Pursuing a formalization of the clustering problem. Answers (and questions) via modal clustering

Giovanna Menardi

Department of Statistical Sciences
University of Padua

menardi@stat.unipd.it

28 November, 2014

Motivating example

How many clusters?

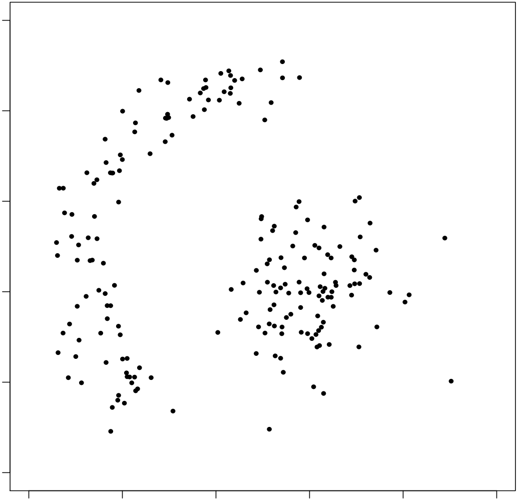


Figure : “Lump and banana” data (Stuetzle and Nugent, 2010).

Motivating example

How many clusters?

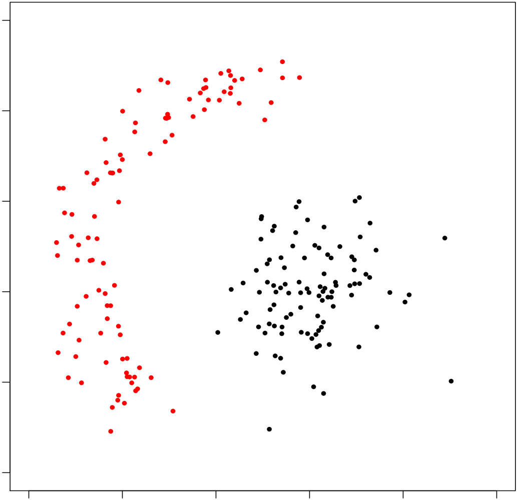


Figure : "Lump and banana" data (Stuetzle and Nugent, 2010).

Motivating example

How many clusters?

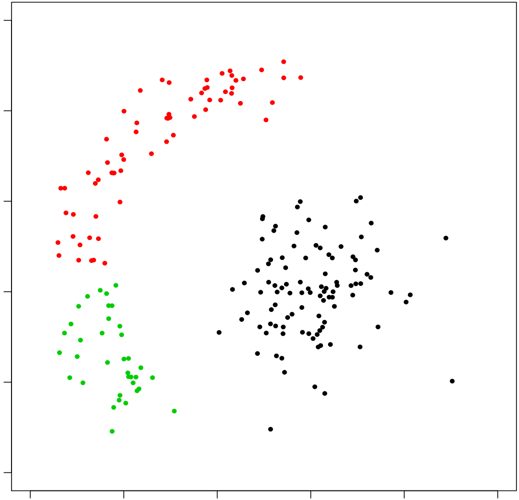


Figure : "Lump and banana" data (Stuetzle and Nugent, 2010).

How many clusters?

- The “true” number of clusters is not obvious even in simple examples
- Both intuition and authomatic methods to determine the optimal number of clusters give different answers
- There is no benchmark to assess the appropriateness of each answer
- How find the best answer without formulating the right question?

To answer the question “how many clusters there are?” we should first ask:

what is a cluster?

What is the underlying problem?

Clustering is an ill-posed-problem

- *For clustering there exists no ground truth* (von Luxburg and Ben-David, 2005)
- *The statistical properties of these methods are generally unknown, precluding the possibility of formal inference* (Fraley and Raftery, 2002)
- *The manner in which data 'should' be clustered depends on the desired resolution* (Domany, 1999)
- *Which [...] definition is appropriate depends on the meaning of the data and the aim of analysis* (Hennig, 2013)

Can we pose it better?

The clustering problem

- Going back to the definition of a statistical problem...
  - $\mathbf{X} = (x_1, \ldots, x_n)'$ sample of observations
  - x_i ,$i = 1, \ldots, n$, *i.i.d* realizations from $x \sim f : \mathcal{X} \subseteq \mathbb{R}^d \mapsto \mathbb{R}$
  - we are interested in some characteristic of f
  - based on $\mathbf{X}$ we make inference on f and, then, on its characteristics

- Why such a reluctance in doing the same in a clustering problem?

- We shall associate clusters to some specific characteristic of f :
  - parametric (model-based) approach:
    - $\mapsto$ clusters are homogeneous distributions combined in a mixture model
  - nonparametric (modal) approach:
    - $\mapsto$ clusters are the domains of attraction of the modes of f

Modal clustering

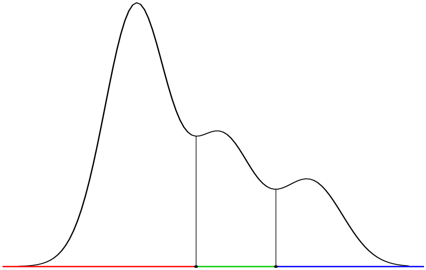
Attempting a formalization

- Clusters correspond to the domain of attraction of the modes of f
- Toward a formalization: Chacon (2013)
  - $d = 1$: set of points bounded by the local minima of f
  - $d > 1$: unstable manifolds of the negative gradient flow corresponding to the local maxima of f (Morse theory)

Modal clustering

Attempting a formalization

- Clusters correspond to the domain of attraction of the modes of f
- Toward a formalization: Chacon (2013)
  - $d = 1$: points at which f has a local maximum are archetypes of the clusters; boundaries are the local minima of f



Modal clustering

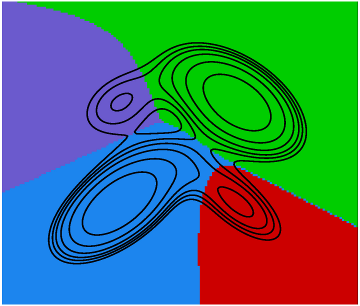
Attempting a formalization

- Clusters correspond to the domain of attraction of the modes of f
- Toward a formalization: Chacon (2013)
  - $d = 1$: points at which f has a local maximum are archetypes of the clusters; boundaries are the local minima of f
  - $d > 1$: unstable manifolds of the negative gradient flow corresponding to the local maxima of f (Morse theory)

Modal clustering

Attempting a formalization

- Clusters correspond to the domain of attraction of the modes of f
- Toward a formalization: Chacon (2013)
  - $d = 1$: points at which f has a local maximum are archetypes of the clusters; boundaries are the local minima of f
  - $d > 1$: unstable manifolds of the negative gradient flow corresponding to the local maxima of f (Morse theory)



Modal clustering

Attempting a formalization

- Clusters correspond to the domain of attraction of the modes of f
- Toward a formalization: Chacon (2013)
  - $d = 1$: points at which f has a local maximum are archetypes of the clusters; boundaries are the local minima of f
  - $d > 1$: unstable manifolds of the negative gradient flow corresponding to the local maxima of f (Morse theory)
- Formalization of these ideas for non-regular densities (non differentiable or densities with plateaux) is more complicated but still possible

Modal clustering

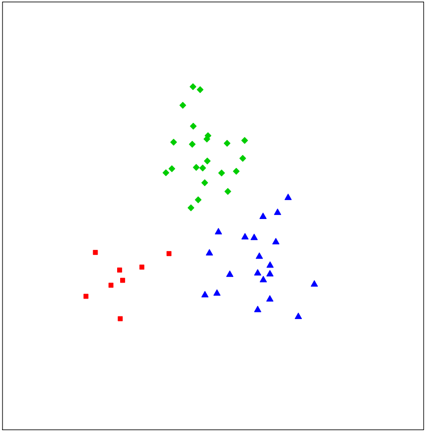
How to?

