Nonparametric Clustering

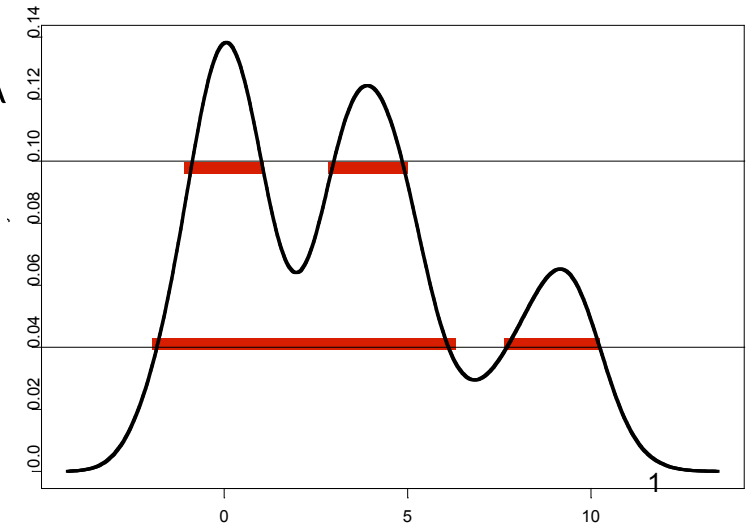
(Wishart 1969; Hartigan 1981; Wong & Lane 1983; Cuevas 2000, 2001)

- Assume a correspondence between groups and modes of the density $p(x)$
- Wishart: methods should "resolve distinct data modes, independently of their shape and variance"
- Define the *cluster tree* of a density as the fundamental quantity to be estimated by nonparametric cluster analysis

"High density clusters" (Hartigan 1975)

- Define a level set of a density $p(x)$ at level λ as the subset of feature space where the density exceeds λ: $L(\lambda; p) = \{x | p(x) > \lambda\}$
- Connected components of level sets have a hierarchical structure.

  For any two conn comp A and B:
  $A \subset B$, $B \subset A$, or $A \bigcap B = \emptyset$



