Multivariate Saddle Point Detection for Statistical Clustering

Dorin Comaniciu[1], Visvanathan Ramesh[1], and Alessio Del Bue[2]

[1] Imaging and Visualization Department
Siemens Corporate Research
755 College Road East, Princeton, NJ 08540, USA
[2] Department of Biophysical and Electronic Engineering
University of Genova
Via All'Opera Pia 11a, 16145 Genova, Italy

Abstract. Decomposition methods based on nonparametric density estimation define a cluster as the basin of attraction of a local maximum (mode) of the density function, with the cluster borders being represented by valleys surrounding the mode. To measure the significance of each delineated cluster we propose a test statistics that compares the estimated density of the mode with the estimated maximum density on the cluster boundary. While for a given kernel bandwidth the modes can be safely obtained by using the mean shift procedure, the detection of maximum density points on the cluster boundary (i.e., the saddle points) is not straightforward for multivariate data. We therefore develop a gradient-based iterative algorithm for saddle point detection and show its effectiveness in various data decomposition tasks. After finding the largest density saddle point associated with each cluster, we compute significance measures that allow formal hypothesis testing of cluster existence. The new statistical framework is extended and tested for the task of image segmentation.

Keywords: grouping and segmentation; image features; nonparametric clustering; cluster significance.

1 Introduction

Data clustering as a problem in pattern recognition and statistics belongs to the class of unsupervised learning. It essentially involves the search through the data for observations that are similar enough to be grouped together. There is a large body of literature on this topic [5,12,17]. Algorithms from graph theory [6, 10], matrix factorization [25,27], deterministic annealing [16], scale-space theory [20,26], and mixture models [7,23] were successfully used to delineate relevant structures within the input data.

However, the clustering task is inherently subjective. There is no accepted definition of the term *cluster* and any clustering algorithm will produce some partitions. Therefore, the ability to statistical characterize the decomposition

A. Heyden et al. (Eds.): ECCV 2002, LNCS 2352, pp. 561–576, 2002.

and to assess the significance of the resulted number of clusters is an important aspect of the problem.

Approaches for estimating the number of clusters can be divided into global and local methods. The former evaluate some measure over the entire data set and optimize it as function of the number of clusters. The latter consider individual pairs of clusters and test whether they should be joined together. A general description of methods used to estimate the number of clusters is provided in [18] while a study described in [21] conducts a Monte Carlo evaluation of 30 indices for cluster validation. These indices are typically functions of the within and between cluster distances and belong to the class of *internal* measures, in the sense that they are computed from the same observation used to create a partition. Consequently, their distribution is intractable and they are not suitable for hypothesis testing.

As a result, the majority of existing methods for estimating the validity of the decomposition do not attempt to perform a formal statistical procedure but rather look for a clustering structure under which the statistic of interest is optimal, i.e, maximize or minimize an objective function [19,24].

Validation methods that do not suffer from this limitation were recently proposed [29,30], but are computationally expensive, since they require simulating multiple datasets from the null distribution.

In this paper we present a new and practical approach to compute the statistical significance (p-value) of each delineated cluster. A nonparametric model is assumed in which the clusters correspond to local maxima (modes) in the probability density function of the data [8, p.533]. The test statistic that we develop compares the estimated density of the mode with the estimated density of the highest saddle point on the cluster boundary. To find the saddle points in the multivariate density surface we develop and test a gradient-based iterative algorithm. Note that for both the mode and saddle point finding only the density estimate and its normalized gradient are used (and not the Hessian matrix of second derivatives). Because of this, our the methods scale well with the space dimension.

The organization of the paper is as follows. Section 2 discusses the importance of the modes and saddle points of the density for characterizing the underlying data structure. The mean shift-based data decomposition is shortly reviewed in Section 3. Section 4 describes the new algorithm for saddle point detection, while the test statistic is developed in Section 5. Experimental results are shown in Section 6. The application of our new statistical framework for image segmentation is discussed in Section 7.

2 Importance of Modes and Saddle Points

Clustering using the nonparametric estimation of the data density is achieved by identifying local maxima (modes) and their basins of attractions in the multivariate surface of the data density function. The modes of the density are usually detected using the gradient ascent mean shift procedure [4,3], discussed in the

next section. All the data points lying in the basin of attraction of a mode will form a separated cluster. In the case of a density function with constant values at a peak, the points on this peak are considered a single mode, called a plateau. Similarly, all the data points lying in the basin of attraction of a plateau form a separated cluster.

The number of observed modes depends on the bandwidth of the kernel used to compute the density estimate. In general the number of modes decreases with the increase of the bandwidth. The most common test for the true number of modes in a population [28] is based on critical bandwidths, the infimum of those bandwidths for which the kernel density estimate is at most m-modal.

A different approach is proposed for the univariate case in [22] where the validity of each mode is tested separately. The test statistic is a measure of the size of the of mode, the absolute integrate difference between the estimated density and the same density with the mode in question removed at the level of the higher of its two surrounding antimodes. The p-value of the test is estimated through resampling. Note that an antimode is defined for the univariate data as the location with the lowest density between two modes. The main advantage of this technique is that each individual suspected mode is examined, while the bandwidth used in the test can be selected adaptively as smallest bandwidth at which the mode still remains a single object.

Our work is related to the technique in [22] by defining a similar procedure for the testing of multivariate data. However, since we would like to avoid resampling, we will define a test statistic whose distribution can be evaluated through statistical inference, by taking into account its sampling properties. In addition, since the antimodes defined for univariate data translate into saddle points for the multivariate case we will need an algorithm for saddle point computation.

