2.2 Scale-space approach to clustering

A continuous multi-scale representation for a measured signal is obtained by embedding a continuous, bounded and integrable signal $\hat{f}(x), x \in \mathbb{R}^d$ into a one-parameter family of derived signals (the scale-space) where the scale parameter $t \in \mathbb{R}_+$ is intended to describe the current level of scale (Witkin [3]). Thus, given a signal $f\colon \mathbb{R} \to \mathbb{R}$, the scale-space representation $I\colon \mathbb{R} \times \mathbb{R}_+ \to \mathbb{R}$ is

$$I(x;t) = K_G(x;t) * f(x) = \int_{-\infty}^{\infty} K_G(\xi;t) f(x-\xi)d\xi \tag{6}$$

where $K_G(x;t)$ is the one dimensional Gaussian kernel defined as

$$K_G(x;t) = \frac{1}{\sqrt{2\pi t}}\exp\left\{-\frac{x^2}{2t}\right\}; x \in \mathbb{R}, t > 0 \tag{7}$$

As $t > 0$ the scale-space representation $I(x;t) \to f(x)\ if\ f(x)$ is continuous. The main idea behind this construction of the Gaussian scale-space representation is that the fine-scale information should be suppressed with increasing values of the scale parameter t. Intuitively, when convolving a signal $f(x)$ with a Gaussian kernel $K_G(x;t)$ the effect is to suppress most of the structures in $f(x)$ with a characteristic length less than $t^{1/2}$. Lindeberg [1] highlights a number of intuitive demands associated with a multi-scale representation such as the semi-group property, rotation invariance, homogeneity and isotropy, separability and scale invariance. We focus on the property

$$L_{t_1+t_2}f = L_{t_1}\left(L_{t_2}f\right), \forall\colon t_{i=1,2}\ \forall\colon f \tag{8}$$

The foregoing property ensures that one can implement the scale-space as a cascade smoothing, which means that, if two kernels convolved with each other the resulting kernel is of the same family. Since each scale-space is simply a family of Gaussian kernel smoothes indexed by the bandwidth t, instead of choosing one level of smoothing, one should consider the full range of smoothes (the whole scale-space) which corresponds to viewing the data as a number of different levels of resolution, each of which may contain useful information. This concept allows to practical solution to the classical problem of choice of the level of smoothing (bandwidth) when it can be viewed in an entirely new way. Scale-space filtering-based approaches to clustering go back some years with examples in Wilson and Spann [7] who proposed an iterative algorithm based on the estimation theory. Their method is a unification of location and mode estimation achieved by considering the effect of spatial scale on estimator. They consider an important shift in the evaluation of the clustering problem in which they require estimates of valid structure within a dataset to be robust to both the presence of outliers and to spatial scale changes of the data. This idea was further developed by Roberts [8] who developed a scale-space filtering-based approach for assessing the probable number of clusters within datasets based on the computation of maxima of the scale-space representation $(x; t)$. Effectively, the approach utilises methods in Lindeberg [1] for extracting robust structures in data and it may be seen as falling within the hierarchical clustering methods based on successive smoothing using a Gaussian kernel function. A combination of the foregoing notions was later developed by Leung *et al.* [9], by mimicking how human eyes unravel intrinsic structures in images.

Clustering by scale-space filtering performs clustering through a blurring (smoothing) process, which treats an image as a dataset with each data point being a light point attached with a uniform luminous flux. We illustrate our approach to clustering via hierarchical clustering in scale-space as follows. The generalized function for the empirical distribution for $x \in \mathbb{R}$ is

$$f_{emp}(x) = n^{-1} \sum_{i=1}^{n} \delta(x - x_i) \tag{9}$$

i.e., for $t > 0$ the smoothed function $I(x; t)$ in (6) is a convolution of the form