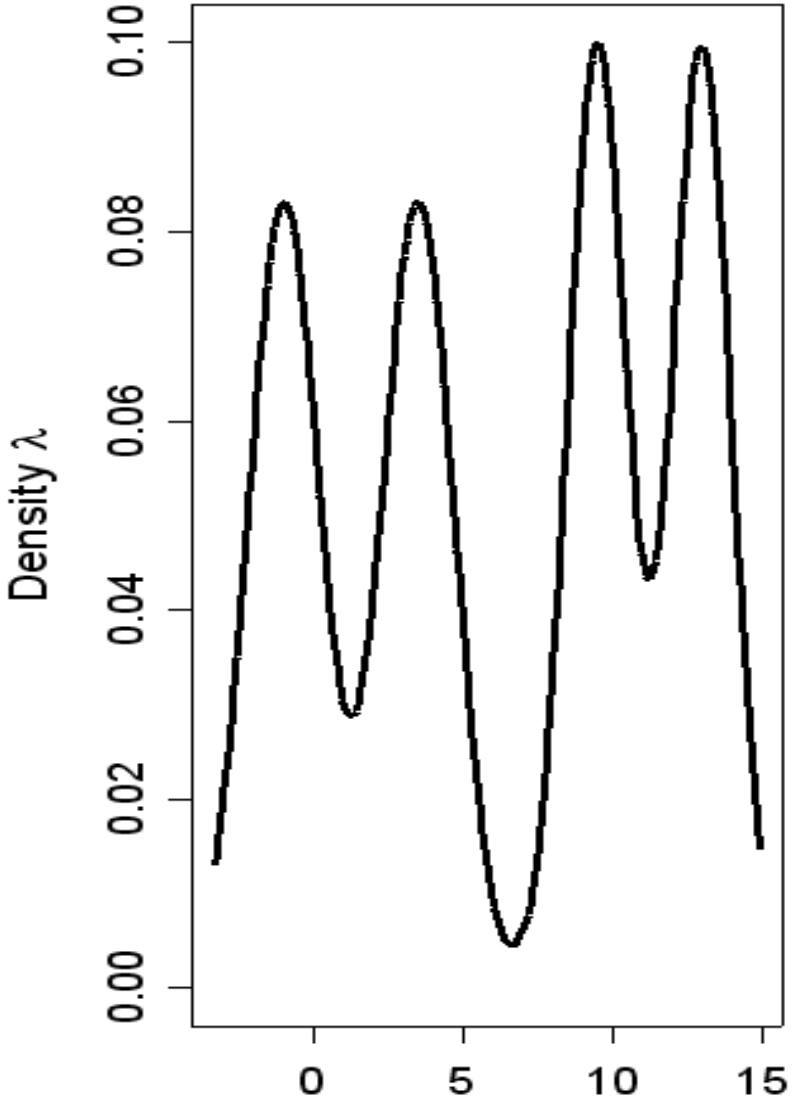
The Cluster Tree of a Density

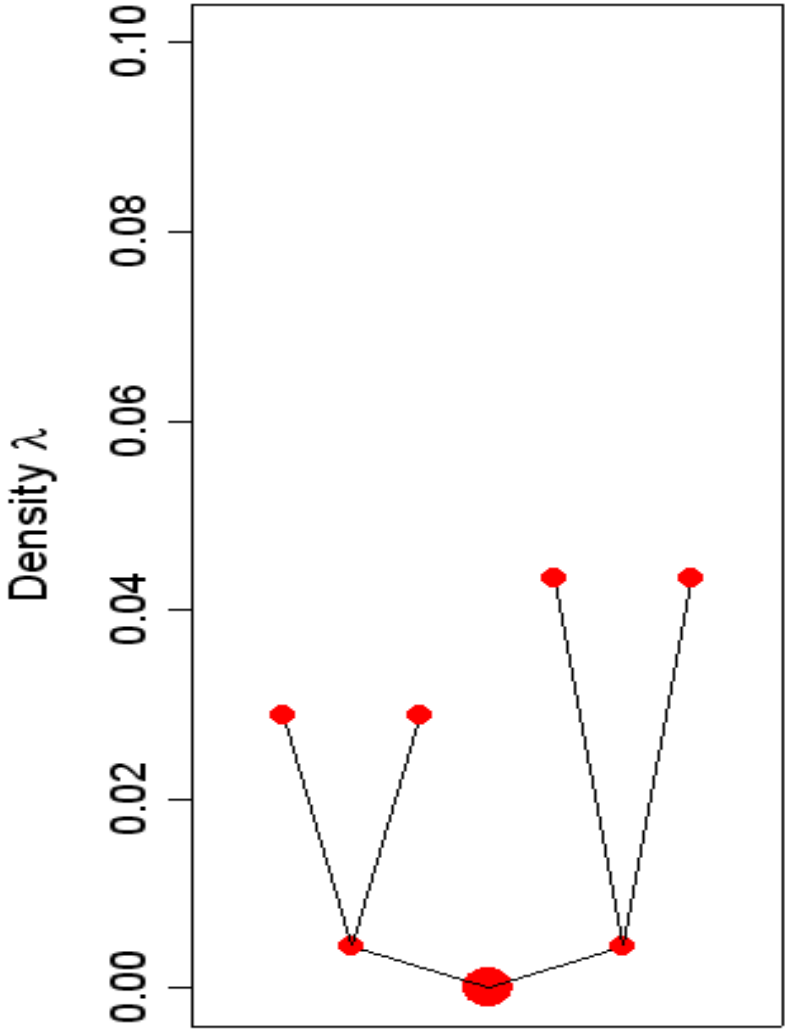
Formal Definition of Cluster Tree (Stuetzle 2003)