To give the reader an initial view on the problem we present in Figure 1a a sample data set drawn from two bivariate normals, while Figure 1b, Figure 1c, and Figure 1d show the corresponding probability density estimate obtained with a two dimensional normal kernel with bandwidth $h = 0.6$, $h = 0.9$, and $h = 1.35$, respectively. The detected modes are marked with green dots, while the saddle points are marked with red dots.

A number of observations can be made using Figure 1. First, for a given bandwidth, the number of observed modes determines the number of distinct structures in the density estimate. The mode density is an indication of the compactness of the associated structure. The difference between the mode density and the saddle point density is an indication of the isolation of the observed structure. In addition, both the mode density and the mode-saddle density difference decrease with the increase of the bandwidth. When the mode density becomes equal to the saddle density the observed structures are amalgamated into a new one. Hence, the appropriate analysis bandwidth should be the smallest bandwidth at which the mode in question still remains a single object.

For a rigorous treatment of the evolution of the zero crossings of the gradient of a function along the bandwidth see [1,32]. The catastrophe theory [11] inves-

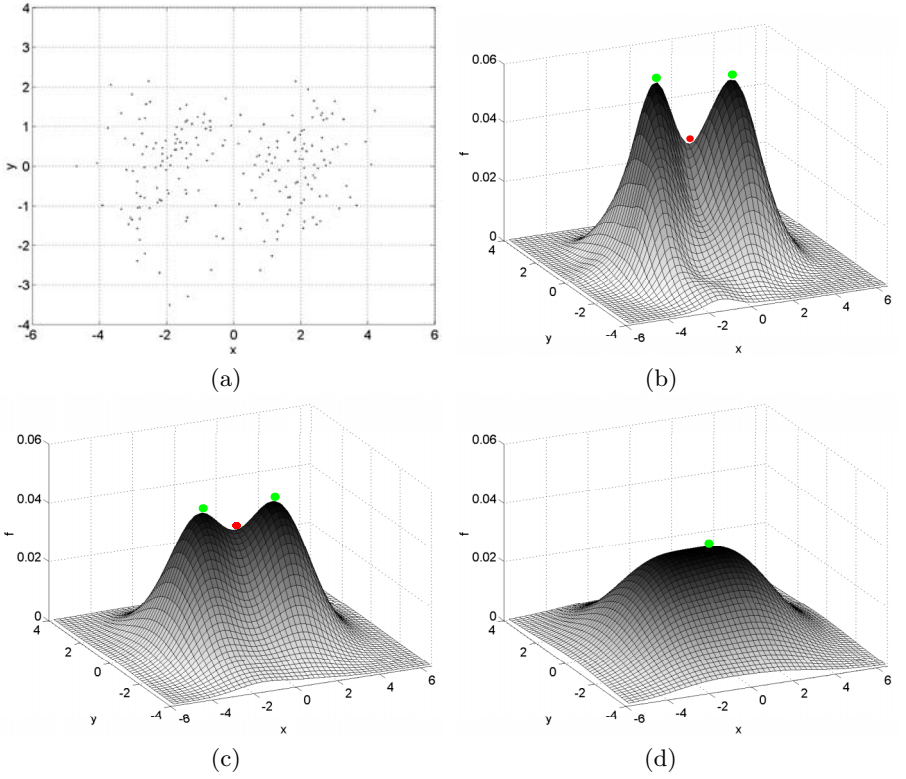

Fig. 1. Importance of modes and saddle points. (a) Input data containing 100 points from each bivariate N([-1.8,0],**I**) and N([1.8,0],**I**). (b) Density estimate with a two dimensional symmetric normal kernel with $h = 0.6$. The modes are marked with green dots, while the saddle point is marked with a red dot. (c) Density estimate with $h = 0.9$. (d) Density estimate with $h = 1.35$. The two modes and the saddle point collapse into one mode.

tigates the behavior of the singularities of a function in families of functions such as the family of densities generated by using various bandwidths.

3 Mean Shift Based Data Decomposition

In this section we define the mean shift vector, introduce the iterative mean shift procedure, and describe its use in the data decomposition.

3.1 The Mean Shift Procedure

Given n data points $\mathbf{x}_i$, $i = 1 \ldots n$ in the d-dimensional space R^d, the multivariate mean shift vector computed with kernel K in the point $\mathbf{x}$ is given by [9,3]

$$\mathbf{m}_K(\mathbf{x}) \equiv \frac{\sum_{i=1}^{n} \mathbf{x}_i K\left(\frac{\mathbf{x}-\mathbf{x}_i}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{\mathbf{x}-\mathbf{x}_i}{h}\right)} - \mathbf{x} \ , \tag{1}$$

where h is the kernel bandwidth. In the following we will use the symmetric normal kernel defined as

$$K(\mathbf{x}) = (2\pi)^{-d/2} \exp\left(-\frac{1}{2}\|\mathbf{x}\|^2\right) \tag{2}$$

It can be shown that the mean shift vector at location $\mathbf{x}$ is proportional to the normalized density gradient estimate computed with kernel K

$$\mathbf{m}_K(\mathbf{x}) = h^2 \frac{\hat{\nabla} f_K(\mathbf{x})}{\hat{f}_K(\mathbf{x})}. \tag{3}$$

The normalization is by the density estimate in $\mathbf{x}$ obtained with kernel K. Note that this formula changes a bit for kernels different from the normal [3].