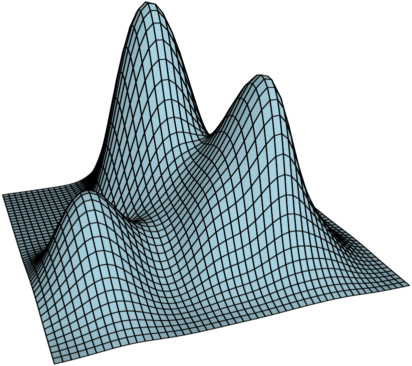
1. Bump hunting:
   $\mapsto$ explicit search of local maxima of the density estimate
   $\mapsto$ gradient ascent algorithms identify, for each observation, its uphill path toward the pertaining mode

   - EM based algorithm: Li, Ray and Lindsay (2007)
   - Mean-shift based algorithms: Cheng (1995), Comaniciu and Meer (2002), Chacon and Duong (2013)

Bumb hunting

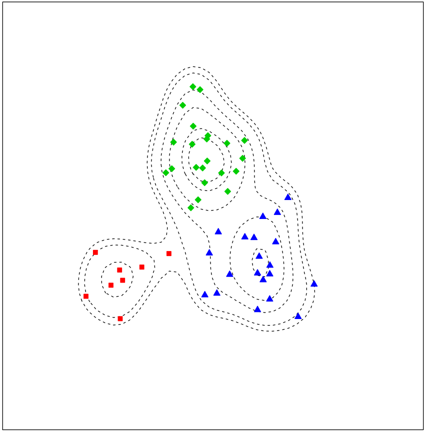
A toy example





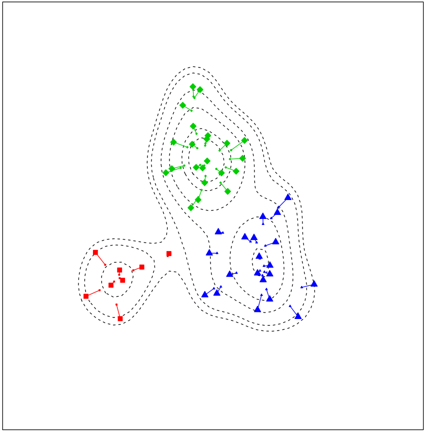
Bumb hunting

A toy example



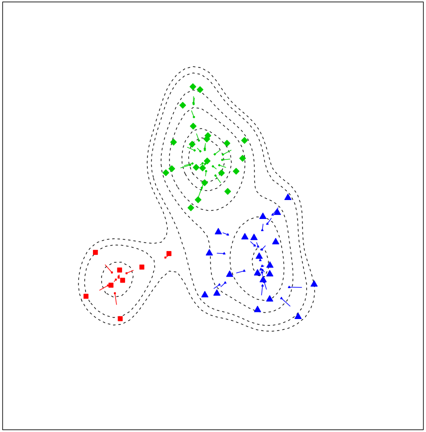
Bumb hunting

A toy example



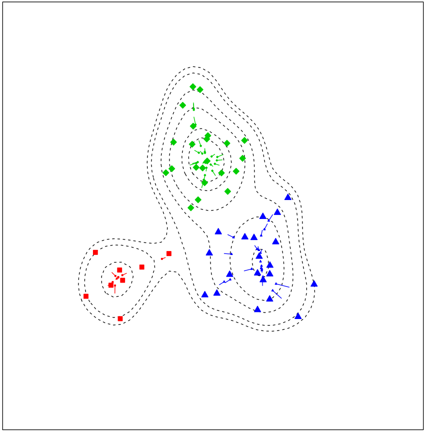
Bumb hunting

A toy example



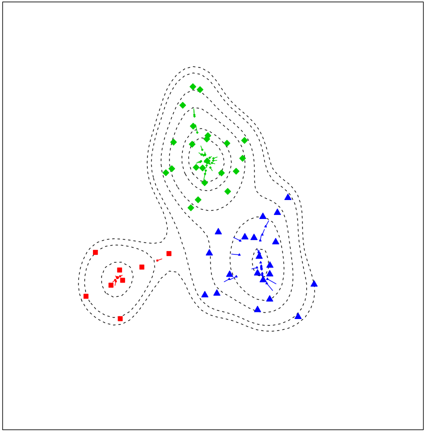
Bumb hunting

A toy example



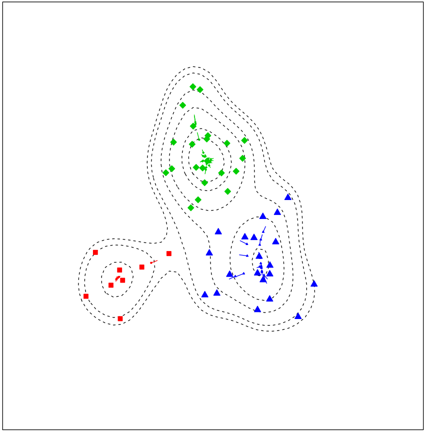
Bumb hunting

A toy example



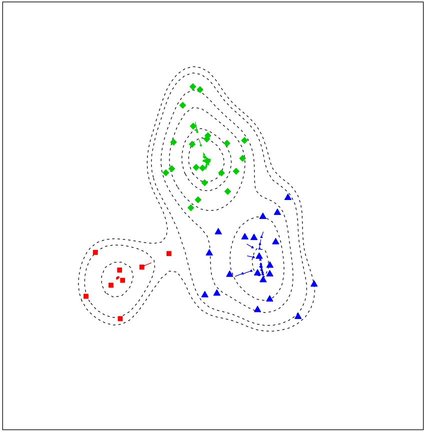
Bumb hunting

A toy example



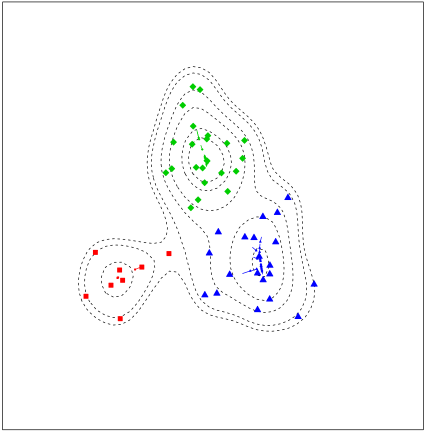
Bumb hunting

A toy example



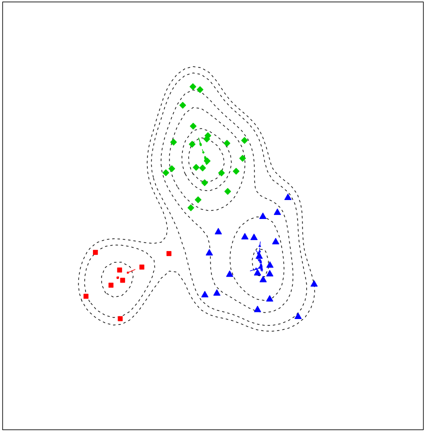
Bumb hunting

A toy example



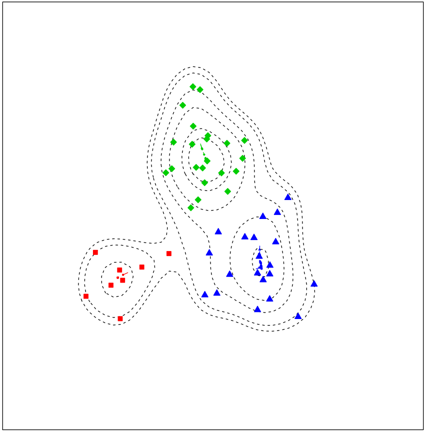
Bumb hunting

A toy example



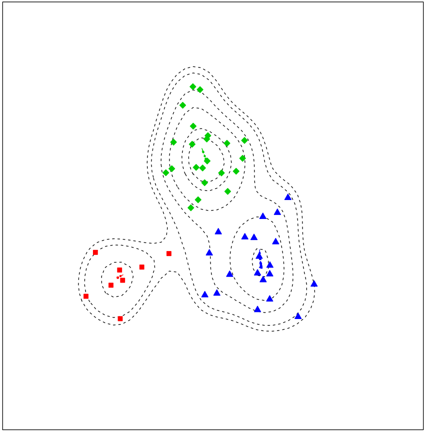
Bumb hunting

A toy example



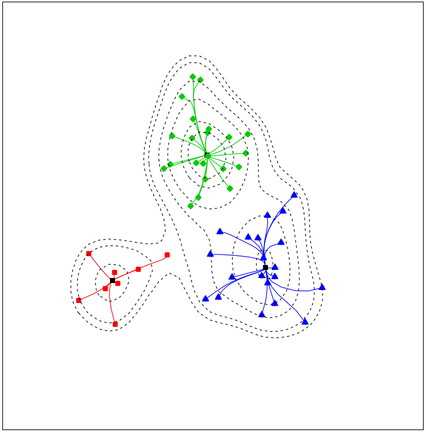
Bumb hunting

A toy example



Bumb hunting

A toy example



Modal clustering

How to?