$$I(x; t) = K_G(x; t) * f_{emp}(x) \tag{10}$$

Thus, data clusters may be defined as peaks in $I(x; t)$ and hence the number of clusters and their locations may be evaluated from the peaks of $I(x; t)$ or the positive-negative zero-crossings of its spatial derivative. Finding the set of zero-crossings of $\partial I(x; t)/\partial x$ is equivalent to estimating the positions of the extrema of $I(x; t)$. As the scale increases, $I(x; t)$ represents a coarser structure. So by applying this scale-space filtering process, a family of smooth images $I(x; t)$ is

$$I(x; t) = n^{-1} \sum_{i=1}^{n} \frac{1}{\sqrt{2\pi t}} \exp\left\{-\frac{(x - x_i)^2}{2t}\right\} \tag{11}$$

Stationary points of $I(x; t)$ can be determined by taking the spatial derivative below and setting to zero

$$\frac{\partial I(x;t)}{\partial x} = (nt)^{-1} \sum_{i=1}^{n} \frac{(x - x_i)}{\sqrt{2\pi t}} exp\left\{-\frac{(x - {x_i}^2)}{2t} = 0\right\} \tag{12}$$

The variation in local maxima as the kernel density bandwidth parameter varies presents a major data clustering challenge. Balancing the lower scale (potential “spurious” information or over-fitting and the higher scale (potential information “masking” or under-fitting (as in Figure 1) – constitutes our main motivation. We now introduce our proposed clustering methodology (BASINS^{-1}).

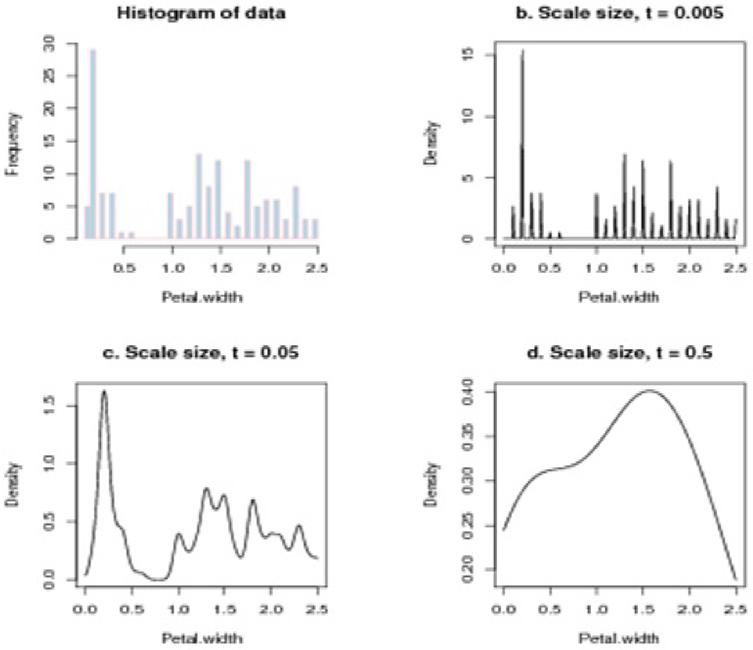

Figure 1: Kernel density estimates for different scales for the iris data.

2.3 Our proposed clustering approach – BASINS^{-1}

Insight into the BASINS^{-1} methodology comes from two traditions in statistics and signal processing – namely Kernel density estimation (KDE) (Silverman [2]) and scale-space theory (Lindeberg [1]). KDE cluster centers are derived by local mode seeking identifying maxima in the normalized density of the dataset while under the scale-space theory, they are identified in different scales – typically finding an important underlying structure at several different levels of smoothing. Our proposed strategy relates the optimal number of clusters to different levels of data smoothing (bandwidth). Thus, we describe the novel clustering approach based on the local extrema (minima and maxima) of a density function for a fixed bandwidth and how these local extrema vary as the bandwidth varies. We show that for a small bandwidth $t,$ there will be n local maxima, one converging to each data point as $t \to 0$ and hence the method is independent of initialization. Further, each local maxima lies between two local minima – that is, there are *n* local maxima and *n+1* local minima including $\pm\infty$. In this case the smoothing function $I(x;t)$ at small t presents a series of *n* steep hills centered at each of the data points and as $t \to \infty$ $I(x;t)$ will become a curve with one global maximum which means that all data points merge together to produce one cluster determined by its global maximum. To describe the clustering process, first we need to present some definitions:

- Let $\mathrm{n_t}$ be the number of local maxima at scale t. Then, for small t, $\mathrm{n_t = n}$.
- Let $\mathrm{m_i(t),\ 1 \leq i \leq n_i}$ denote the values of the local $\mathrm{I(x;t)}$ at scale t such that for small $\mathrm{t, m_i(t) \to x_i}$ as $\mathrm{t \to 0}$. Hence, there are $\mathrm{n_t = n}$ local maxima and the notation is unambiguous.
- Let $\pi_\mathrm{j} \in \{1 \ldots \mathrm{n}\}$ be the local maximum which disappears through $\mathrm{t_j}$ and $\mathrm{n_t}$ decreases by 1 at $\mathrm{t_j}$.
- Strictly speaking $\mathrm{m_{\pi j}(t)}$ is undefined for $\mathrm{t} > t_\mathrm{j}$. Hence for labeling purposes we set $\mathrm{m_{\pi j}(t)} = \Delta$ where Δ denotes a "coffin bin" for $\mathrm{t} > t_\mathrm{j}$.

One of the scale-space properties stipulates that there exist "smooth paths" $m_{\pi j}(t) \in \mathbb{R}$ that link the local maxima together at different levels of scale t. The paths (Figure 3) are called "maximal curves" $m_{\pi j}(t), 0 \leq t \leq t_j < \infty$ such that:

- $\mathrm{m_i(t)}$ present cluster locations at scale t.
- No new local maxima can appear as t increases (causality).
- At scale t, $\mathrm{n - n_t}$ values of $\mathrm{m_i(t)}$ take the value Δ.

The local extrema of (12) can produce a pattern of clustering of a smoothed scale-space representation $I(x;t)$ at a given scale t by which each data point $x \in \mathbb{R}$ can be allocated to a cluster based on the notion of "inverted rainfall basins" (BASINS^{-1}). The basins can also be referred to as "domains for attraction" as they attract the set of all local maxima for an increasing scale t to the same mode. Hence, for a small t we can define a family of domains $D_i^{(t)}$, as intervals between local minima, exist for $0 < t < t_j$, such that:

- For small enough t if $x_i s$ are in increasing order, each x_i has a corresponding $D_i^{(t)}$.
- As $t \to 0$, $D_i^{(t)}$ converges to an interval, $\left[\frac{x_{i-1}+x_i}{2}, \frac{x_i+x_{i+1}}{2}\right]$ including $\pm\infty$.
- $D_i^{(t)}$ is defined for $m_{\pi j}(t) \neq \Delta$.
- Given scale t, for each local maximum $m_i(t)$, $j = 1 \ldots n_t$, there exists a "domain of attraction" $D_{j<1}^{(t)}$.

We can now define $C_i^{(t_i)}$ clusters at each scale t_i where each $C_i^{(t_i)}$ is determined by its local maximum $m_i(t) \neq \Delta$ such that $C_i^{(t_i)}$ is the union of domains over the local maxima

$$\boldsymbol{C}_{\boldsymbol{i}}^{(\boldsymbol{t}_{\boldsymbol{i}})} = \bigcup\nolimits_{\boldsymbol{m}_{\boldsymbol{j}}(\boldsymbol{t}) \in \boldsymbol{D}_{\boldsymbol{i}}^{(\boldsymbol{t})}} \boldsymbol{D}_{\boldsymbol{j}}^{(\boldsymbol{t})} \tag{13}$$

