Identifying connected components in Gaussian finite mixture models for clustering

Luca Scrucca
Department of Economics
Università degli Studi di Perugia
luca@stat.unipg.it

Workshop on Clustering Methods and Their Applications
Free University of Bozen-Bolzano
Bolzano, November 28, 2014
http://pro1.unibz.it/projects/Clustering_Methods_2014

Introduction

- ✓ In model-based clustering each component of a finite mixture model is associated to a group or cluster.
- ✓ Let $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$ be a sample of n iid observations, whose distribution can be specified by a pdf/pmf of the following form

$$f(\boldsymbol{x}; \boldsymbol{\Psi}) = \sum_{k=1}^{G} \pi_k f_k(\boldsymbol{x}; \boldsymbol{\theta}_k),$$

with parameters $\boldsymbol{\Psi} = \{(\pi_k, \boldsymbol{\theta}_k), k = 1, \ldots, G\}$ ($\pi_k > 0$, $\sum \pi_k = 1$), and G is the number of mixture components.

- ✓ Implicit assumption: **a mixture component** $\leftrightarrow$ **a cluster**
- ✓ Often, finite mixture of Gaussian densities are used for continuous data. However, a non-Gaussian cluster may require more than a single mixture Gaussian component.

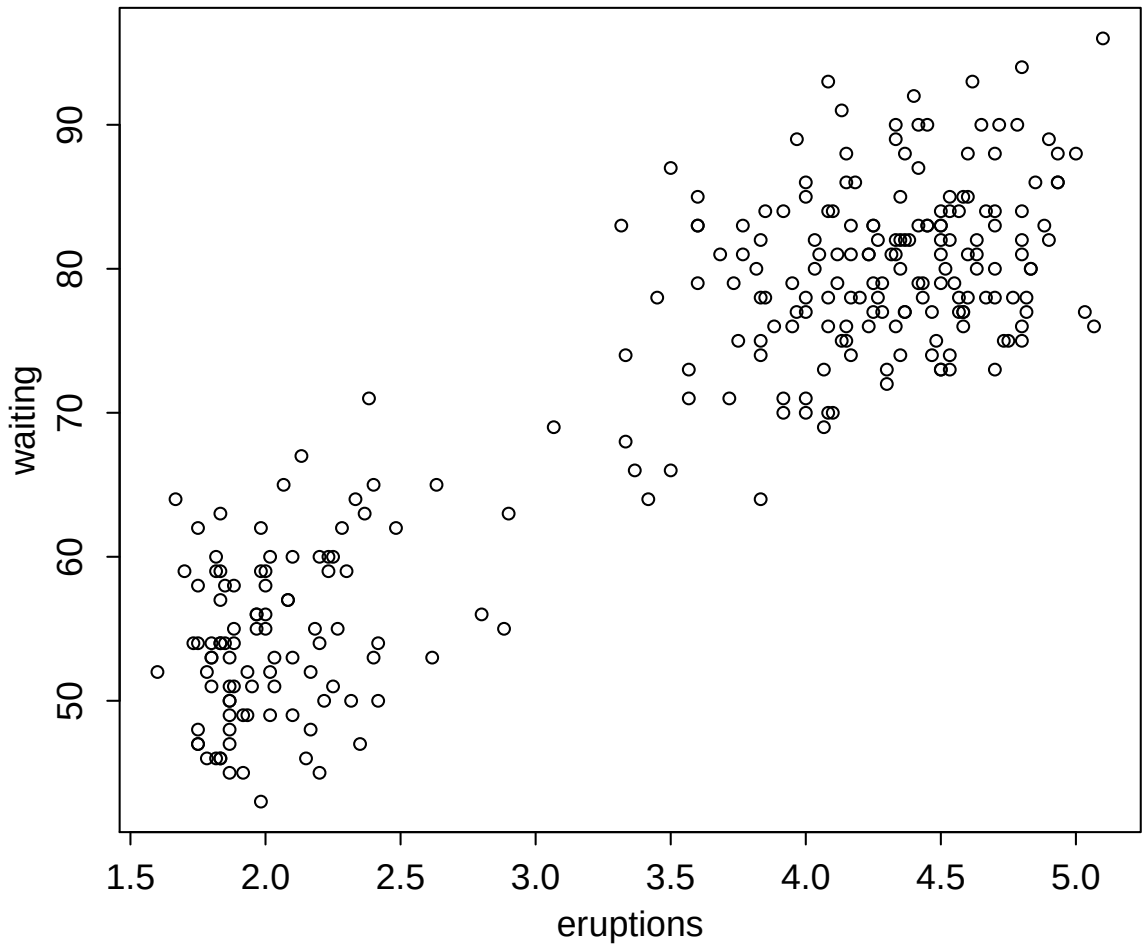
> *"it can be misleading to identify the number of Gaussian components with the number of clusters" (Hennig, 2010, p. 5).*