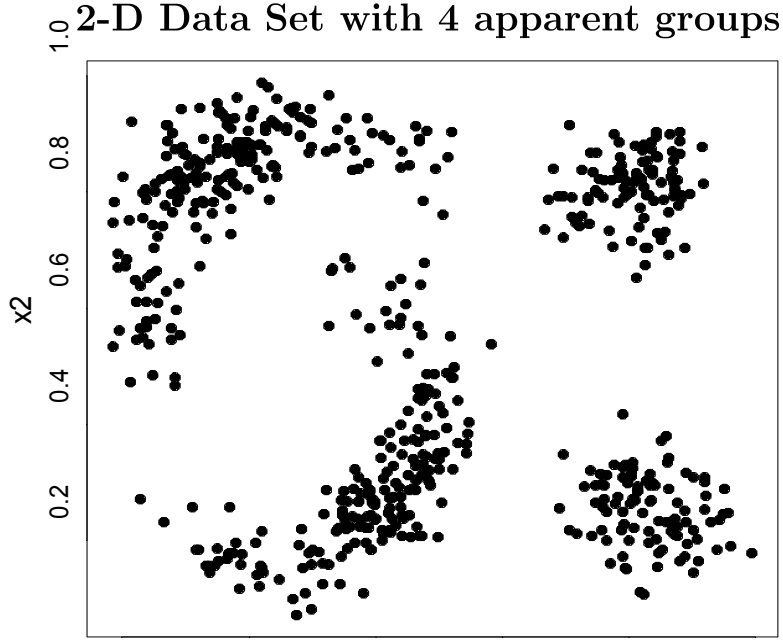
- Each node N of the tree represents a subset $D(N)$ of the support $L(0;p)$ of p (a *high density cluster* of p) and is associated with a density level $\lambda(N)$.
- Root node represents entire support of p; density level $\lambda(N) = 0$.
- Determining descendants of a node N:
  - Find lowest level λ_d for which $L(\lambda;p) \bigcap D(N)$ has two or more conn comp.
  - If no such λ_d exists, p has only one mode in $D(N)$; N is a leaf of the tree
  - Otherwise, let $C_1, C_2, ...C_k$ be the conn comp of $L(\lambda_d;p) \bigcap D(N)$. Create two daughter nodes for C_1 and C_2 (or $C_2 \bigcup ... \bigcup C_k$), each at level λ_d.
  - Apply definition recursively to daughter nodes.

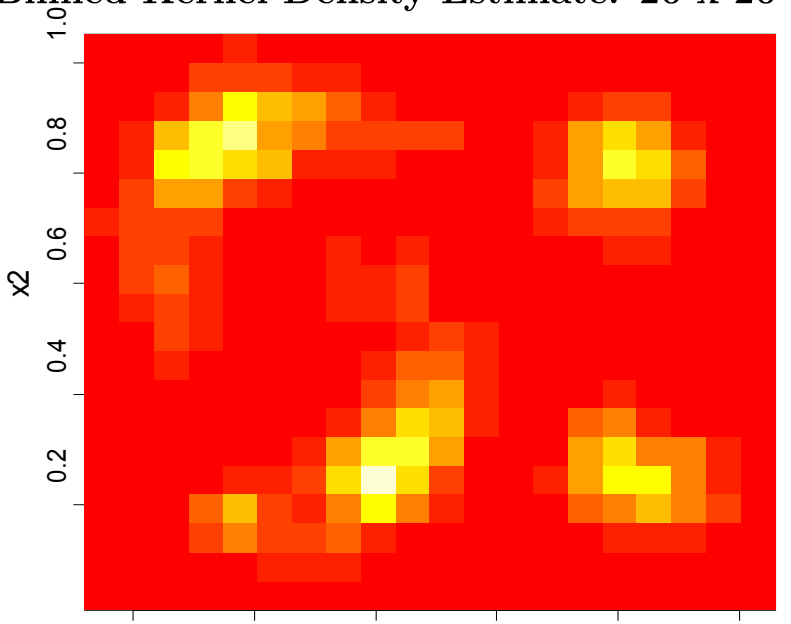
Representing the Connected Components' Hierarchical Structure:

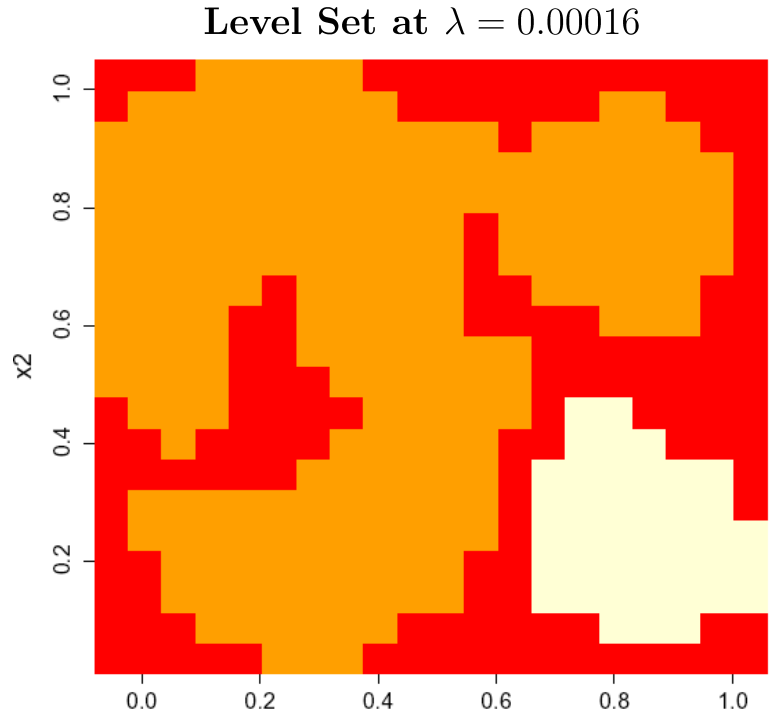
The Cluster Tree

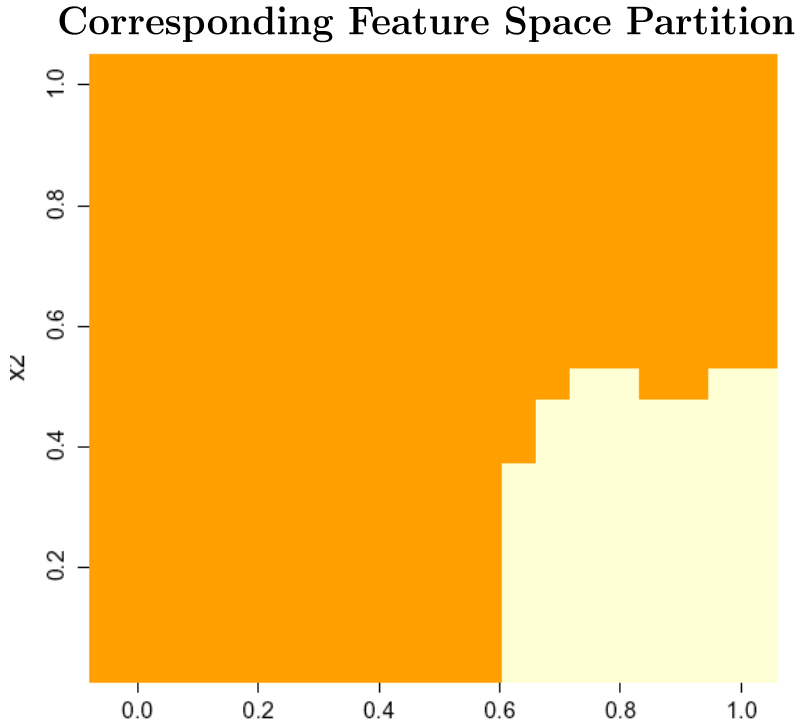




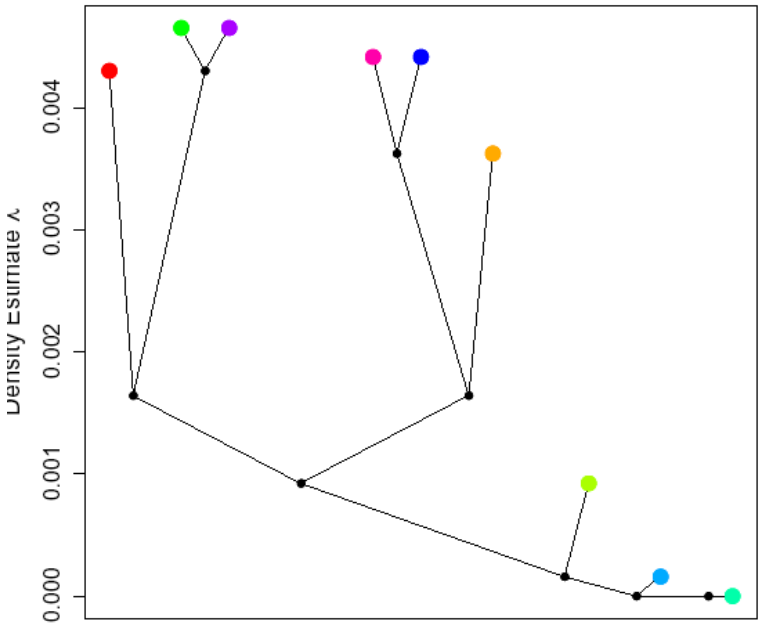

2-D Data Set with 4 apparent groups
x2
0.2
0.4
0.6
0.8
1.0


Binned Kernel Density Estimate: 20 x 20 grid
x2