The relation captured in (3) is intuitive, the local mean is shifted toward the region in which the majority of the points reside. Since the mean shift vector is aligned with the local gradient estimate it can define a trajectory leading to a stationary point of the estimated density. Local maxima of the underlying density, i.e., the modes, are such stationary points. The *mean shift procedure* is obtained by successive

- computation of the mean shift vector $\mathbf{m}_K(\mathbf{x})$,
- translation of the kernel $K(\mathbf{x})$ by $\mathbf{m}_K(\mathbf{x})$,

and is guaranteed to converge at a nearby point where the density estimate has zero gradient [3].

3.2 Data Decomposition

Let us denote by $\left\{\mathbf{y}^j\right\}_{j=1,2\ldots}$ the sequence of successive locations of the kernel K, where

$$\mathbf{y}^{j+1} = \frac{\sum_{i=1}^{n} \mathbf{x}_i K\left(\frac{\mathbf{y}^j-\mathbf{x}_i}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{\mathbf{y}^j-\mathbf{x}_i}{h}\right)}, \quad j = 1, 2 \ldots \tag{4}$$

is the weighted mean at $\mathbf{y}^j$ computed with kernel K and $\mathbf{y}^1$ is the center of the initial kernel.

By running the mean shift procedure for all input data, each point $\mathbf{x}_i$, $i = 1, \ldots, n$ becomes associated to a point of convergence denoted by $\mathbf{y}_i$ where the underlying density has zero gradient. A test for local maximum is therefore needed. This test can involve a check on the eigenvalues of the Hessian matrix

of second derivatives, or a check for the stability of the convergence point. The latter property can be tested by perturbing the convergence point by a random vector of small norm, and letting the mean shift procedure to converge again. Should the convergence point be the same, the point is a local maximum.

Depending on the local structure of the density hypersurface, the convergence points can form ridges or plateaus. Therefore, the mean shift procedure should be followed be a simple clustering which links together the convergence points that are sufficiently close to each other. The algorithm is given below [3].

Mean Shift Based Decomposition

1. For each $i = 1, \ldots, n$ run the mean shift procedure for $\mathbf{x}_i$ and store the convergence point in $\mathbf{y}_i$.
2. Identify clusters $\{\mathbf{B}_u\}_{u=1\ldots m}$ of convergence points by linking together all $\mathbf{y}_i$ which are closer than h from each other.
3. For each $u = 1 \ldots m$ join together in cluster $\mathbf{D}_u$ all the data points $\mathbf{x}_i$ having the corresponding convergence point in $\mathbf{B}_u$.

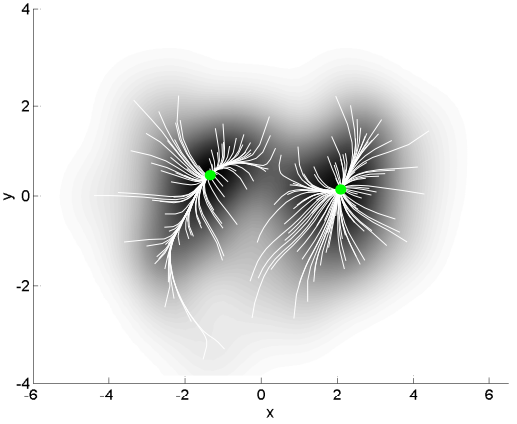

Fig. 2. Decomposition example. The mean shift procedure with $h = 0.6$ is applied for the data set presented in Figure 1a. The trajectory of each point is shown together with the two modes superimposed on the density surface. The view angle is from above.

A decomposition example is shown in Figure 2. Observe that the mean shift trajectories are smooth, verifying a property remarked in [3], that the cosine of the angle between two consecutive mean shift vectors is strictly positive when a normal kernel is employed.

The advantage this type of decomposition is twofold. First is requires a weak assumption about the underlying data structure, namely, that a probability density can be estimated nonparametrically. In addition, the method scales well with the space dimension, since the mean shift vector is computed directly from the data.

4 Saddle Point Detection

This section presents an algorithm for finding the saddle points associated with a given bandwidth h and a partition $\{\mathbf{D}_u\}_{u=1\ldots m}$ obtained through mean shift decomposition. We are interested in the detection of saddle points of first order, having the Hessian matrix with one positive eigenvalue and all other eigenvalues negative.

Select a cluster index v and define the complementary cluster set

$$\mathbf{C}_v \equiv \bigcup_{u \neq v} \mathbf{D}_u. \tag{5}$$

In the following we will drop the index v for the simplicity of the equations. We define two functions

$$\hat{f}_{D,K}(\mathbf{x}) = \frac{1}{nh^d} \sum_{\mathbf{x}_D \in D} K\left(\frac{\mathbf{x} - \mathbf{x}_D}{h}\right) \tag{6}$$

and

$$\hat{f}_{C,K}(\mathbf{x}) = \frac{1}{nh^d} \sum_{\mathbf{x}_C \in C} K\left(\frac{\mathbf{x} - \mathbf{x}_C}{h}\right) \tag{7}$$

whose superposition at $\mathbf{x}$ equals the density estimate at $\mathbf{x}$

$$\hat{f}_K(\mathbf{x}) \equiv \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{\mathbf{x} - \mathbf{x}_i}{h}\right) = \hat{f}_{D,K}(\mathbf{x}) + \hat{f}_{C,K}(\mathbf{x}). \tag{8}$$