Motivating example: Old Faithful data

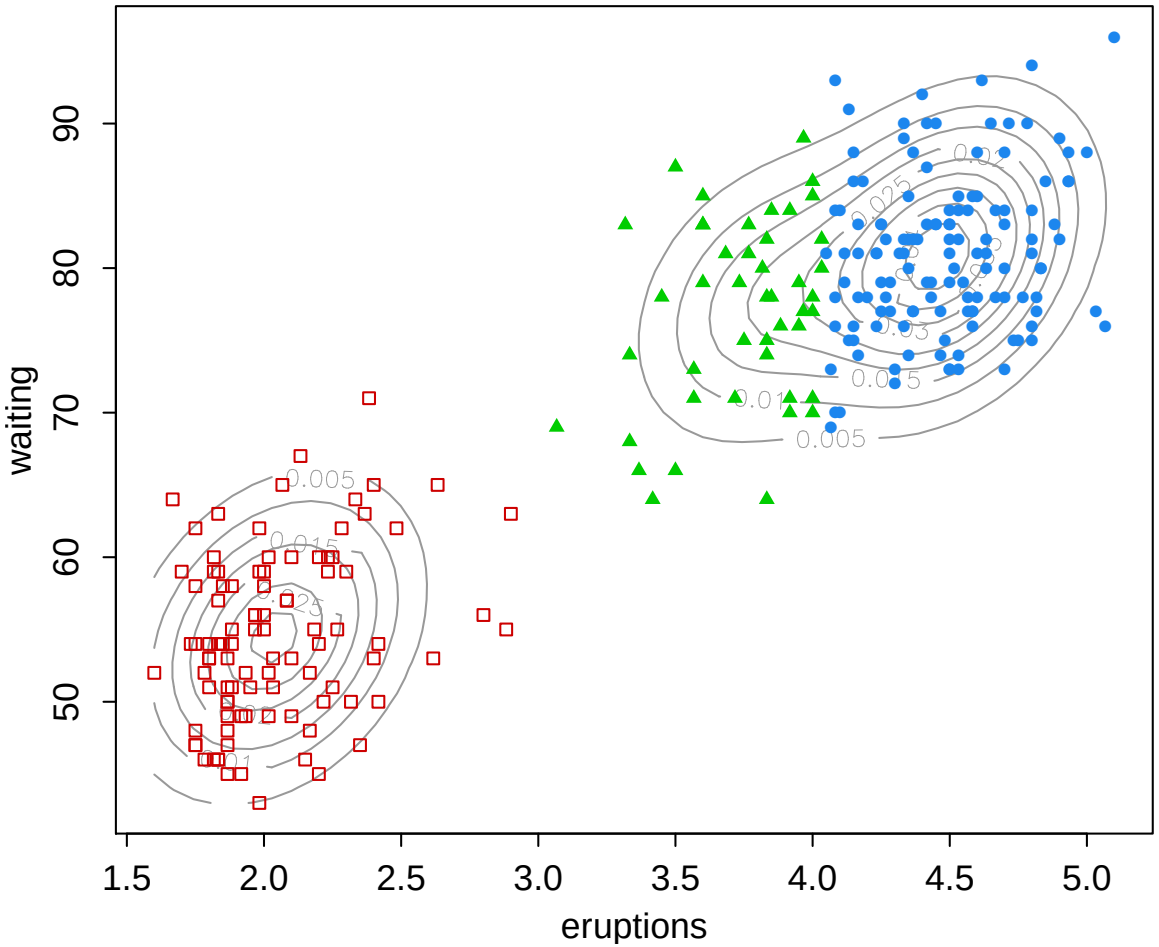
Old Faithful is a geyser located in Yellowstone National Park, Wyoming, US.

http://en.wikipedia.org/wiki/Old_Faithful
http://www.nps.gov/features/yell/webcam/oldFaithfulStreaming.html