0.2
0.4
0.6
0.8
1.0

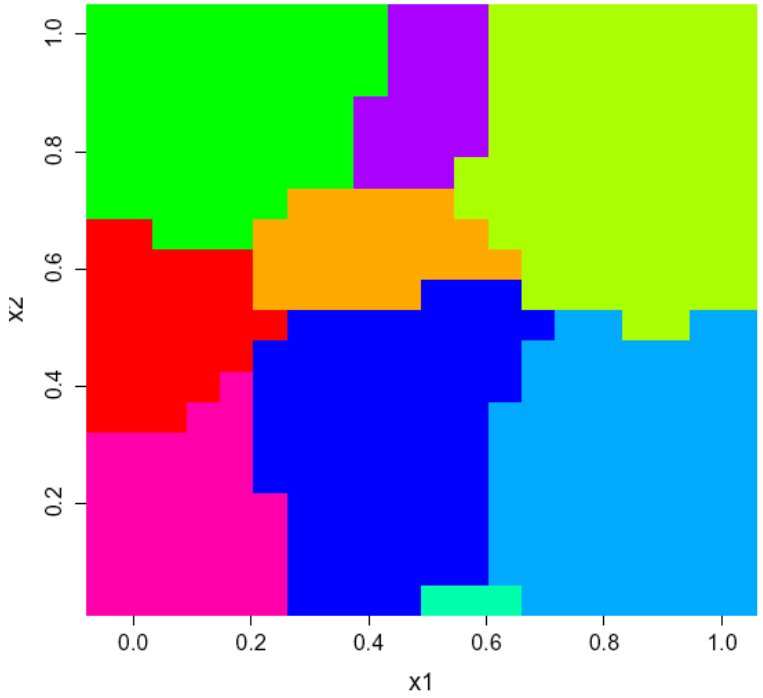

Level Set at $\lambda = 0.00016$
x2
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0


Corresponding Feature Space Partition
x2
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0