Computing now the gradient of expression (8), multiplying by h^2, and normalizing by $\hat{f}_K$ it results that

$$\mathbf{m}_K(\mathbf{x}) = h^2 \frac{\hat{\nabla} f_K(\mathbf{x})}{\hat{f}_K(\mathbf{x})} = \alpha_D(\mathbf{x}) \mathbf{m}_{D,K}(\mathbf{x}) + \alpha_C(\mathbf{x}) \mathbf{m}_{C,K}(\mathbf{x}) \tag{9}$$

where

$$\mathbf{m}_{D,K}(\mathbf{x}) = \frac{\sum_{\mathbf{x}_D \in D} \mathbf{x}_D K\left(\frac{\mathbf{x} - \mathbf{x}_D}{h}\right)}{\sum_{\mathbf{x}_D \in D} K\left(\frac{\mathbf{x} - \mathbf{x}_D}{h}\right)} - \mathbf{x} \tag{10}$$

$$\mathbf{m}_{C,K}(\mathbf{x}) = \frac{\sum_{\mathbf{x}_C \in C} \mathbf{x}_C K\left(\frac{\mathbf{x} - \mathbf{x}_C}{h}\right)}{\sum_{\mathbf{x}_C \in C} K\left(\frac{\mathbf{x} - \mathbf{x}_C}{h}\right)} - \mathbf{x} \tag{11}$$

are the mean shift vectors computed only within the sets $\mathbf{D}$ and $\mathbf{C}$ respectively, and

$$\alpha_D(\mathbf{x}) = \frac{\hat{f}_{D,K}(\mathbf{x})}{\hat{f}_K(\mathbf{x})} \qquad \alpha_C(\mathbf{x}) = \frac{\hat{f}_{C,K}(\mathbf{x})}{\hat{f}_K(\mathbf{x})} \tag{12}$$

with $\alpha_D(\mathbf{x}) + \alpha_C(\mathbf{x}) = 1$. Equation (9) shows that the mean shift vector at any point $\mathbf{x}$ is a weighted sum of the mean shift vectors computed separately for the points in the sets $\mathbf{D}$ and $\mathbf{C}$.

Our idea is to exploit this property for the finding of saddle points. Let us assume that $\mathbf{x}_s$ is a saddle point of first order located on the boundary between **D** and **C**. The boundary condition is

$$\mathbf{m}_K(\mathbf{x}_s) = \mathbf{0} \tag{13}$$

which means that the vectors $\alpha_D(\mathbf{x}_s)\mathbf{m}_{D,K}(\mathbf{x}_s)$ and $\alpha_C(\mathbf{x}_s)\mathbf{m}_{C,K}(\mathbf{x}_s)$ have equal magnitude, are collinear, but point towards opposite directions. This explains the instability of $\mathbf{x}_s$. Indeed, a slight perturbation of $\mathbf{x}_s$ towards **C** and along the line defined by $\alpha_D(\mathbf{x}_s)\mathbf{m}_{D,K}(\mathbf{x}_s)$ and $\alpha_C(\mathbf{x}_s)\mathbf{m}_{C,K}(\mathbf{x}_s)$ will induce an increase in the magnitude of $\alpha_C(\mathbf{x}_s)\mathbf{m}_{C,K}(\mathbf{x}_s)$, which will move more $\mathbf{x}_s$ towards **C**, and so on. The same effect stands when $\mathbf{x}_s$ is perturbed towards **D**. Note that the process is one dimensional, no matter the dimensionality d of the space.

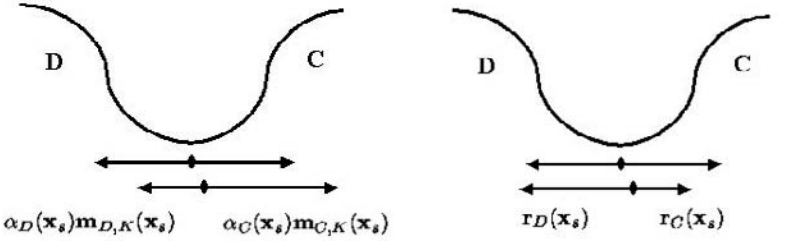

Fig. 3. Solving first order saddle point instability. (a) A slight perturbation of $\mathbf{x}_s$ towards C along the line defined by $\alpha_D(\mathbf{x}_s)\mathbf{m}_{D,K}(\mathbf{x}_s)$ and $\alpha_C(\mathbf{x}_s)\mathbf{m}_{C,K}(\mathbf{x}_s)$ will determine the point $\mathbf{x}_s$ to start moving towards C. (b) By employing the new vectors $\mathbf{r}_D(\mathbf{x}_s)$ and $\mathbf{r}_C(\mathbf{x}_S)$ the saddle point becomes stable and has a basin of attraction.