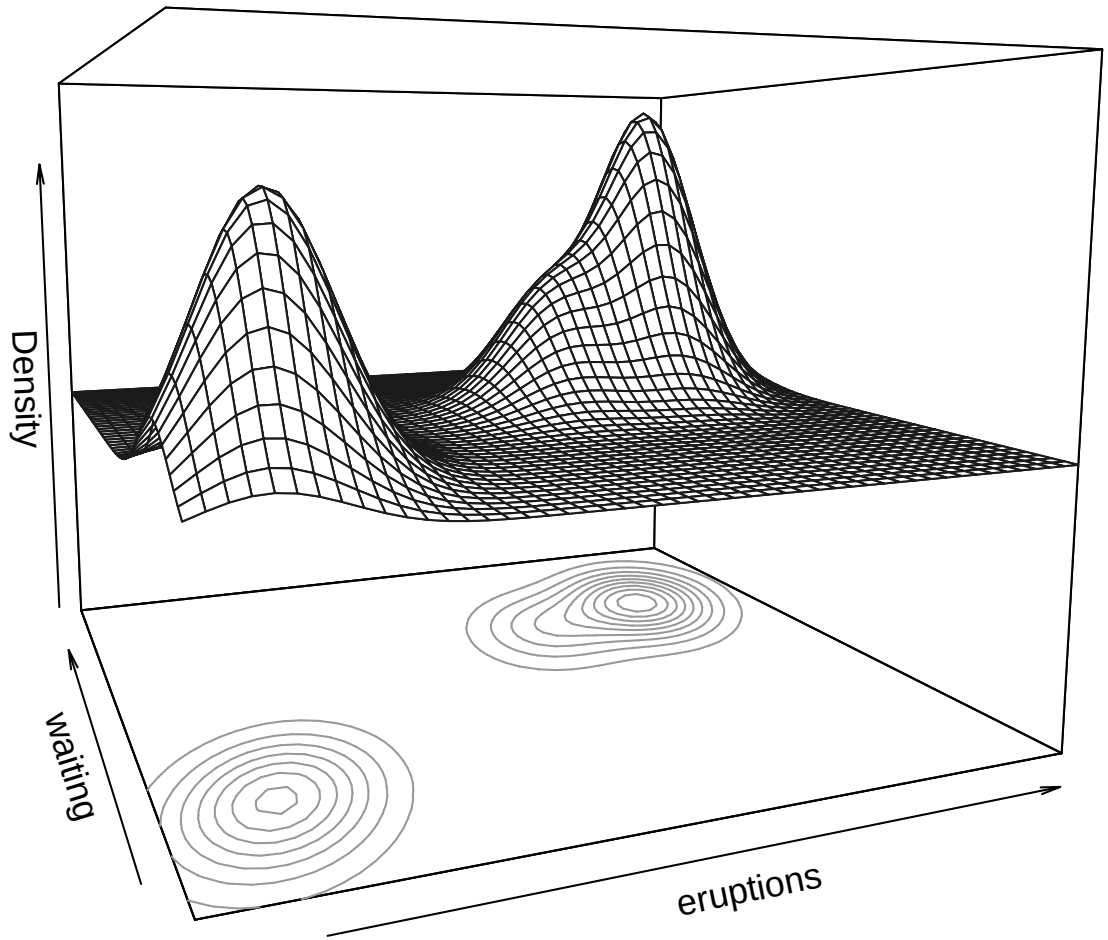
There is a direct relationship between the duration of Old Faithful's eruption (`eruptions`, time in mins) and the waiting time before it erupts again (`waiting`, time in mins).



Best GMM according to BIC is (EEE,3)



However, the bivariate density estimate clearly indicates the presence of two separate regions of high density:



Cluster definition

Working definition of clusters

Clusters may be thought of as regions of high density separated from other such regions by regions of low density. (Hartigan, 1975, p. 205)

- ✓ Fukunaga & Hostetler (1975) proposed the mean shift algorithm for detecting the modes of a nonparametric density estimate;
- ✓ Stuetzle (2003) presented a method which exploits the connection between the minimum spanning tree and the nearest neighbour density estimatie;
- ✓ Stuetzle & Nugent (2010) introduced level set clustering to find the hierarchical structure of connected components of a density level set;
- ✓ Azzalini & Torelli (2007) proposed a method based on nonparametric density estimation to find regions of high density. This has been extended to higher dimensionality by Menardi & Azzalini (2014).

Here I present a proposal which, using the working definition of clusters given by Hartigan, adapts the methodology of Azzalini & Torelli (2007) to model-based clustering.

Methodology

Level set

- ✓ For any threshold $c > 0$, the *upper level set* is defined as

$$L(c) = \{\boldsymbol{x} : \boldsymbol{x} \in \mathbb{R}^p, f(\boldsymbol{x}) > c\},$$

  i.e. the subset of $\mathbb{R}^p$ whose density is greater than c, with associated probability $p_c = \int_{L(c)} f(\boldsymbol{x}) d\boldsymbol{x}$.
- ✓ A level set $L(c)$ may be connected or not. In the latter case two or more regions of high density are detected.
- ✓ Hartigan (1975) defined the *high density clusters at level* c as the connected components of $L(c)$.

Mode function

- ✓ A step function $m(p)$ which gives the number of connected components of $L(c)$ as p varies in $(0, 1)$.
- ✓ Some properties:
  - $m(p) \geq 1$ for $p \in (0, 1)$;
  - by definition, $m(p) = 0$ for $p = 0$ and $p = 1$;
  - the number of modes M is given by the total number of increments of $m(p)$, counted with their multiplicity;
  - if the density f is unimodal, $M = 1$, then $m(p) = 1$ for $p \in (0, 1)$;
  - as c varies the connected components of $L(c)$ generate a hierarchical structure (i.e. a tree).

Sample data and density estimation via mixture modelling

- ✓ Given a iid sample $\mathcal{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n; \boldsymbol{x} \in \mathbb{R}^p\}$ drawn from a distribution with density $f(\boldsymbol{x})$, we may approximate this density using a GMM with G components of the form

$$f(\boldsymbol{x}) \approx \sum_{k=1}^{G} \pi_k \phi_k(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

where π_k = mixing probabilities ($\pi_k > 0$, $\sum_{k=1}^{G} \pi_k = 1$)
$\phi_k(\cdot)$ = multivariate Gaussian density of the k-th component with parameters $(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$.