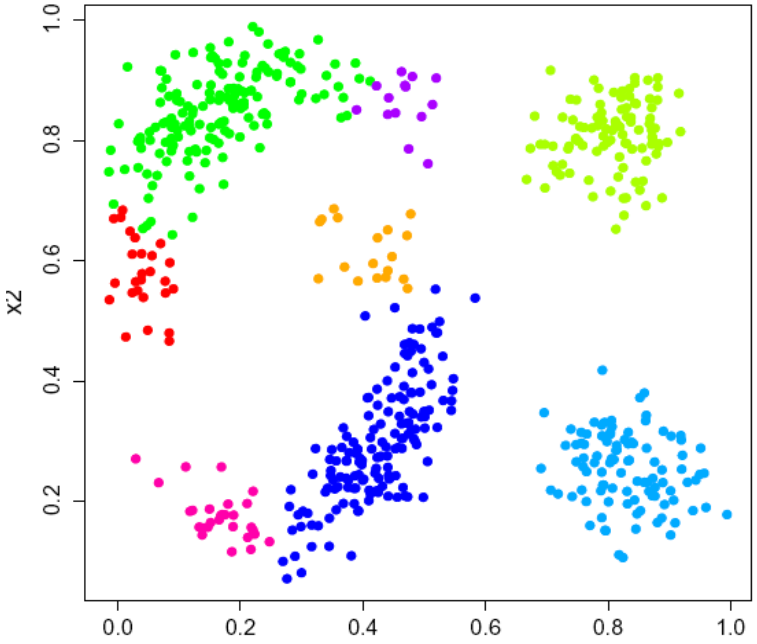
Cluster Tree for the BKDE



Partitioned Feature Space

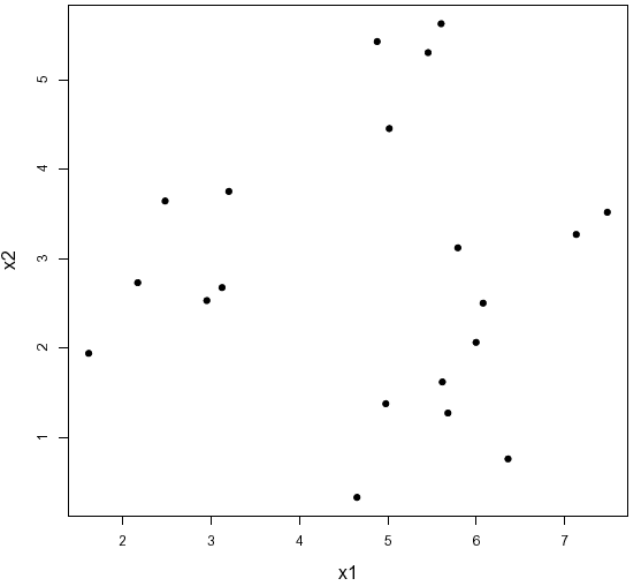


Final Cluster Assignments

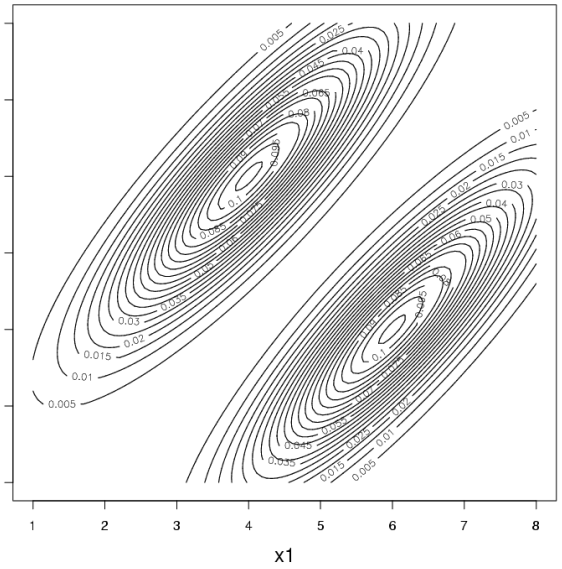


- Procedure identifies nine modes, eight clusters;
  cluster tree is exact/accurate for density estimate
- First split is artifact of density estimate;
  next three splits identified the four groups;
  subsequent splits correspond to spurious modes