A simple modification, however can make a saddle point stable. Let us define the vectors

$$\mathbf{r}_D(\mathbf{x}) = \frac{\|\alpha_C(\mathbf{x})\mathbf{m}_{C,K}(\mathbf{x})\|}{\|\alpha_D(\mathbf{x})\mathbf{m}_{D,K}(\mathbf{x})\|}\alpha_D(\mathbf{x})\mathbf{m}_{D,K}(\mathbf{x}) \tag{14}$$

and

$$\mathbf{r}_C(\mathbf{x}) = \frac{\|\alpha_D(\mathbf{x})\mathbf{m}_{D,K}(\mathbf{x})\|}{\|\alpha_C(\mathbf{x})\mathbf{m}_{C,K}(\mathbf{x})\|}\alpha_C(\mathbf{x})\mathbf{m}_{C,K}(\mathbf{x}) \tag{15}$$

obtained by switching the norms of $\alpha_D(\mathbf{x}_s)\mathbf{m}_{D,K}(\mathbf{x}_s)$ and $\alpha_C(\mathbf{x}_s)\mathbf{m}_{C,K}(\mathbf{x}_s)$. This time, in the case of a perturbation, the resultant

$$\mathbf{r}(\mathbf{x}) = \mathbf{r}_D(\mathbf{x}) + \mathbf{r}_C(\mathbf{x}) \tag{16}$$

will point towards the saddle point and not away from the saddle point. Since the saddle point is of first order, it will be also stable for the directions perpendicular to $\mathbf{r}(x)$ hence it will be a stable point with basin of attraction.

Our algorithm will use the newly defined basin of attraction to converge to the saddle point. The convergence proof will be given in a subsequent paper, the

flavor of the proof being similar to the mean shift convergence discussed in [4, 3].

However, while all the mean shift paths converge to a local stationary point, the saddle point detection algorithm should be started close to a valley, i.e., at locations having divergent mean shift vectors coming from the sets $\mathbf{D}$ and $\mathbf{C}$

$$\alpha_D(\mathbf{x})\alpha_C(\mathbf{x})\mathbf{m}_{D,K}(\mathbf{x})^\top \mathbf{m}_{C,K}(\mathbf{x}) < 0 \tag{17}$$

Since the data is already partitioned it is simple to search for points that verify condition (17). If one starts the search from a point in $\mathbf{D}$ just follow the mean shift path defined by $\mathbf{m}_{C,K}(\mathbf{x})$ till the condition (17) is satisfied. Nevertheless, if the cluster $\mathbf{D}$ is isolated, the function $\hat{f}_{C,K}(\mathbf{x})$ (7) will be close to zero for the data points belonging to $\mathbf{x} \in \mathbf{D}$ and can generate numerical instability. Therefore a threshold should be imposed on this function before computing $\mathbf{m}_{C,K}(\mathbf{x})$.

The overall algorithm for finding the saddle points lying on the border of $\mathbf{D}$ is given below.

Saddle Point Detection

Given a data partitioning into a cluster $\mathbf{D}$ and another set $\mathbf{C}$ containing the rest of the data points. For each $\mathbf{x}_D \in \mathbf{D}$, if the value of $\hat{f}_{C,K}(\mathbf{x}_D)$ (7) is larger than a threshold

1. Follow the mean shift path defined by $\mathbf{m}_{C,K}(\mathbf{x})$ (11) until the condition (17) is satisfied.
2. Follow the mean shift path defined by $\mathbf{r}(\mathbf{x})$ (16) until convergence.

Observe that similarly to the mean shift iterations, the saddle point iterations can stop at saddle point of order larger than one. Therefore, the convergence point should be tested, either using the Hessian, or through the stability method discussed in Section 3.2.

Also note that the algorithm from above uses multiple initializations in the search for the saddle points. Prior knowledge regarding the border of cluster $\mathbf{D}$ can be used to reduce the number of executions.

An example of saddle point finding is shown in Figure 4. The two stages of the algorithm are visible for some of the trajectories. When the second part of the algorithm is initialized, the trajectory can have a sharp turn, to return back towards the valley. Afterwards, however, the trajectory converges smoothly to a saddle point.

5 Significance Test for Cluster Validity

Denote by $\mathbf{x}_s$ the saddle point with the largest density lying of the border of a given cluster characterized by the mode $\mathbf{y}_m$. The point $\mathbf{x}_s$ represents the "weakest" point of the cluster border. It requires the least amount of probability mass which should be taken from the neighborhood of $\mathbf{y}_m$ and placed in the neighborhood of $\mathbf{x}_s$ such that the cluster mode disappears, as described in [22].

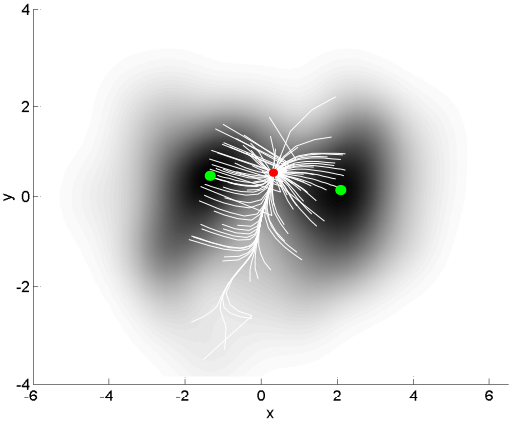

Fig. 4. Saddle point finding example. The algorithm was applied twice, for the data clustered as in figure Figure 2, once for the left cluster and once for the right cluster. The trajectories are shown together with the two modes and the detected saddle point superimposed on the density surface. The view angle is from above.

To characterize this process, we will assume in the following that the amount of probability mass in the neighborhood of the mode is proportional with $\hat{f}_K(\mathbf{y}_m)$, the probability density at the mode location, and the amount of probability mass in the neighborhood of the saddle point is proportional to $\hat{f}_K(\mathbf{x}_s)$, the density at $\mathbf{x}_s$. This approximation is shown in Figure 5, which presents a vertical slice in the density function.

Note that more evolved formulas can be derived based on the mean shift trajectory starting from the saddle point, however, for larger dimensions it is difficult to compute the exact amount of probability mass in a neighborhood.