As t passes through t_i; $t_1 < t_2 < \cdots . < t_n$ one of the clusters $C_i^{(t_i)}$ disappears and gets re-allocated to merge with the left or right $D_j^{(t)}$ such that $m_j(t_i^-) \in D_j^{(t^+)}$ and at the limit (threshold t_i), $C_j^{(t_i)} = D_j^{(t^+)}$. That is, if we regard t_i as a merging threshold, then the scale-space evolution of local maxima in $I(x;t)$, may be regarded as a form of dendogram - a tree showing a sequence of clustering. As the scale t changes, a hierarchical clustering is formed with a dendogram output similar to that formed from ordinary hierarchical clustering methods. Thus, clustering of each datum amounts to evaluating which "domain of attraction" each x_i belongs to. Figure 2 provides a graphical illustration of hierarchical clustering in the context of our proposed method.

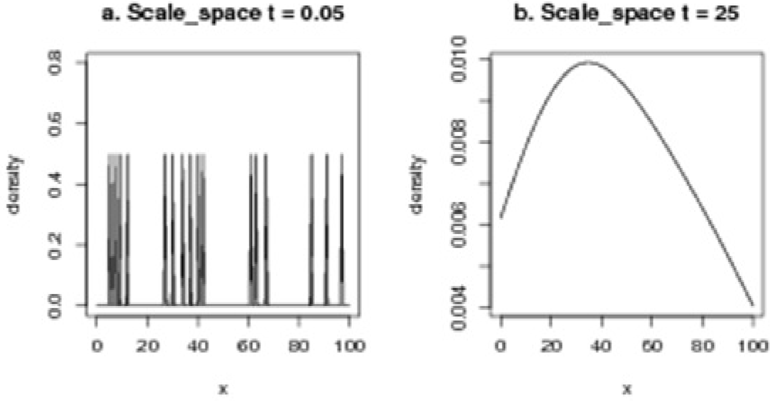

Figure 2: Steep hills (LHS) and one global maximum (RHS).

As in Leung *et al.* [9], hierarchical clustering may be categorised into *nested* and *non-nested*. In the former, once a cluster is formed, its members cannot be separated as $\boldsymbol{t}$ increases while in the latter, each datum in a formed cluster may change its boundaries with changing scale. The underlying mechanics of BASINS^{-1} can be described under both forms of hierarchical clustering.

2.3.1 BASINS^{-1} under nested hierarchical clustering

The main idea here is that changing the scale t and looking at each scale t_i separately produces a nested hierarchical clustering $C_i^{(t_i)}$, where once any two clusters merge, they stay merged. The nested hierarchical clustering process based on $D_i^{(0)}$ can be described as follows. For small t, $C_i^{(t)} = D_i^{(0)}$ as t passes through t_j, and for some i, $m_{\pi j}(t_j) \in D_j^{(t_j^+)}$ set $C_i^{(t_j^+)} = C_i^{(t_j^-)} \cup D_{\pi j}^{(0)}$, where other $C_i^{(t_j^+)}$remain unchanged as t passes through t_j – i.e., constant clusters for $t_j < t < t_{j+1}$. The process follows the following steps

$$0 \leftarrow (\text{Step 1}) \rightarrow t_1 \leftarrow (\text{Step 2}) \rightarrow t_2 \leftarrow (\text{Step 3}) \rightarrow \cdots \leftarrow t_{n-1} \leftarrow (\text{Step n}) \rightarrow t_n \tag{14}$$

where at the last step, n, only one local maximum is left – the global maximum and described as follows.