2. Detection of connected components of the density level sets
   - ↦ for $0 \leq k \leq \max \ f$, define the level set $R(k)$ as:

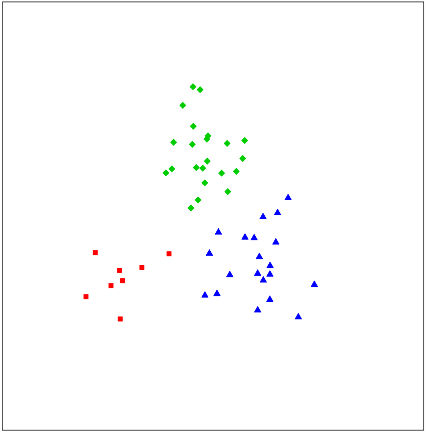
$$R(k) = \{x \in \mathbb{R}^d : f(x) \geq k\}$$

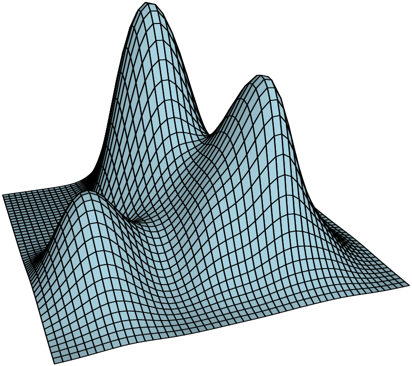
   - ↦ clusters correspond to the maximum connected components of $R(k)$
   - ↦ when k varies, the number of connected components of $R(k)$ varies and a hierarchical tree structure is generated.

   - methods mainly differ for the procedure to find the connected components: Stuetzle (2003), Azzalini and Torelli (2007), Stuetzle and Nugent (2010), Menardi and Azzalini (2014)