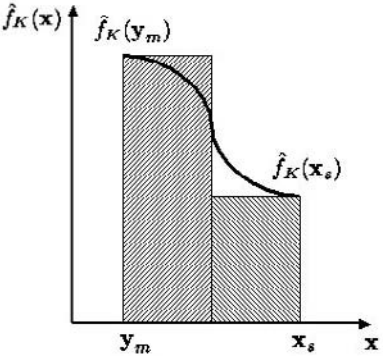

Fig. 5. Approximation of the probability mass. The probability mass in the neighborhood of $\mathbf{y}_m$ and $\mathbf{x}_s$ is assumed proportional to $\hat{f}_K(\mathbf{y}_m)$ and $\hat{f}_K(\mathbf{x}_s)$, respectively.

Using the approximation from above, we model the location of data points belonging to the cluster of cardinality n_c as a Bernoulli random variable which has a probability of

$$\hat{p} = \frac{\hat{f}_K(\mathbf{y}_m)}{\hat{f}_K(\mathbf{y}_m) + \hat{f}_K(\mathbf{x}_s)} \tag{18}$$

to lie in the mode neighborhood, and a probability of

$$\hat{q} = 1 - \hat{p} = \frac{\hat{f}_K(\mathbf{x}_s)}{\hat{f}_K(\mathbf{y}_m) + \hat{f}_K(\mathbf{x}_s)} \tag{19}$$

to lie in the saddle point neighborhood. Taking now into account the sampling properties of the estimator $\hat{p}$ (seen here as a random variable), the distribution of $\hat{p}$ can be approximated under weak conditions as normal, with mean and variance given by

$$\mu_p = p \qquad\qquad \sigma_p^2 = \frac{\hat{p}(1-\hat{p})}{n_c} \tag{20}$$

The null hypothesis which we test is the mode existence

$$H_0 : \ p \geq 0.5 \quad \text{versus} \quad H_1 : \ p < 0.5 \tag{21}$$

Hence, the test statistic is written as

$$z = \frac{\hat{p} - 0.5}{\sigma_p} \tag{22}$$

and using (18) and (20) we have

$$z = \frac{\sqrt{n_c}}{2} \frac{\hat{f}_K(\mathbf{y}_m) - \hat{f}_K(\mathbf{x}_s)}{\sqrt{\hat{f}_K(\mathbf{y}_m)\hat{f}_K(\mathbf{x}_s)}} \tag{23}$$

The p-value of the test is the probability that z, which is distributed with $N(0,1)$, is positive

$$\text{Prob}(z \geq 0) = \frac{1}{\sqrt{2\pi}} \int_{-z}^{\infty} \exp(-t^2/2)dt \tag{24}$$

A confidence of 0.95 is achieved when $z = 1.65$.

Using the framework from above, the clusters delineated with $h = 0.6$ shown in Figure 1b have a confidence of 0.99 and 0.98, respectively, derived using the mode densities of 0.0614 and 0.0598 and a saddle point density of 0.0384.

When $h = 0.9$ (Figure 1c) the two clusters have a confidence of 0.82 and 0.86, their mode densities are 0.0444 and 0.0460, while the saddle point density is 0.0369.

6 Clustering Experiments

Ideally, the input data should be analyzed for many different bandwidths and the confidence of each delineated cluster computed. This will guarantee the detection of significant clusters even when they exhibit different scales. An alternative method, less expensive is to choose one scale and join the least significant clusters until they become significant. We should, however, be cautious in joining too many clusters, because the approximation used in the computation of the p-value of the test assumes a certain balance between the peak and the saddle point neighborhood.

We applied the agglomerative strategy for the decomposition of the nonlinear structures presented in Figure 6a. A bandwidth $h = 0.15$ was employed. In the final configuration two modes and two saddle point were detected. The two clusters have a confidence equal to 1.00. The mode densities are 0.7744 and 0.8514 while the saddle densities are 0.2199 and 0.1957. The right saddle point has the largest density. Note that the density values are one order of magnitude larger than in the previous experiment. This is not a concern, since both coordinates were rescaled. Also, note that our test statistic (23) accepts the rescaling of the measured density.

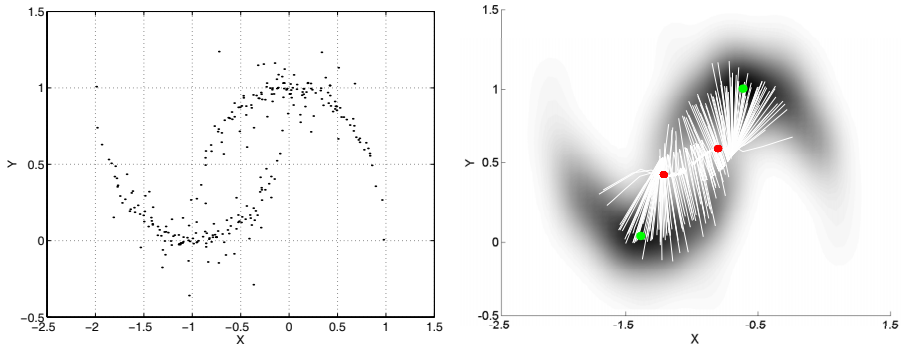

Fig. 6. Clustering of nonlinear structures. (a) Input data containing 270 data points. (b) The trajectories for saddle point detection are shown. Our algorithm detected two modes and two saddle points. The view angle is from above.