1) Start with n domains of attraction, $D_i^{(0^+)} = C_j^{(Step\ 1)}$, $i = 1, 2, \dots, n-1, n$.
2) At scale $\mathrm{t} = \mathrm{t}_1$, the cluster $\mathrm{C}_{(1)}^{(\mathrm{Step\ 1})}$ disappears as it is absorbed into an adjacent cluster (left or right), $\mathrm{C}_{(1)}^{(\mathrm{Step\ 1})}$, say, which means that the local maximum $m_1(t_1) \in D_{(1)}^{(t_1^+)}$ - that is, for t a bit bigger than $t_1, m_1(t_1)$ disappears and hence falls into another domain $D_{(1)}^{(t_1^+)}$.
3) For $t_1 < t < t_2$ and $k < l$ two clusters are merged as

$$\begin{cases} C_{(k)}^{(Step\ 2)} = C_{(k)}^{(Step\ 1)}\ if\ k \neq 1 \text{ or } l \\ C_{(k)}^{(Step\ 2)} = \phi \\ C_{(l)}^{(Step\ 2)} = C_{(1)}^{(Step\ 1)} \bigcup C_{(1)}^{(Step\ 1)} \end{cases} \tag{15}$$

4) Repeat step 3 until at $\mathrm{t} = \mathrm{t_n}$ all clusters merge into one cluster.

2.3.2 BASINS^{-1} under non-nested hierarchical clustering

As in the previous case, changing the scale t continuously and looking at the range of continuous changes of $t_i < t < t_{i+1}$ produce a non-nested hierarchical clustering $C_j^{(t)}$ where boundaries may move. The non-nested hierarchical clustering process is described by $C_i^{(t)} = D_i^{(t)}$ and defined for $i \in \{1,2 \dots n-1, n\}$ such that $m_i(t) \neq \Delta$. All clusters $C_j^{(t)}$ at a given scale t are unions of the initial clusters $C_i^{(0)} = x_i : \forall_i$. We require that “true” clusters in the dataset be stable over a range of scale parameters, t.

2.3.3 Obtaining the optimal scale parameter

In order to determine the optimal scale parameter $t_{(opt)}$, the Gaussian kernel is essential as it is the only kernel which does not introduce new maxima as the scale increases (Babaud *et al.* [10]; Koenderink [11]; Roberts [8]; Silverman [2] and Yuille and Poggio [12]). We suggest the following simple method:

1) For the Gaussian kernel, it can be shown in one dimension that the number of modes, n_t, say, is a decreasing function of the bandwidth, t.

2) Given β, let $t_{max}(\beta)$ and $t_{min}(\beta)$ denote the maximum and minimum scales respectively yielding β clusters. Further, let $r(\beta) = t_{max}(\beta) - t_{min}(\beta)$ denote the range of t values over which $n_t = \beta$.
3) We simply look for values β for which $r(\beta)$ (linear in β) is the largest (Figure 4). Thus, we can then set the optimal scale parameter as

$$t_{opt} = \frac{t_{max}(\beta) + t_{min}(\beta)}{2} \tag{16}$$

3 Implementation of the method, results and discussions

We analyse both simulated and real data. For the former, we ran the algorithm over 100 simulations for each distribution in Table 1.

Table 1: Normal distribution with corresponding smoothing parameters.

No	Density	$\text{f(x)}=\sum_{j=1}^{J} p_l \emptyset_{\sigma_i}(x-\mu_i)$	Sample size n
1	One Gaussian	$N(0,1)$	100
2	Two Gaussian	$\frac{1}{2}N(0,1)+\frac{1}{2}N(3,1)$	200
3	Three Gaussian	$\frac{1}{3}N(0,1)+\frac{1}{3}N(3,1)+\frac{1}{3}N(6,1)$	300
4	Skewed Unimodal	$\frac{1}{5}N(0,1)+\frac{1}{5}N\left(\frac{1}{2},\frac{4}{9}\right)+\frac{3}{5}N\left(\frac{13}{12},\frac{25}{82}\right)$	300
5	Skewed Bimodal	$\frac{3}{4}N(0,1)+\frac{1}{4}N\left(\frac{3}{2},\frac{1}{9}\right)$	200
6	Trimodal	$\frac{9}{20}N\left(\frac{-6}{5},\frac{9}{25}\right)+\frac{9}{20}N\left(\frac{36}{25},\frac{9}{25}\right)+\frac{1}{10}N\left(0,\frac{1}{16}\right)$	600
7	Claw	$\frac{1}{2}N(0,1)+\frac{1}{10}\sum_{i=0}^{4}N\left(\frac{i}{2}-1,\frac{1}{100}\right)$	300
8	Smooth Comb	$\sum_{i=0}^{5}\frac{2^{5-i}}{63}N\left(65-96*\frac{i^2}{2},\frac{32/63}{2^{2i}}\right)$	300