Level set connected component detection

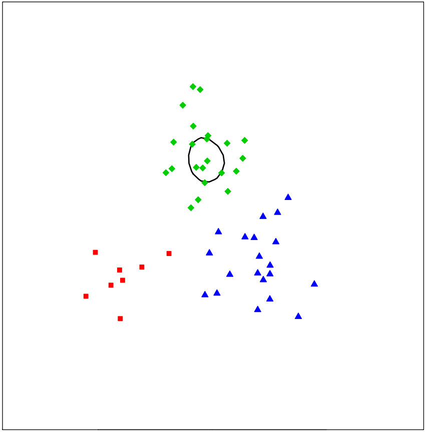
A toy example

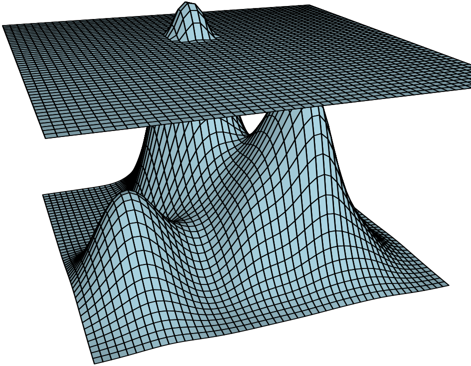




Level set connected component detection

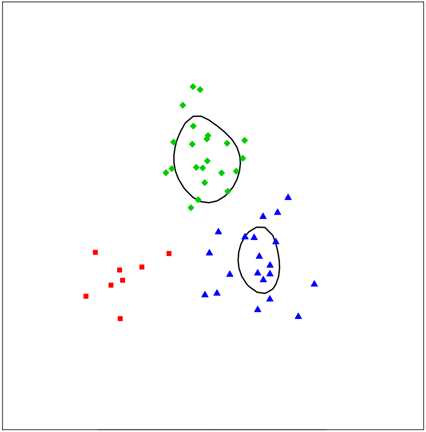
A toy example

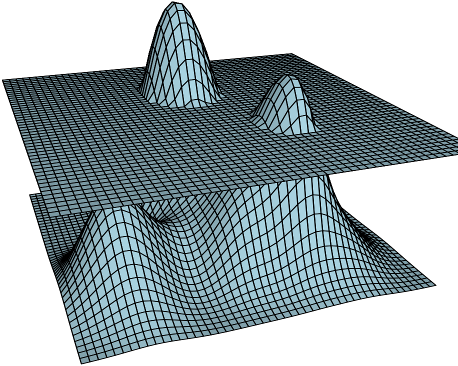




Level set connected component detection

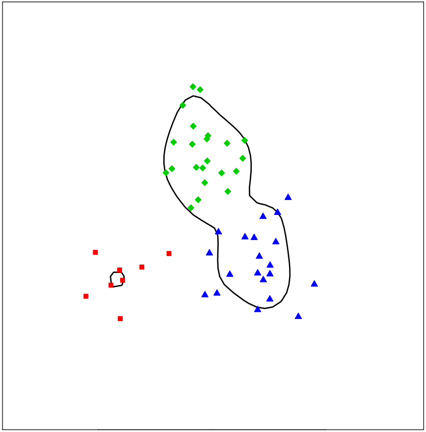
A toy example

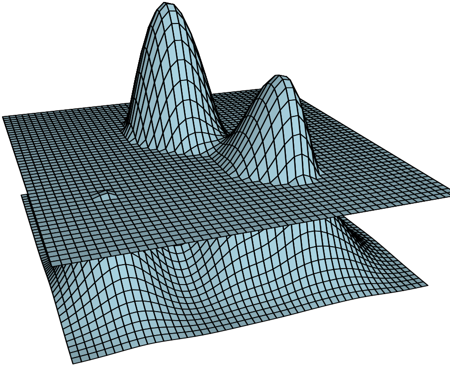




Level set connected component detection

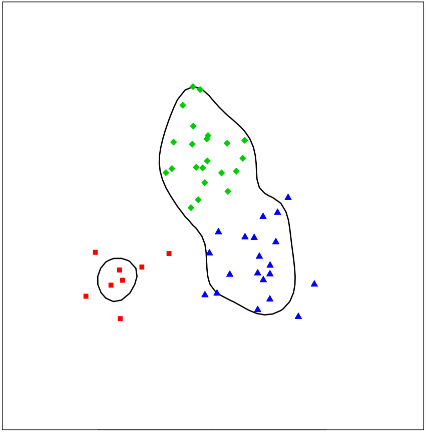
A toy example

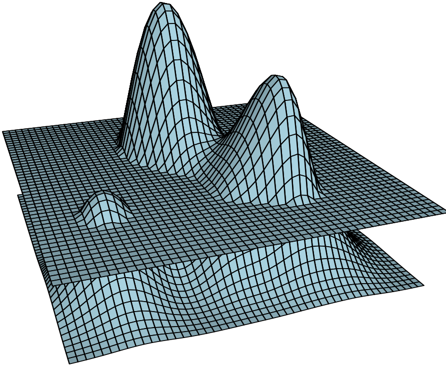




Level set connected component detection

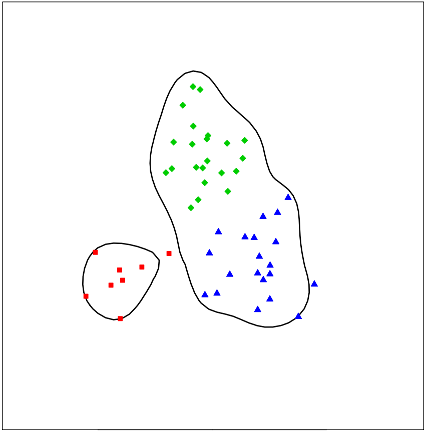
A toy example

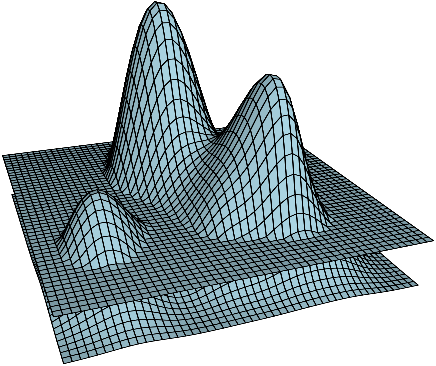




Level set connected component detection

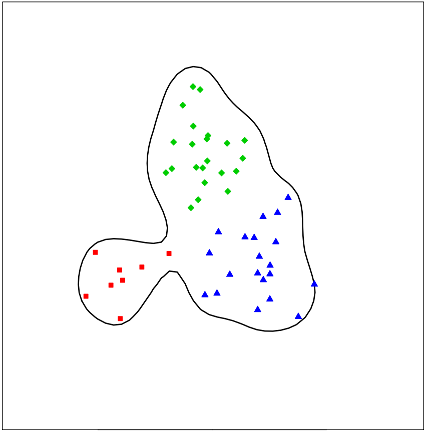
A toy example

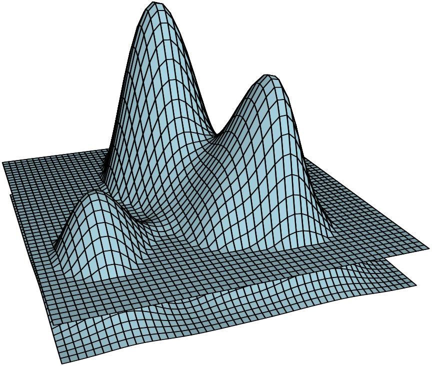




Level set connected component detection

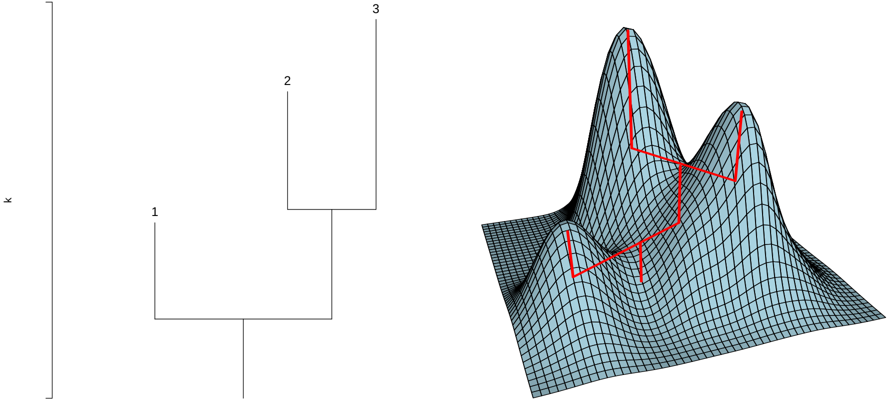
A toy example





Level set connected component detection

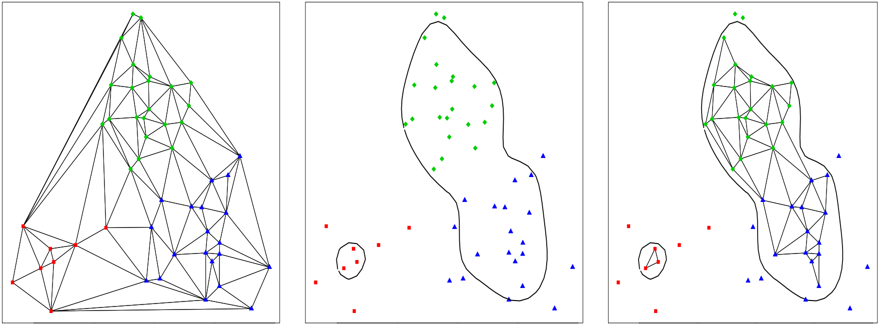
A toy example



Level set connected component detection

Graph-based connected components

- detection of the connected components of each level set is operationally performed by finding the connected components of a suitable graph built on tha data



Strengths of the approach

- Precise notion of cluster, associated with an intrinsic property of the data density
  - definition of a ground truth in the clustering task
  - number of clusters is conceptually defined
    - $\mapsto$ determining the number of clusters is a circumscribed problem of estimation
    - $\mapsto$ no detection of clusters (i.e. number of clusters equal to 1) is possible
    - $\mapsto$ the number of clusters is determined by the procedure
  - a probabilistic notion of cluster allows for providing each observation with a degree of confidence of belonging to the clusters
    - $\mapsto$ soft clustering schemes or cluster diagnostics
- Appealing notion of cluster
  - clusters are not bounded to have a particular shape
    - $\mapsto$ operationally: nonparametric density estimation allows to maintain this freedom
  - clusters ideally close to "natural groups" in data
  - the cluster tree naturally defines different levels of cluster resolution

Not all that glitters is gold

- Density estimation
  - governs the number and the shape of the clusters
- The nature of the data
  - categorical/mixed data are precluded
- Computational issues
  - actual implementation of the approach is often burdensome
- Conceptual questions
  - is the nonparametric approach always appropriate?

Density estimation

- Shape, number and composition of the clusters depend on the density estimate
- Use of nonparametric methods to allow for maximum flexibility; *e.g.* kernel estimator:

$$\hat{f}(x) = \sum_{i=1}^{n} \frac{1}{nh_1 \cdots h_d} \prod_{j=1}^{d} K\left(\frac{x^{(j)} - x_i^{(j)}}{h_j}\right),$$

  $x^{(j)}$, j-th component of x.
- Main concerns:
  - choice of the smoothing parameters
  - curse of dimensionality

Density estimation

Choice of the smoothing parameters

- The number of clusters is affected by the choice of the bandwidths
  - large bandwidths tend to oversmooth the density, possibly hiding some modes
  - small bandwidths tend to undersmooth the density, and favor the appearance of spurious modes
- How to choose the bandwidths?
  - critical in density estimation
  - less influential than expected in clustering
  - rule of thumb selections often work
  - clustering robust to a quite wide range of values (depending on cluster separation)

Density estimation

Choice of the smoothing parameters - Example (1)

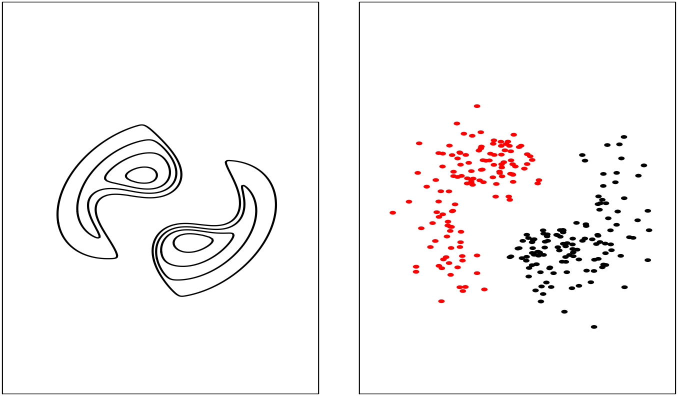


Figure : Density function (left) and associated (true) data clusters.

Density estimation

Choice of the smoothing parameters - Example (1)

Bandwidth: h_{NORM}
(normal reference rule: optimal for gaussian data)

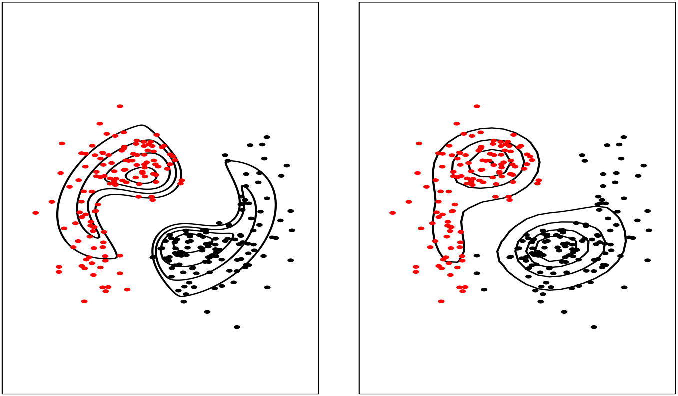


Figure : Density function and associated clusters: true (left) and estimated (right).