The next experiment was performed with $h = 0.6$ for the data shown in Figure 7a. Initially the algorithm detected four peaks that can be seen in the density surface shown in Figure 7b. However, the upper left peaks were joined together, their clusters having low statistical confidence (0.73 and 0.57). The cluster confidence for the final configuration are 0.921, 0.96 and 0.95. One can observe that the upper left cluster has the lowest confidence.

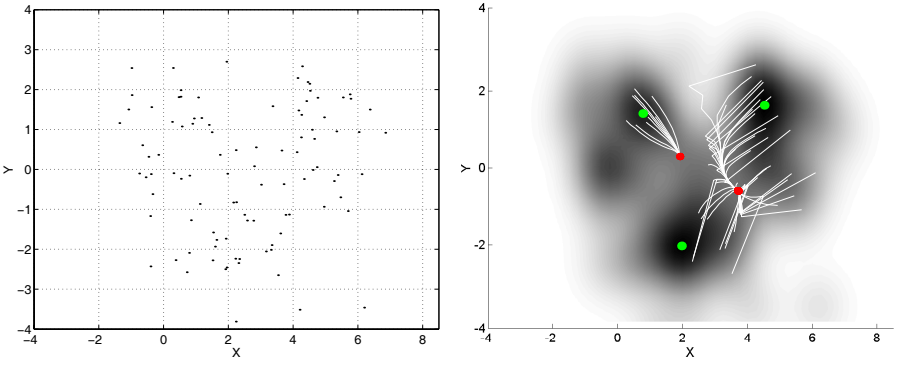

Fig. 7. Three clusters. (a) Input data containing 100 data points. (b) The trajectories for saddle point detection are shown. Our algorithm detected three modes with high confidence. Only the largest density saddle point for each cluster is shown. The view angle is from above.

7 Application to Image Segmentation

We apply the image segmentation framework described in [2], which employs mean shift to delineate clusters in a joint space of dimension $d = r + 2$ which includes the spatial coordinates. For the gray level case, $r = 1$, while for color images $r = 3$.

A clustering example for image-like data is shown in Figure 8. The 200 points have the x coordinate data points in increasing order (for each unit x coordinate there is one data point of variable y). A bandwidth of $h = 0.4$ was employed. The algorithm detected first 5 clusters of confidence 0.59, 0.79, 0.61, 0.99, and 0.99, which were reduced to 4 clusters of confidence 0.83, 0.61, 0.99, and 0.99, and finally to three clusters of confidence 1.00, 0.99, and 0.99. All the clusters that were merged belong to the elongated structure from the left.

In addition, we performed a Monte Carlo test to verify the stability of the decomposition. Out of 50 simulations only in 7 cases the number of clusters was different different from three. A confidence of 0.9 was used.

The same agglomerative algorithm was employed for the segmentation of two test images. Only clusters with confidence larger than 0.9 were retained. The contours of the decomposition are shown in Figure 9. The same bandwidth $h_r = 20$ was used for the color information, while for the spatial domain we used $h_s = 3$ and $h_s = 4$ (the second test image is 512 pixels, much larger than the first one). The segmentation exhibits high quality contours. Compare for example our result on *Woman* image with the segmentation in [31].

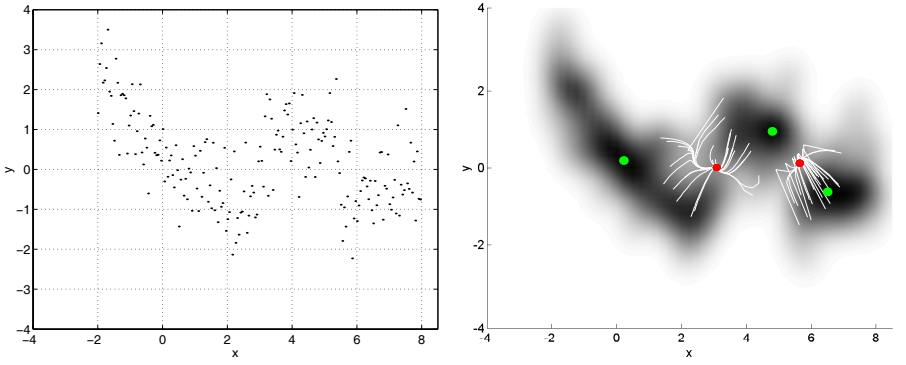

Fig. 8. Clustering of image-like data. (a) Input data containing 200 data points. (b) The trajectories for saddle point detection are shown. Our algorithms detected three modes with high confidence. Only the largest density saddle point for each cluster is shown. The view angle is from above.

Fig. 9. Clustering of image data. (a) Woman. (b) Lenna. Only clusters with confidence larger than 0.9 are selected.

8 Discussion

The results presented in this paper show that hypothesis testing for nonparametric clustering is a promising direction for solving decomposition problems and evaluating the significance of the results. Although our simulations are not comprehensive, we believe that the proposed algorithms are powerful tools for image data analysis. The natural way to continue this research is to investigate the

data in a multiscale approach and use our confidence measure to select clusters across scales.

We recently discovered that the problem of finding the saddle points of a multivariate surface appears in condensed matter physics and theoretical chemistry [13]. The computation of the energy barrier for the atomic transitions from one stable configuration to another requires the detection of the saddle point of the potential energy surface corresponding to a maximum along a minimum energy path. Numerical algorithms for solving this problem were developed for the case when both the initial and final states of the transitions are known [15] or only the initial state of the transition is known [14]. Compared to these methods which perform constrained optimization on one surface, our technique exploits the clustering of the data points the guide the optimization relative to two surfaces whose superposition represents the initial surface.