The values of t were obtained by a sequence of length 200, uniformly spaced and for the modes, the kernel density estimates were computed over an equally spaced grid of 500 points. The values in the table represent 8 different normal densities, each of which can be expressed as mixture distribution as follows.

$$f(x) = \sum_{j=1}^{J} p_l \emptyset_{\sigma_i}(x - \mu_i) \tag{17}$$

Here $P = \{(p_1, \dots \dots, p_J): \sum_{j=1}^{J} = 1 : p_j \geq 0,\ j = 1,2, \dots ., J-1, J\}$ are the mixture component proportions, whereas $\{\mu_j\}_{j=1}^{J}$ and $\{\sigma_j\}_{j=1}^{J}$ are the population mean and variance of the normal density in (17) respectively. The clustering results are in Figure 3.

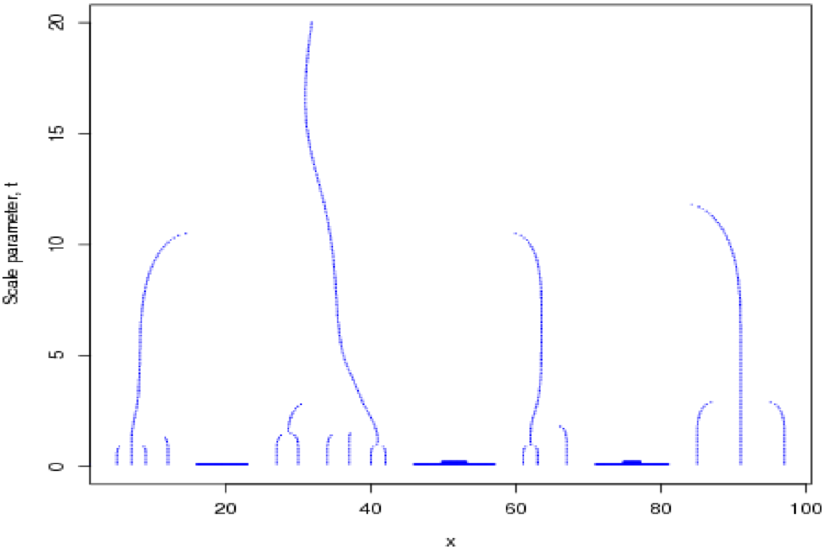

Figure 3: Paths of maximal curves through scale t.

As noted earlier, the “smooth paths” $m_{\pi j}(t) \in \mathbb{R}$ linking the local maxima together at different levels of bandwidth are necessary for cluster detection. Figure 3 presents these paths (maximal curves) at variable bandwidths $m_{\pi j}(t), 0 \leq t \leq t_j < \infty$ whereby each $\mathrm{m_i(t)}$ is a cluster location at scale t with no new local maxima appearing as the bandwidth t increases – avoiding spurious clusters. The results in Figure 3 correspond to the simulated data points in Table 1 and they provide an insight into the mechanism and movement of maximal curves as a function of the bandwidth. Note how the maximal curves vanish as they are siphoned into the nearest maximal curve based on the adopted density – i.e., the higher the density the longer the duration. For this reason, the number of maximal curves (clusters) decreases sequentially, until the last maximal curve (cluster) – one with the highest density is reached.