Density estimation

Choice of the smoothing parameters - Example (1)

Bandwidth: 0.9 $\times h_{NORM}$
(normal reference rule: optimal for gaussian data)

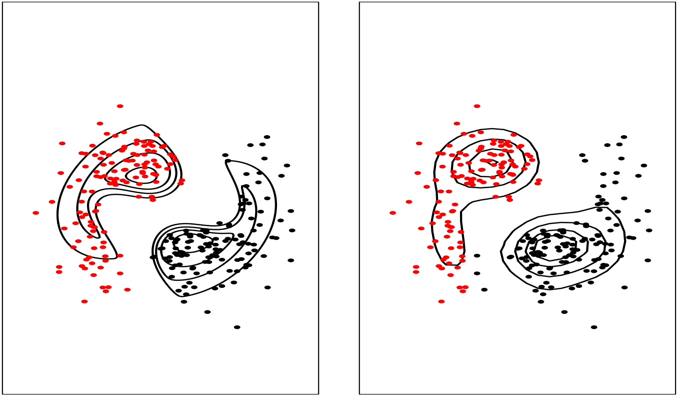


Figure : Density function and associated clusters: true (left) and estimated (right).

Density estimation

Choice of the smoothing parameters - Example (1)

Bandwidth: 0.8 $\times h_{NORM}$
(normal reference rule: optimal for gaussian data)

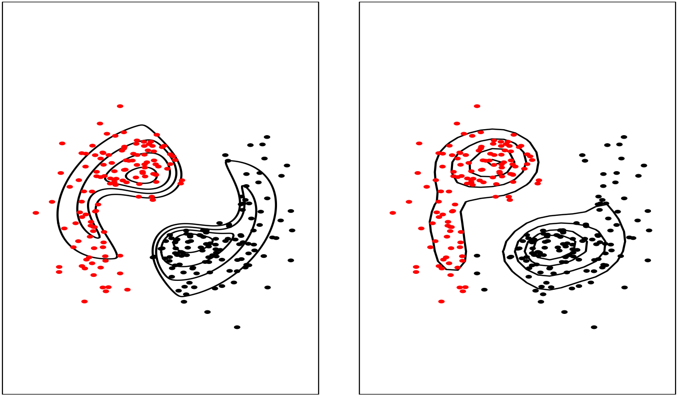


Figure : Density function and associated clusters: true (left) and estimated (right).

Density estimation

Choice of the smoothing parameters - Example (1)

Bandwidth: 0.7 $\times h_{NORM}$
(normal reference rule: optimal for gaussian data)

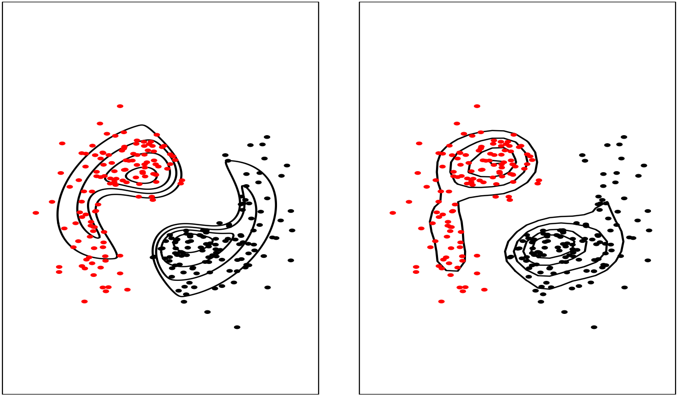


Figure : Density function and associated clusters: true (left) and estimated (right).

Density estimation

Choice of the smoothing parameters - Example (1)

Bandwidth: 0.6 $\times h_{NORM}$
(normal reference rule: optimal for gaussian data)

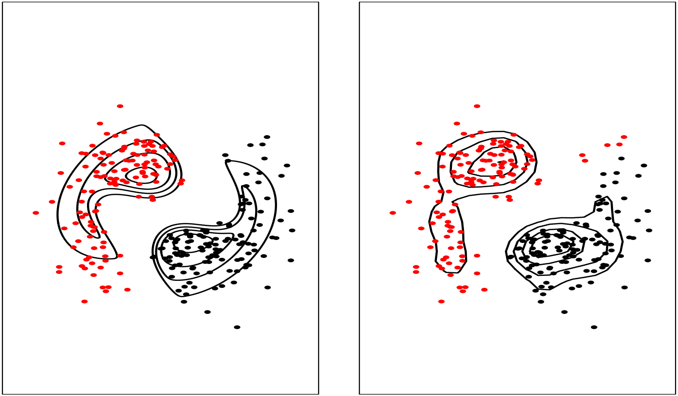


Figure : Density function and associated clusters: true (left) and estimated (right).

Density estimation

Choice of the smoothing parameters - Example (1)

Bandwidth: 0.5 $\times h_{NORM}$
(normal reference rule: optimal for gaussian data)

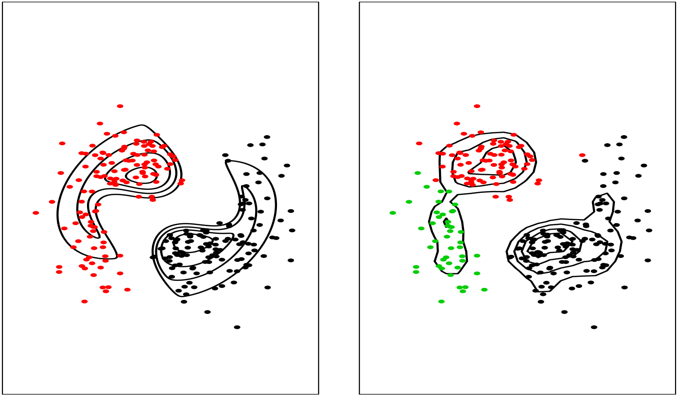


Figure : Density function and associated clusters: true (left) and estimated (right).

Density estimation

Choice of the smoothing parameters - Example (1)

Bandwidth: $1 \times h_{NORM}$
(normal reference rule: optimal for gaussian data)

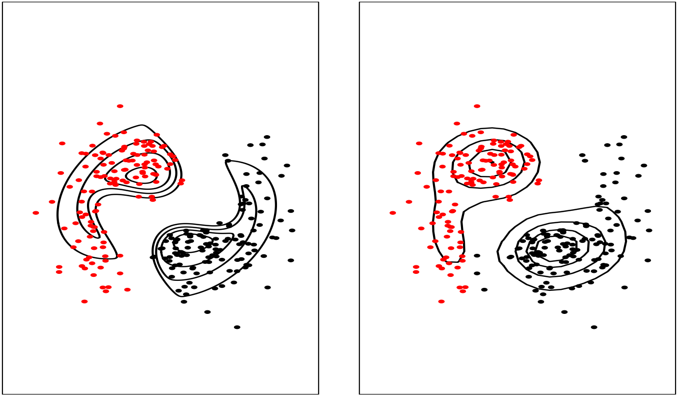


Figure : Density function and associated clusters: true (left) and estimated (right).

Density estimation

Choice of the smoothing parameters - Example (1)

Bandwidth: 1.1 $\times h_{NORM}$
(normal reference rule: optimal for gaussian data)

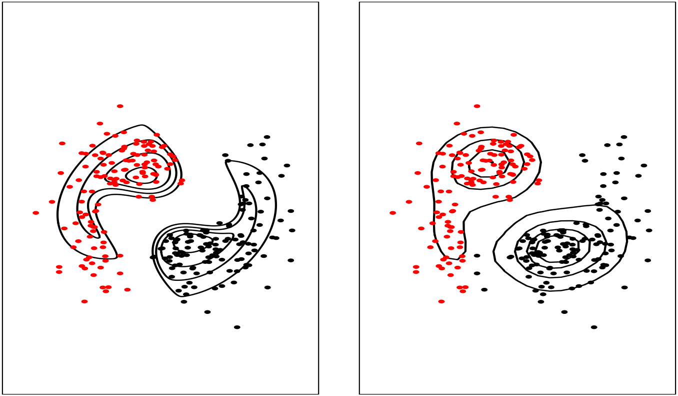


Figure : Density function and associated clusters: true (left) and estimated (right).

Density estimation

Choice of the smoothing parameters - Example (1)

Bandwidth: 1.2 $\times h_{NORM}$
(normal reference rule: optimal for gaussian data)

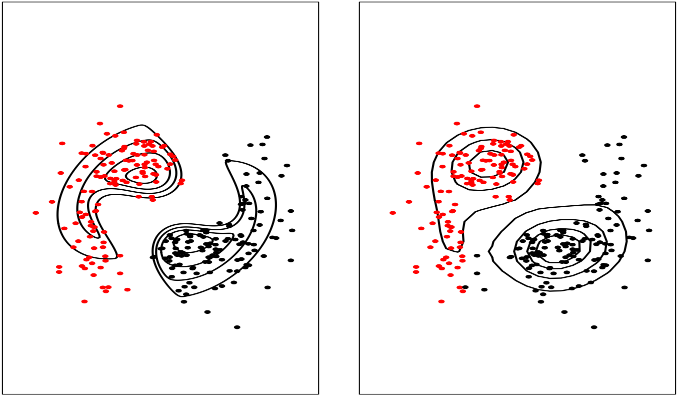


Figure : Density function and associated clusters: true (left) and estimated (right).