- ✓ Parsimonious parametrisation of the component-covariance matrix is obtained using the eigen-decomposition $\boldsymbol{\Sigma}_k = \lambda_k \boldsymbol{D}_k \boldsymbol{A}_k \boldsymbol{D}_k^\top$ (Banfield & Raftery, 1993; Celeux & Govaert, 1995).
- ✓ MLEs are usually computed via the EM algorithm (McLachlan & Peel, 2000; Fraley & Raftery, 2002), while a standard model selection procedure (wrt number of mixture components and covariance matrix parametrisation) may be based on BIC (Schwartz, 1978).

Sample level set

- ✓ Consider the level set for the observed sample data:

$$S(c) = \{\boldsymbol{x}_i : \boldsymbol{x}_i \in \mathcal{X}, \widehat{f}(\boldsymbol{x}_i) > c\}, \qquad 0 < c \leq \max \widehat{f}$$

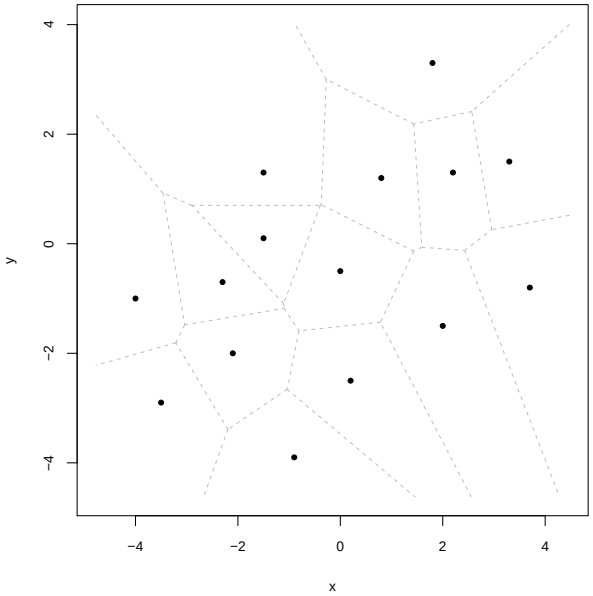
with associated relative frequency $\widehat{p}_c = \dfrac{|S(c)|}{n}$.

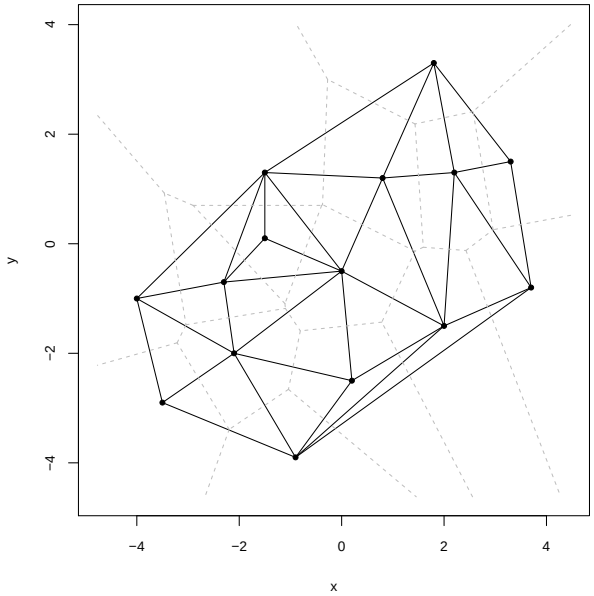
Connected sets

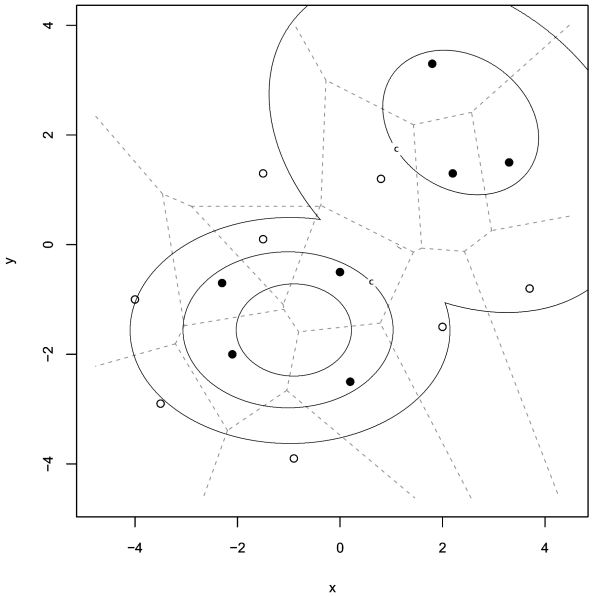
Connected sets are the connected components of $S(c)$ as c varies:

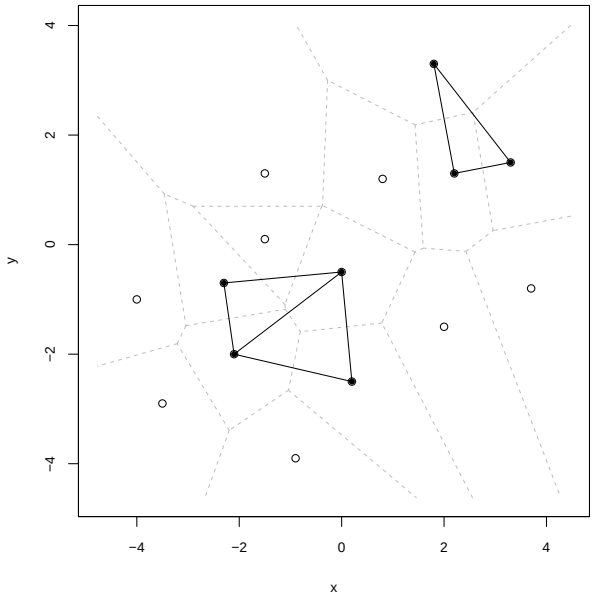
- ✓ Consider the Delaunay triangulation of sample points $\boldsymbol{x}_i$ obtained from *Voronoi tesselation* (see graphs).
- ✓ After removing the sample points $\boldsymbol{x}_i \notin S(c)$ and all the edges with at least one vertex among these points, a set of points is obtained which can form one or more connected components.
- ✓ Each connected component is a mode at density level c.
- ✓ Note that Delaunay triangulation can be obtained directly (and efficiently) without building the Voronoi diagram.

Sample data and density estimation via mixture modelling

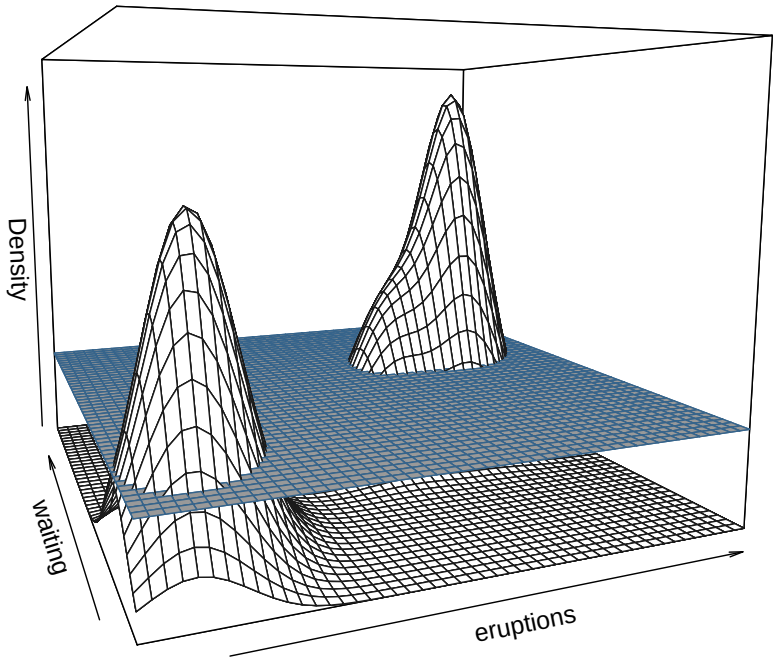




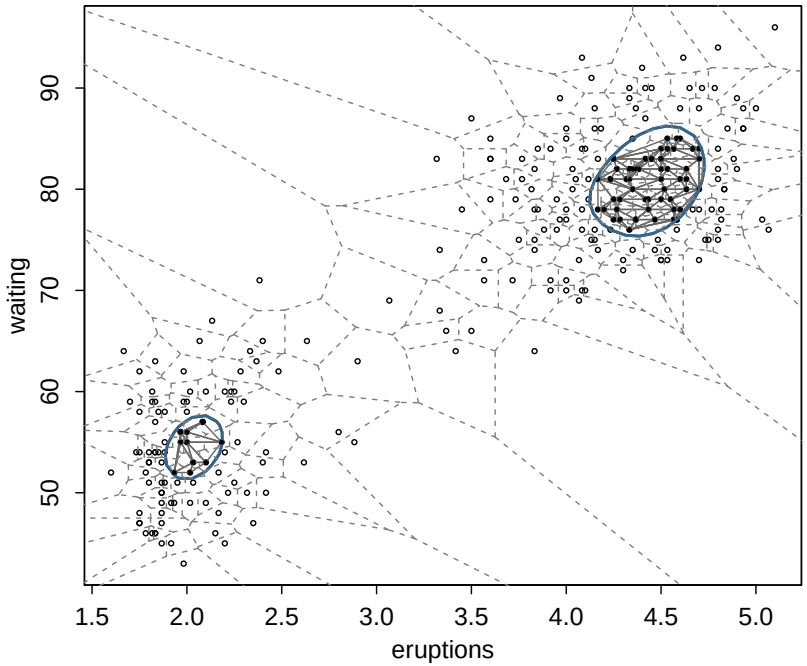




Example: Old Faithful data (continued)



Estimated density with a cutting plane at density level c.