Note that for one Gaussian density, it is somewhat difficult to distinguish between the duration of two and three modes. Also, the underlying distribution of the duration of local maxima for four modes and more yields less than four local maxima indicating that the data consists of one mode. The duration of two local maxima is the longest for the two-Gaussian density while that of three local maxima is the longest for the three-Gaussian density. For the skewed unimodal density, two modes seem to dominate – suggesting that two modes are clearly considered for a skewed unimodal density. For the trimodal density, the size of middle cluster is somewhat smaller than the other two and not well separated from them - the longest duration is noted for two local maxima. As we get to the most difficult cases, the plot suggests one mode for the claw density with the durations of other local maxima standing very close to each other and not disappearing fast as is the case in the other densities. For the smooth comb density, the longest duration of mode is for two, three and four modes – sequentially. The reason for this is that the first two clusters are well separated in comparison with the rest of the clusters. However this pattern is not sustained as the number of clusters increase. Based on plots of median modes for the majority of the underlying densities, we found that the number of local maxima increases sharply as long as the real structure of the densities is smoothed away. Figure 4 presents clustering results of the real data – petal-width of iris data (Fisher [13]).

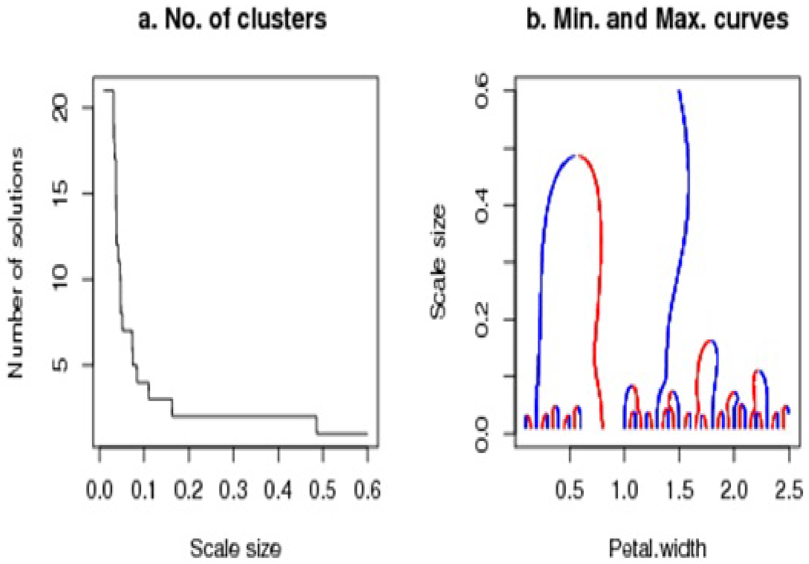

Figure 4: The number of local maxima for an increasing scale (LHS) and the corresponding positions of local maxima from the iris data (petal width) (RHS).

The two panels in Figure 4 are both based on the iris data (petal width) and they highlight the crucial issue of choosing the optimal scale parameter $t_{(opt)}$, based on the Gaussian kernel and our proposed method as described above. The maximal curves and density durations can be assessed based on the bandwidth and it can clearly be seen (RHS panel) that the number of modes n_t is a decreasing function of the bandwidth. The LHS panel plots the scale size against the number of solutions (clusters) and it can be seen that range of scale size 0.175 to 0.490 corresponds to three clusters. The optimal scale size is defined on the range $r(\beta) = t_{max}(\beta) - t_{min}(\beta)$ over which $n_t = \beta$ and since we look for values β for which $r(\beta)$ is the largest, the optimal scale parameter is