Density estimation

Choice of the smoothing parameters - Example (1)

Bandwidth: 1.3 $\times h_{NORM}$
(normal reference rule: optimal for gaussian data)

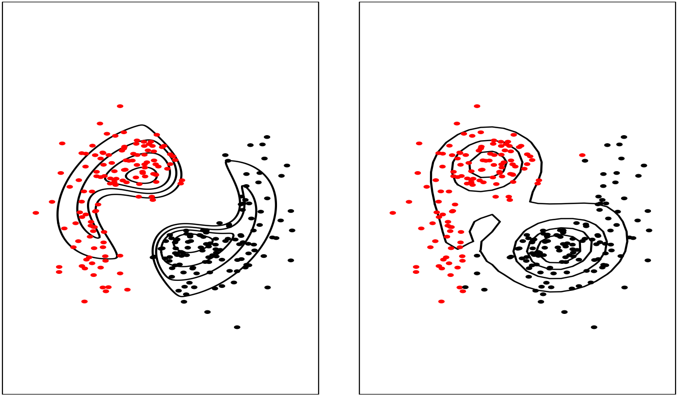


Figure : Density function and associated clusters: true (left) and estimated (right).

Density estimation

Choice of the smoothing parameters - Example (1)

Bandwidth: 1.4 $\times h_{NORM}$
(normal reference rule: optimal for gaussian data)

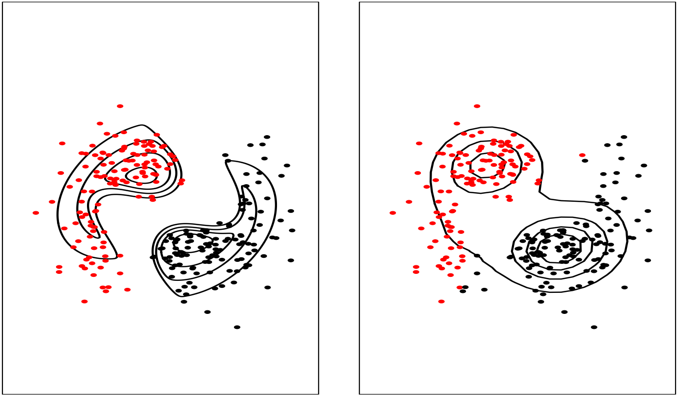


Figure : Density function and associated clusters: true (left) and estimated (right).

Density estimation

Choice of the smoothing parameters - Example (1)

Bandwidth: 1.5 $\times h_{NORM}$
(normal reference rule: optimal for gaussian data)

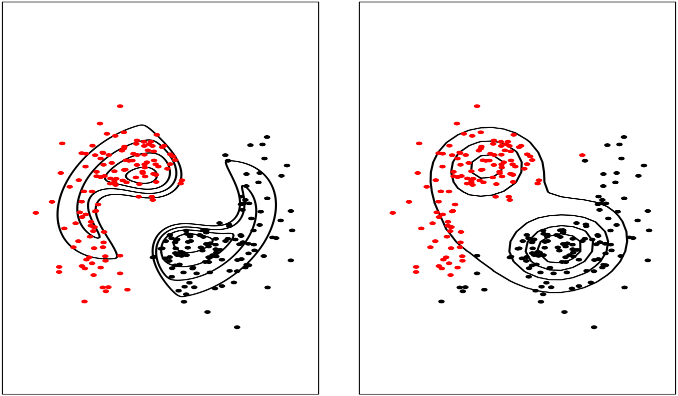


Figure : Density function and associated clusters: true (left) and estimated (right).

Density estimation

Choice of the smoothing parameters - Example (2)

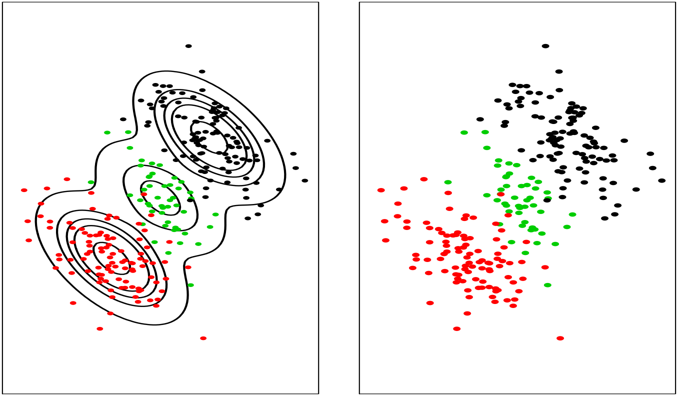


Figure : Density function (left) and associated (true) data clusters.

Density estimation

Choice of the smoothing parameters - Example (2)

Bandwidth: h_{NORM}
(normal reference rule: optimal for gaussian data)

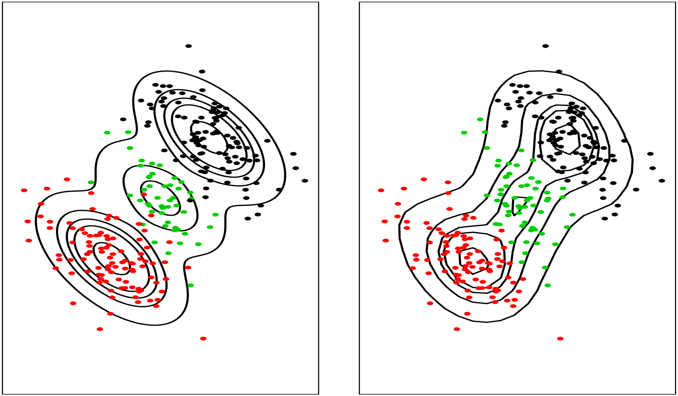


Figure : Density function and associated clusters: true (left) and estimated (right).

Density estimation

Choice of the smoothing parameters - Example (2)

Bandwidth: 0.9 $\times h_{NORM}$
(normal reference rule: optimal for gaussian data)

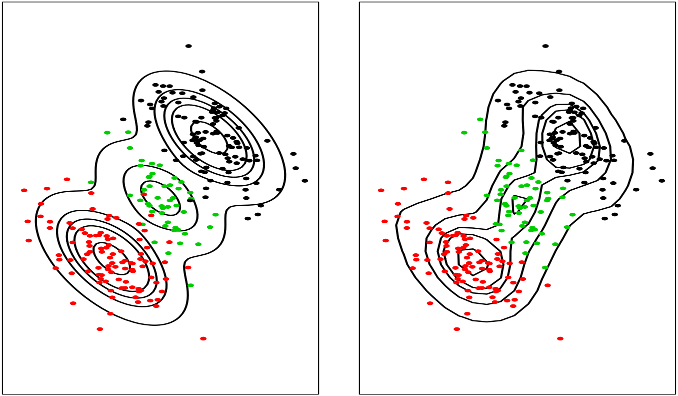


Figure : Density function and associated clusters: true (left) and estimated (right).

Density estimation

Choice of the smoothing parameters - Example (2)

Bandwidth: 0.8 $\times h_{NORM}$
(normal reference rule: optimal for gaussian data)

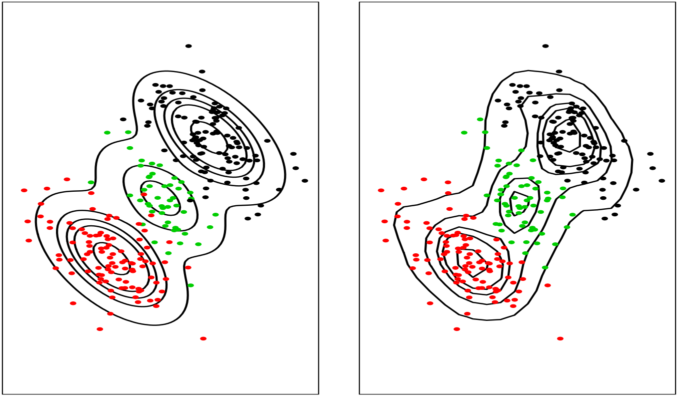


Figure : Density function and associated clusters: true (left) and estimated (right).