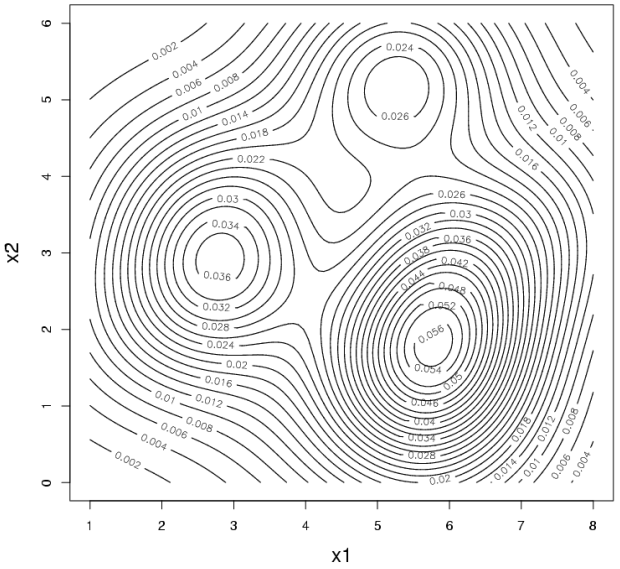
Original Data



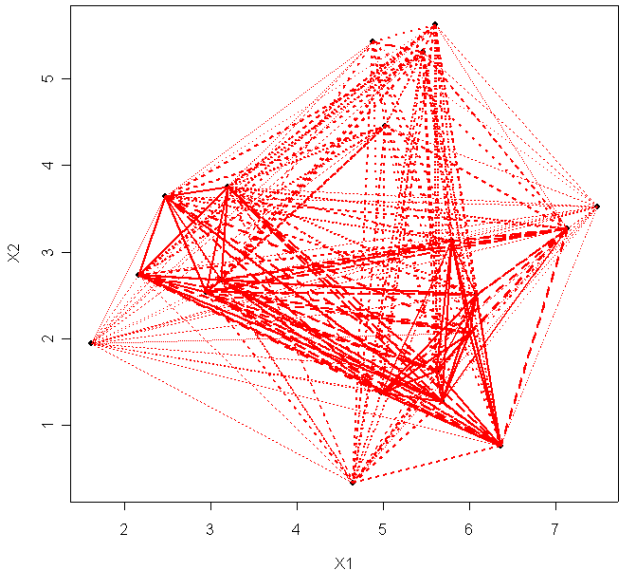
True Underlying Density



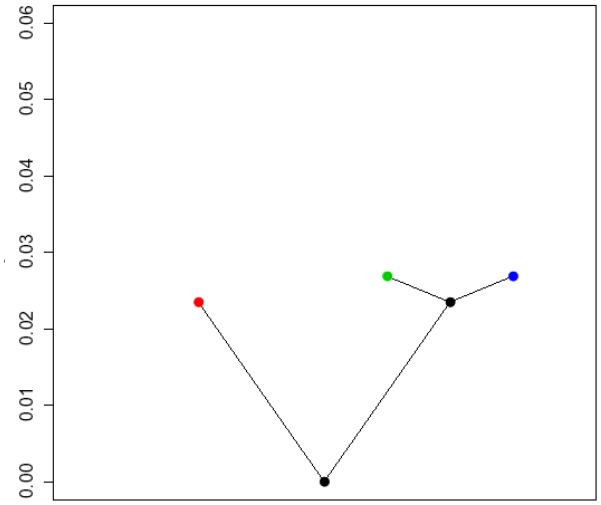
Gaussian Adaptive Kernel Density Estimate