$$t_{opt} = \frac{0.175 + 0.49}{2} = 0.33 \tag{18}$$

4 Conclusion

Determining the natural number of clusters in data is one of the most difficult problems in cluster analysis and the relevance and challenges of mode-finding algorithms in data clustering applications are well-documented. This paper presented a novel method (BASINS^{-1}) for mode-finding based on conventional kernel and space-scale methodologies. Our proposed method was motivated by two conventional approaches to mode-finding – kernel density estimation (KDE) and scale-space theory (SST) – which we adapted to possible clustering approaches based on local maxima and minimal features in the data. Its basic idea is that modes can be associated with these important structures in empirical distributions. The relation between KDE and scale-space theory is well-investigated in the literature. The potential clustering power of the kernel density estimation is embedded in its estimation of the kernel from the original dataset and specification of an isotropic bandwidth which controls the amount of smoothing. These potentials were combined with the fundamental features of scale-space filtering (clustering) in which the kernel density $\hat{f}_t(x)$ corresponds to a smooth function $I(x;t)$ generated by convolution of the original dataset with a Gaussian kernel and a fixed isotropic scale parameter $t > 0$, provided that semi-group property holds. The novelty in our method derives from the fact that most clustering algorithms do not incorporate a mechanism for finding an optimal scale parameter that corresponds to an appropriate number of clusters.

Typically, the optimal scale parameter is normally performed via the so-called “strength of clusters” or “cluster validity” which quite often relies on computed distances among data points. Our clustering approach does not require the use of any “strength of clusters” criteria. Basically, assessing the strength of clusters is performed by examining and comparing the duration or survival period of clusters for increasing scale parameter t. We used simulated and real data to demonstrate how durations can be used to indicate the significance of a particular set of clusters and we were able to determine the optimal scale parameter and correspondingly the optimal number of clusters as a function of the longest lasting cluster set. We expect that the proposed approach will contribute to the ever growing portfolio of enhanced data mining methods for different applications – particularly those of a survival nature. Multi-disciplinary applications of the novel method on multivariate data should provide further validation of its accuracy and reliability (Mwitondi *et al.* [14]).

References

[1] Lindeberg, T.: Scale-space Theory in Computer Vision, Kluwer Academic Publishers, Dordrecht, Netherlands (1994).

[2] Silverman, B. W.: Density Estimation for Statistics and Data Analysis, Chapman and Hall, (1986).

[3] Witkin, A. P.: Scale-space Filtering, Proceedings of the Eighth International Joint Conference on Artificial Intelligence, Karlsruche, W. Germany, pp 1019-1022 (1983).

[4] Bowman, A. W. and Azzalini, A.: Applied Smoothing Techniques for Data Analysis: The Kernel Approach with S-Plus Illustrations, Oxford University Press, London; New York (1997).

[5] Wand, M. P. and Jones, M. C.: Kernel Smoothing, Chapman and Hall, (1995).

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

[7] Wilson, R. and Spann, M.: A New Approach to Clustering, Pattern Recognition, 23, pp 1413-425 (1990).

[8] Roberts, S. J.: Parametric and Non-parametric Unsupervised Cluster Analysis, Pattern Recognition, 30, 5, pp 261-272 (1997).

[9] Leung, Y., Zhang, J. and Xu, Z.: Clustering by Scale-space Filtering, IEEE Transactions on Pattern Analysis and Machine Intelligence, 22, 12, pp 1396-1410 (2000).

[10] Babaud, J., Witken, A. P. and Baudin, M.: Uniqueness of the Gaussian for Scale-space Filtering, IEEE Transactions on Pattern Analysis and Machine Intelligence, 8, 1, pp 26-33 (1986).

[11] Koenderink, J. J.: The Structure of Images, Biological Cybernetics, 50, pp 363-370 (1984).

[12] Yuille, A. L. and Poggio, T.A.: Scaling Theorems for Zero Crossings, IEEE Transactions on Pattern Analysis and Machine Intelligence, 8(1), pp 15-25 (1986).

[13] Fisher, R. A.: The Use of Multiple Measurements in Taxonomic Problems, Annals of Eugenics, 7, 4 pp 179-188 (1936).

[14] Mwitondi, K., Said, R. and Yousif, A.: A sequential data mining method for modelling solar magnetic cycles; Neural Information Processing, LNCS, Vol. 7663, pp 296-304, Springer (2012).