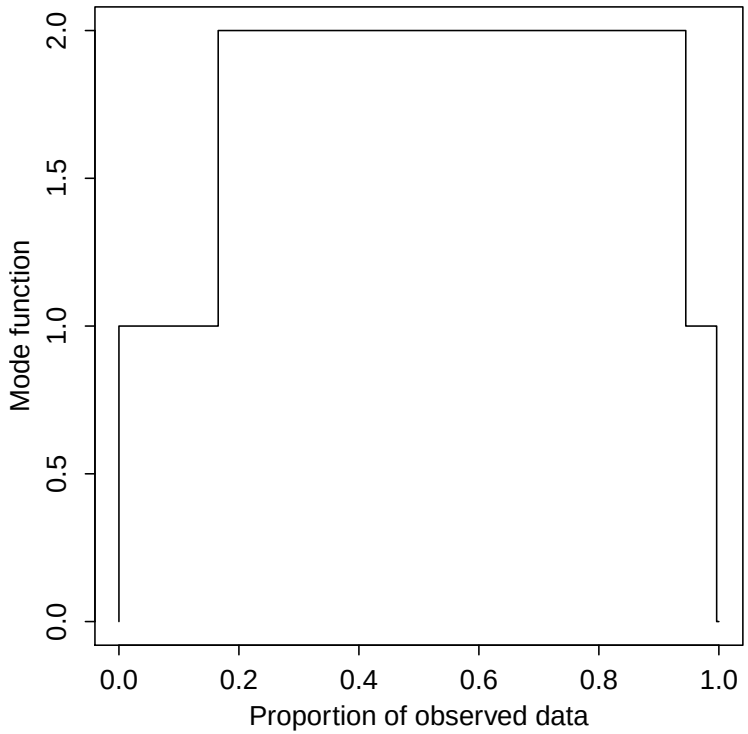
Voronoi diagrams for all the points, and Delaunay triangulation for $\boldsymbol{x}_i \in S(c)$ with $\widehat{p}_c = 0.26$.

Two connected components are clearly identified corresponding to local modes of the estimated density.

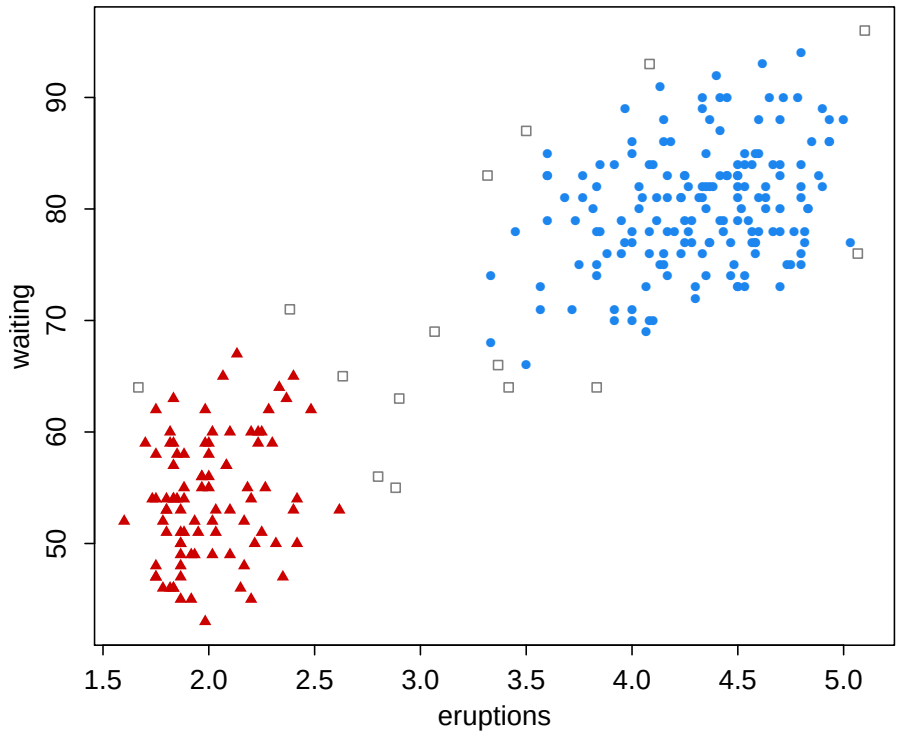
Identifying cluster cores

- ✓ For each p on an equally spaced grid in the range $(0, 1)$, the *sample level set* $S(c_p)$ is obtained.
- ✓ The empirical mode function $\widehat{m}(p)$ is obtained by counting the corresponding number of connected components.
- ✓ The total number of increments of $\widehat{m}(p)$, counted with their multiplicity, is equal to the number of modes M.
- ✓ Cluster cores are formed by the data lying in the regions around the detected modes.
- ✓ The number of clusters is estimated by identifying the connected components corresponding to the largest empirical mode $\widehat{m}(p)$, counted with their multiplicity.

Example: Old Faithful data (continued)



Empirical mode function showing the number of modes found as a function of the proportion of data points above a given density level.



Identified cluster cores (marked as ● and ▲) and remaining unlabelled data (□).