Density estimation

Choice of the smoothing parameters - Example (2)

Bandwidth: 0.7 $\times h_{NORM}$
(normal reference rule: optimal for gaussian data)

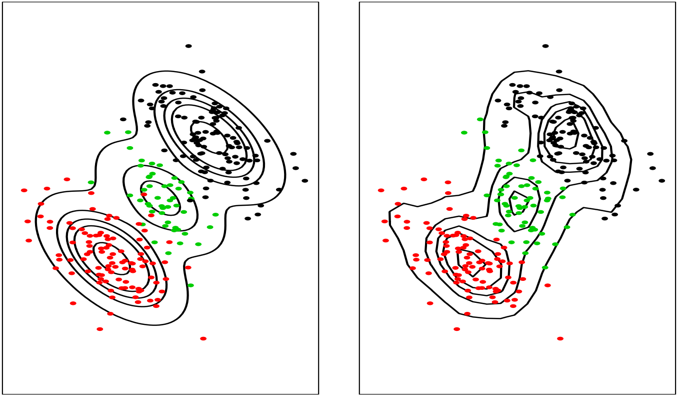


Figure : Density function and associated clusters: true (left) and estimated (right).

Density estimation

Choice of the smoothing parameters - Example (2)

Bandwidth: 0.6 $\times h_{NORM}$
(normal reference rule: optimal for gaussian data)

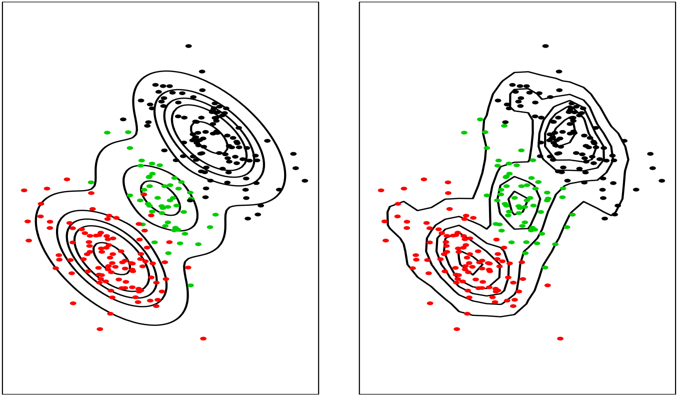


Figure : Density function and associated clusters: true (left) and estimated (right).

Density estimation

Choice of the smoothing parameters - Example (2)

Bandwidth: 0.5 $\times h_{NORM}$
(normal reference rule: optimal for gaussian data)

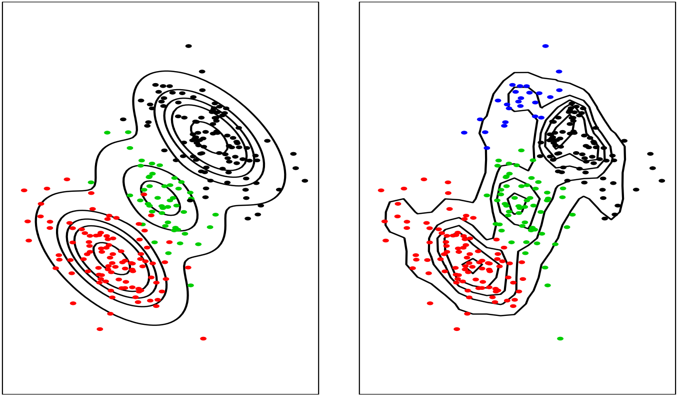


Figure : Density function and associated clusters: true (left) and estimated (right).

Density estimation

Choice of the smoothing parameters - Example (2)

Bandwidth: h_{NORM}
(normal reference rule: optimal for gaussian data)

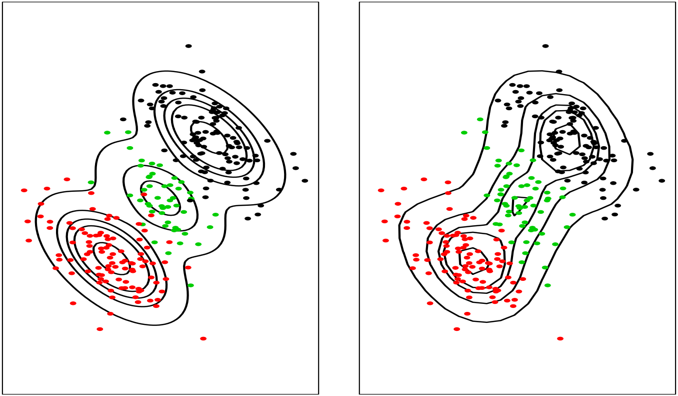


Figure : Density function and associated clusters: true (left) and estimated (right).

Density estimation

Choice of the smoothing parameters - Example (2)

Bandwidth: 1.1 $\times h_{NORM}$
(normal reference rule: optimal for gaussian data)

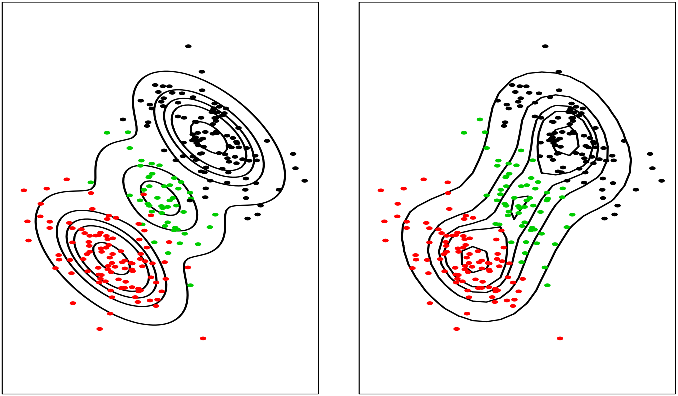


Figure : Density function and associated clusters: true (left) and estimated (right).

Density estimation

Curse of dimensionality

- Nonparametric density estimate degrades as the dimensionality increases
- The sparsity of data produces empty neighborhoods, especially in the low-density regions
  - birth of spurious clusters
- Modal clustering is jeopardized in high dimensions but for moderately high dimensions (tenths of variables):
  - kernel estimator can still reveal the modes for fairly separated clusters
    - ↦ oversmooth the density estimate
    - ↦ use of adaptive estimator
  - remedies to remove spurious clusters may help
    - ↦ Methods for pruning the cluster tree based on evaluation of mode "relevance"(Stuetzle, 2003; Li et al., 2007; Stuetzle and Nugent, 2010)
    - ↦ Evaluation of valley relevance based on introducing some tolerance parameter in graph building (Menardi and Azzalini, 2014)

The nature of the data

- Modal clustering hinges on the notions of probability density function and connected regions.
  - $\mapsto$ intrinsically designed for continuous data
- Real data are usually of mixed nature (categorical/numeric)

How to circumvent the assumption of continuity?

The nature of the data

Handling categorical data: a possible solution[1]

- Categorical data may be thought of as a simplified representation of some continuous latent variables
- A latent numerical configuration can be found by means of multidimensional scaling (MDS)
  - reflects the dissimilarities among points
  - shares the starting point of traditional clustering methods
- Numerical coordinates are then passed to the density-based clustering procedure.