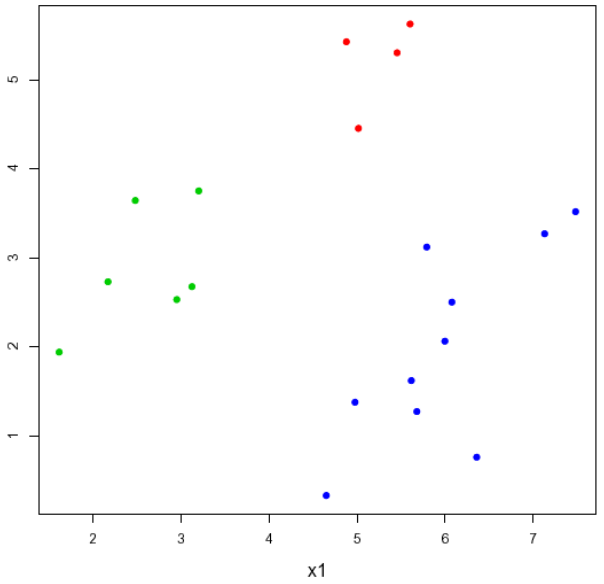
Complete Minimum Density Graph

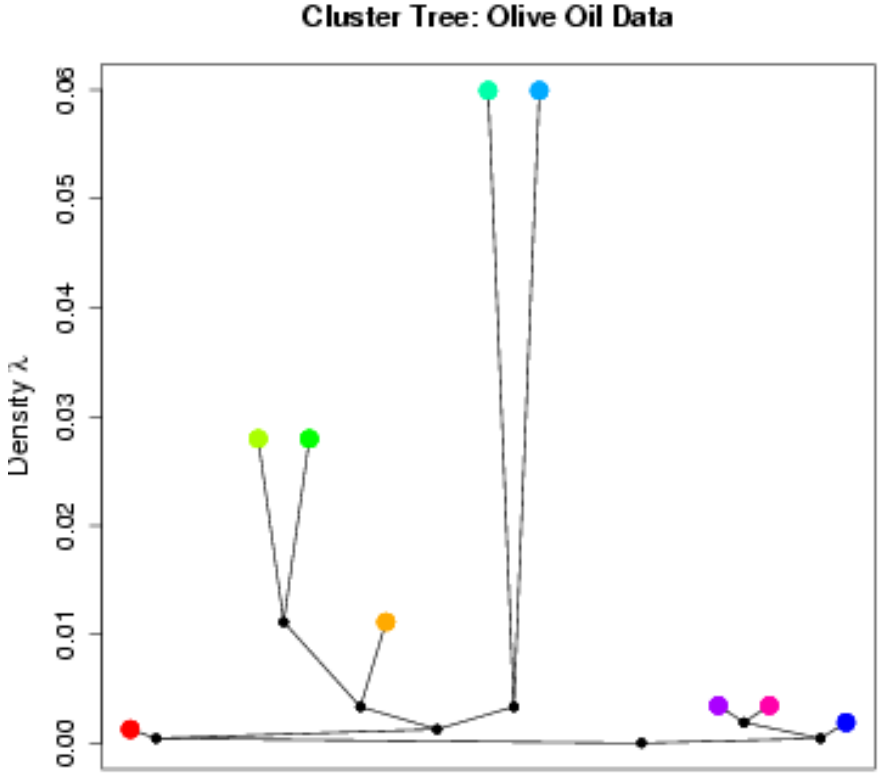


Cluster Tree



Final Cluster Assignment





All Observations: ARI = 0.587

	C1	C2	C3	C4	C5	C6	C7	C8	C9
A1	23	0	0	1	0	0	0	1	0
A2	0	47	0	8	0	0	1	0	0
A3	0	0	57	149	0	0	0	0	0
A4	5	17	0	13	0	0	0	1	0
A5	0	0	0	0	61	4	0	0	0
A6	0	0	0	0	5	28	0	0	0
A7	1	1	1	0	0	0	1	15	31
A8	0	0	0	0	0	0	0	50	0
A9	0	0	0	0	0	0	51	0	0

- Region 3 separates early; Areas 7, 8, 9 split shortly therafter
- Area 1 isolates itself from the remainder of Regions 1 and 2
- Followed by a split of Areas 2, 3, 4 from Region 2
- Areas 2 and 3 form clusters; Area 4 is split among clusters
- Region 2 splits into clusters for Areas 5 and 6
- Some misclassification of areas within regions;
  Very little misclassification across regions