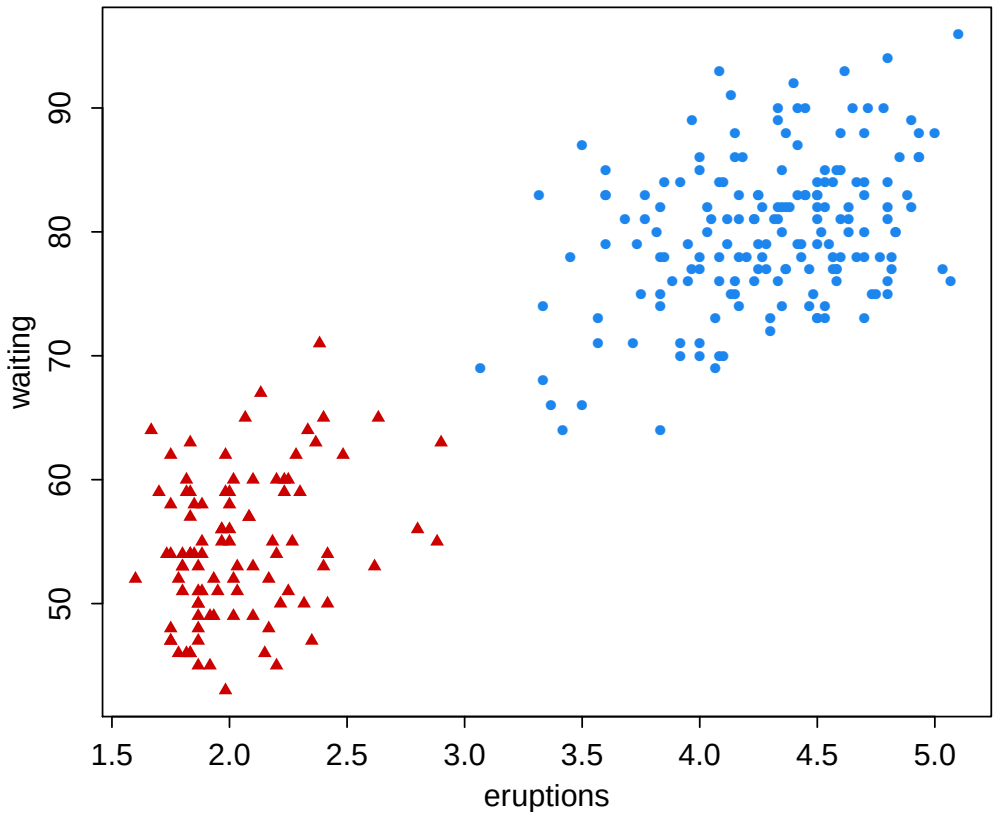
Classification of unallocated points

- ✓ Once cluster cores have been identified, some observations usually remain unlabelled and need to be classified.
- ✓ Semi-supervised learning is a class of techniques that make use of both unlabelled and labelled data for building a classifier (Zhu & Goldberg, 2009, Ch. 3; McLachlan & Peel, 2000, Sec. 2.19, named as "partial classification").
- ✓ However, *unallocated points are not positioned randomly in the feature space*, but are placed on the outskirts of cluster cores.
- ✓ Several algorithms could be adopted for this particular semi-supervised classification task. Here, we propose to fit a Gaussian mixture model (GMM) on the cluster cores and assign the unlabelled points to the cluster with the highest posterior probability in a block assignment procedure.

Algorithm for the assignment of unallocated points

1. Fit a supervised GMM using observations from the cluster cores and their labels.
2. From the GMM estimated on the n_{inc} allocated points, calculate the conditional probability $\widehat{z}_{ik} = \Pr(\boldsymbol{x}_i \notin C_k | \mathcal{X}_{\text{inc}} \subset \mathcal{X})$;
3. compute the log-ratios $r_{ik} = \log(\widehat{z}_{ik}/(1 - \widehat{z}_{ik}))$ for all the unallocated observations;
4. update the classification by assigning those observations whose $r_{ik} \geq q_k$ to cluster core C_k for which $\widehat{z}_{ik}$ is the maximum, where q_k is the $\sqrt{n_{\text{inc}}/n}$ quantile of the empirical distribution of log-ratios r_{ik} within group k;
5. if $n_{\text{inc}} < n$ repeat steps 2–4, where n_{inc} is the updated number of allocated points.