- involves some arbitrariness
- but still has been showing fair results

[1] Joint work in progress with Adelchi Azzalini

The nature of the data

Handling categorical data: an example

- *Crabs data*: 200 observations describing 5 morphological measurements from each of two colour forms and both sexes of the species *Leptograpsus variegatus* collected at Fremantle, W. Australia.
- Clustering based on use of the 5 continuous variables only cannot reconstruct neither the species or the gender

		Blue F	Blue M	Orange F	Orange M
clusters	1	28	21	9	22
	2	22	29	41	28

- Reconstruction is achieved by merging the observed numerical variables with a bidimensional configuration detected by classic MDS

		Blue F	Blue M	Orange F	Orange M
clusters	1	46	5	5	1
	2	0	45	0	0
	3	4	0	45	0
	4	0	0	0	49

Computational issues

- Computational complexity is strongly algorithm-dependent
- Bumb hunting methods
  - main source of computation is the required iterative density estimation
- Connected components detection
  - main source of computation is in detection of connected components (burdensome in multidimensional spaces)
    - $\mapsto$ some methods require to establish if each pair of observations is connected by an edge in the associated graph $\rightarrow$ computational complexity grows quadratically with the sample size
    - $\mapsto$ other methods have computational complexity that is mildly affected by the sample size but grows exponentially with the dimensionality (Azzalini and Torelli, 2007)
- Reducing the computational complexity still remains an open problem, usually faced with ad hoc solutions

Conceptual questions

- Although often corresponding to a common-sense idea of group, the association between clusters and modes of a density function can be sometimes questioned
- Example (Hennig, 2010)

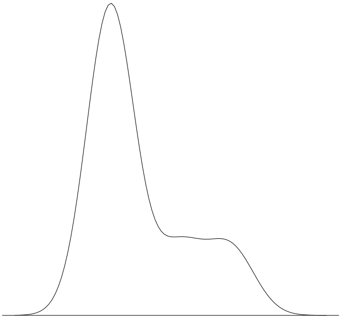


Figure : Unimodal distribution with evidence of more than one group

Conceptual questions

- Although often corresponding to a common-sense idea of group, the association between clusters and modes of a density function can be sometimes questioned
- Example (Hennig, 2010)

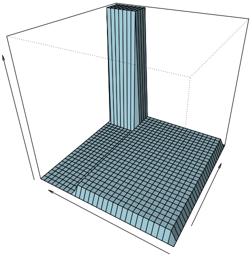


Figure : Unimodal distribution with evidence of more than one group

Conceptual questions

- Although often corresponding to a common-sense idea of group, the association between clusters and modes of a density function can be sometimes questioned
- Example (Hennig, 2010)

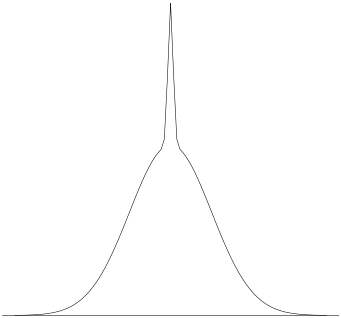


Figure : Unimodal distribution with evidence of more than one group

Conceptual questions

- Although often corresponding to a common-sense idea of group, the association between clusters and modes of a density function can be sometimes questioned
- Example (Stuetzle, 2003)

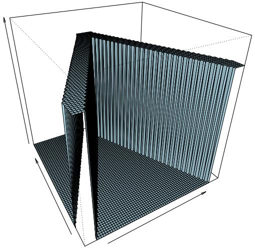


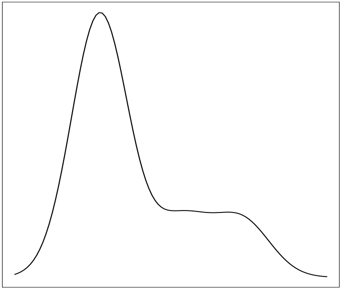
Figure : Unimodal distribution with evidence of more than one group

Conceptual questions

- Hand-made and very instable examples

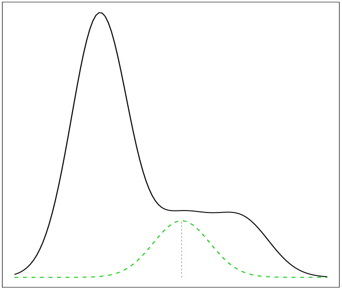
Conceptual questions

- Hand-made and very instable examples



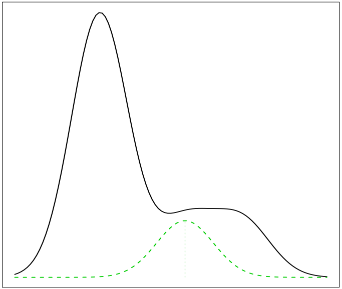
Conceptual questions

- Hand-made and very instable examples



Conceptual questions

- Hand-made and very instable examples



Conceptual questions

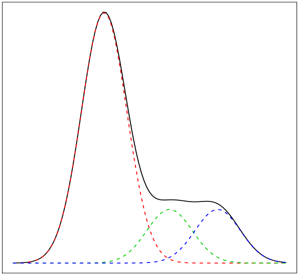
- Hand-made and very instable examples
- In order to say that two or more clusters exist we need to have in mind some concept of cluster $\rightarrow$ go back to the starting point: the clustering problem needs to be precisely defined

Conceptual questions

- Hand-made and very instable examples
- In order to say that two or more clusters exist we need to have in mind some concept of cluster $\rightarrow$ go back to the starting point: the clustering problem needs to be precisely defined
- For every precise notion of cluster, there exist some limit situations that cannot be caught

Conceptual questions

- Hand-made and very instable examples
- In order to say that two or more clusters exist we need to have in mind some concept of cluster $\rightarrow$ go back to the starting point: the clustering problem needs to be precisely defined
- For every precise notion of cluster, there exist some limit situations that cannot be caught



To sum up

Modal clustering

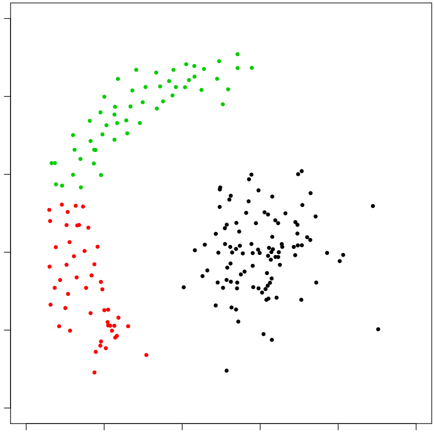
- not meant to be the definitive answer to the clustering problem
  - depending on clustering aim it may be suitable or not
  - knowledge of the phenomenon must be still the priority guide
  - human intervention is often unavoidable

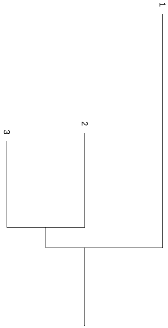
$\longrightarrow$

- but still a sound attempt to keep the arbitrariness low
  - natural clusters are a good-sense solution when aim of clustering is vague/unknown
  - knowledge of the phenomenon should guide human intervention
  - human intervention (usually) limited to some detailed aspects

Concluding example

(just for curiosity)





References

- Azzalini, A. and Torelli, N. Clustering via nonparametric density estimation. *Stat. Comp.*, 17, 2007
- Cheng, Y. Mean Shift, Mode Seeking, and Clustering. *IEEE Trans. Patt. An. Mach. Intell.*, 17, 1995.
- Chacn, J.,Clusters and water flows: a novel approach to modal clustering through morse theory. arXiv preprint arXiv:1212.1384, 2013.
- Chacn, J., Duong, T. Data-driven density derivative estimation, with applications to nonparametric clustering and bump hunting. *Electron. J. Stat.*, 7, 2013.
- Comaniciu, d, Meer, P. Mean Shift: A Robust Approach toward Feature Space Analysis, *IEEE Trans. Patt. An. Mach. Intell.*, 24, 2002
- Domany, E. Superparamagnetic clustering of data - The definitive solution of an ill-posed problem. *Physica, A: Stat. Mech. and its Appl.*, 263, 1999.
- Hennig, C. Methods for merging Gaussian mixture components. *Adv. Data An. & Clas.*, 2010.
- Hennig, C. How many bee species? A case study in determining the number of clusters. *Proc. GfKl-2012*, 2013.
- Li. J, Ray. S, Lindsay. B. G, A nonparametric statistical approach to clustering via mode identfication, *J. Mach. Learn. Res.*, 8, 2007.
- Fraley, C.,Raftery, A.E. Model-Based Clustering, Discriminant Analysis, and Density Estimation. *J. Am. Stat. Ass.*, 97, 2002.
- Menardi, G., Azzalini, A. An advancement in clustering via nonparametric density estimation. *Stat. Comp.*, 24, 2014.
- Stuetzle, W. Estimating the cluster tree of a density by analyzing the minimal spanning tree of a sample. *J. Classif.*, 20, 2003.
- Stuetzle, W. and Nugent, R. A generalized single linkage method for estimating the cluster tree of a density. *J. Comp. Graph. Stat.*, 19,2010.
- Von Luxburg, U., Ben-David, S. Towards a statistical theory for clustering. *PASCAL Workshop on Stat. and Optim. of Clustering*, 2005.