References

1. J. Babaud, A. Witkin, M. Baudin, and R. Duda. Uniqueness of the gaussian kernel for scale-space filtering. *IEEE Trans. Pattern Anal. Machine Intell.*, 8(1):26–33, 1986.
2. D. Comaniciu and P. Meer. Mean shift analysis and applications. In *Proc. Int. Conf. Computer Vision,* Kerkyra, Greece, pages 1197–1203, September 1999.
3. D. Comaniciu and P. Meer. Mean shift: A robust approach toward feature space analysis. *IEEE Trans. Pattern Anal. Machine Intell.*, 24(5):To appear, 2002.
4. D. Comaniciu, V. Ramesh, and P. Meer. The variable bandwidth mean shift and data-driven scale selection. In *Proceedings International Conference on Computer Vision,* Vancouver, Canada, volume I, pages 438–445, July 2001.
5. R. Duda, P. Hart, and D. Stork. *Pattern Classification.* Wiley, second edition, 2001.
6. P. F. Felzenszwalb and D. P. Huttenlocher. Image segmentation using local variation. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition,* Santa Barbara, CA, pages 98–103, June 1998.
7. C. Fraley and A. Raftery. How many clusters? which clustering method? - answers via model-based cluster analysis. *Computer Journal*, 41:578–588, 1998.
8. K. Fukunaga. *Introduction to Statistical Pattern Recognition.* Academic Press, second edition, 1990.
9. K. Fukunaga and L. D. Hostetler. The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Trans. Information Theory*, 21:32–40, 1975.
10. Y. Gdalyahu, D. Weinshall, and M. Werman. Self organization in vision: Stochastic clustering for image segmentation, perceptual grouping, and image database organization. *IEEE Trans. Pattern Anal. Machine Intell.*, 23(10):1053–1074, 2001.
11. R. Gilmore. *Catastrophe Theory for Scientists and Engineers.* Dover, 1993.
12. J. Hartigan. Statistical theory in clustering. *Journal of Classification*, 2:63–76, 1985.
13. G. Henkelman, G. Johannesson, and H. Jonsson. Methods for finding saddle points and minimum energy paths. In S. Schwartz, editor, *Progress on Theoretical Chemistry and Physics*, pages 269–300. Kluwer, 2000.

14. G. Henkelman and H. Jonsson. A dimer method for finding saddle points on high dimensional potential surfaces using only first derivatives. *Journal of Chemical Physics*, 111:7010–7022, 1999.
15. G. Henkelman, B. Uberuaga, and H. Jonsson. A climbing image nudged elastic band method for finding saddle points and minimum energy paths. *Journal of Chemical Physics*, 113:9901–9904, 2000.
16. T. Hofmann and J. Buhmann. Pairwise data clustering by deterministic annealing. *IEEE Trans. Pattern Anal. Machine Intell.*, 19(1):1–13, 1997.
17. A. Jain, M. Murty, and P. Flyn. Data clustering: A review. *ACM Computing Surveys*, 31(3):264–323, 1999.
18. A. K. Jain and R. C. Dubes. *Algorithms for Clustering Data*. Prentice Hall, 1988.
19. L. Kauffman and P. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*. J. Wiley & Sons, 1990.
20. Y. Leung, J. Zhang, and Z. Xu. Clustering by scale-space filtering. *IEEE Trans. Pattern Anal. Machine Intell.*, 22(12):1396–1410, 2000.
21. G. Milligan and M. Cooper. An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, 50:159–179, 1985.
22. M. Minnotte. Nonparametric testing of the existence of modes. *The Annals of Statistics*, 25(4):1646–1660, 1997.
23. A. Moore. Very fast EM-based mixture model clustering using multiresolution kd-trees. In *Advances in Neural Information Processing Systems 11*. MIT Press, 1999.
24. E. J. Pauwels and G. Frederix. Finding salient regions in images. *Computer Vision and Image Understanding*, 75:73–85, 1999.
25. P. Perona and W. Freeman. A factorization approach to grouping. In *Proceedings 5th European Conference on Computer Vision,* Freiburg, Germany, pages 655–670, 1998.
26. S. J. Roberts. Parametric and non-parametric unsupervised cluster analysis. *Pattern Recog.*, 30:261–272, 1997.
27. J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Trans. Pattern Anal. Machine Intell.*, 22(8):888–905, 2000.
28. B. Silverman. Using kernel density estimates to investigate multimodality. *J. R. Statist. Soc. B*, 43(1):97–99, 1981.
29. R. Tibshirani, G. Walther, D. Botstein, and P. Brown. Cluster validation by prediction strength. Technical Report 21, Dept. of Statistics, Stanford University, 2001.
30. R. Tibshirani, G. Walther, and T. Hastie. Estimating the number of clusters in a data set via the gap statistic. Technical Report 208, Dept. of Statistics, Stanford University, 2000.
31. Z. Tu, S. Zhu, and H. Shum. Image segmentation by data driven markov chain monte carlo. In *Proceedings International Conference on Computer Vision,* Vancouver, Canada, volume II, pages 131–138, July 2001.
32. A. Yuille and T. Poggio. Scaling theorems for zero crossings. *IEEE Trans. Pattern Anal. Machine Intell.*, 8(1):15–25, 1986.