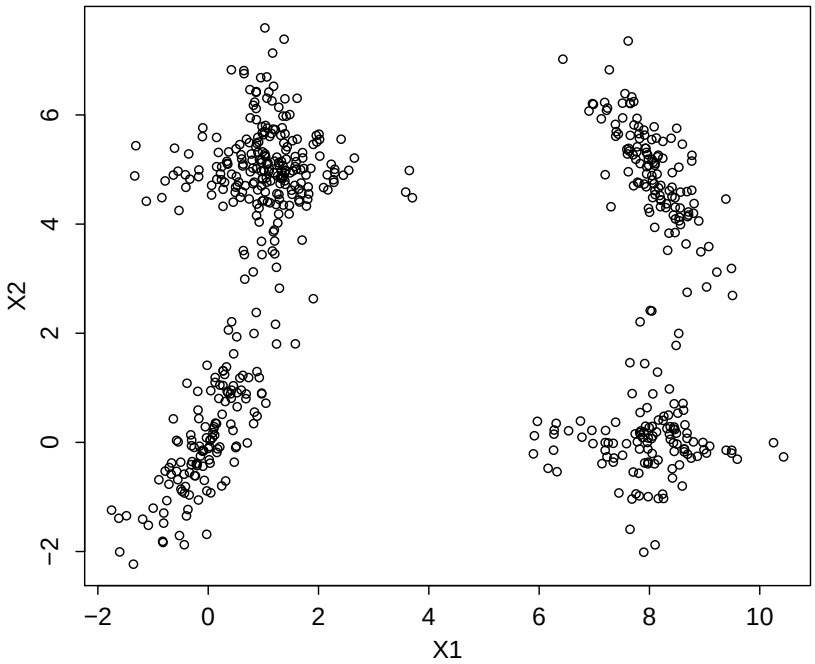
Example: Old Faithful data (continued)



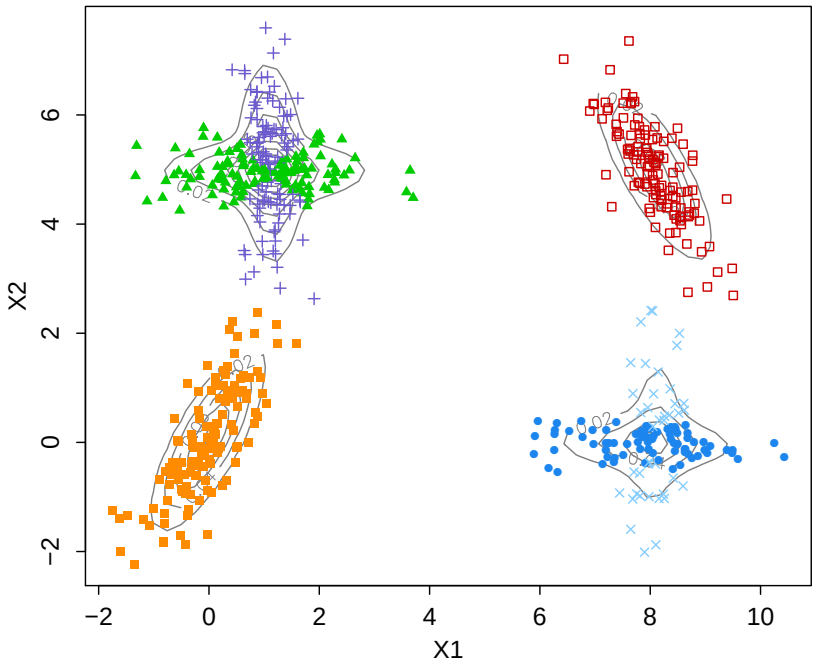
Plot of final clustering for the Old Faithful data obtained after unlabelled data have been assigned to one of the cluster cores.

Example: synthetic data with overlapping components

Consider a simulated sample of 600 observations generated from a bivariate mixture of six Gaussian components (Baudry et al., 2010):

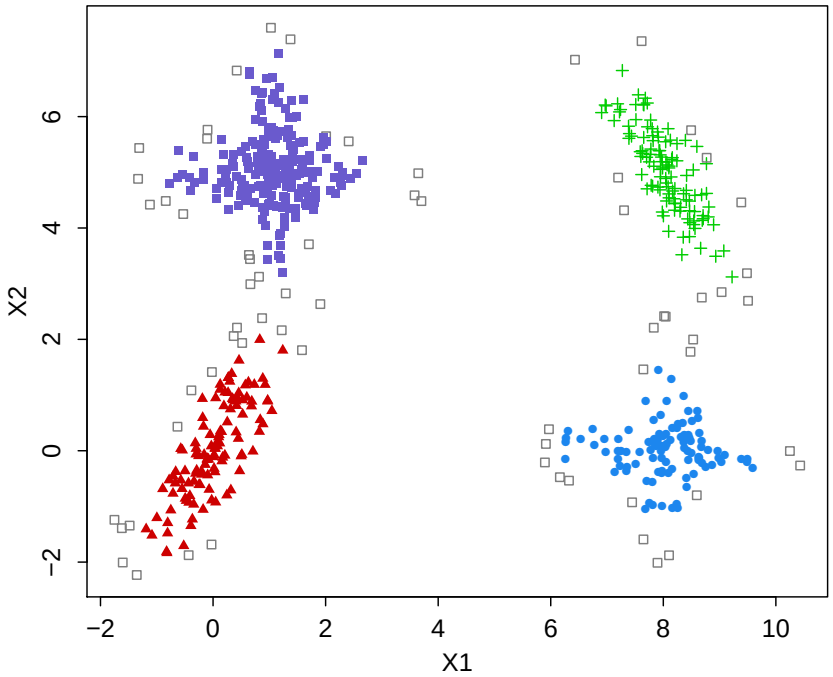


A visual inspection suggests a four clusters solution

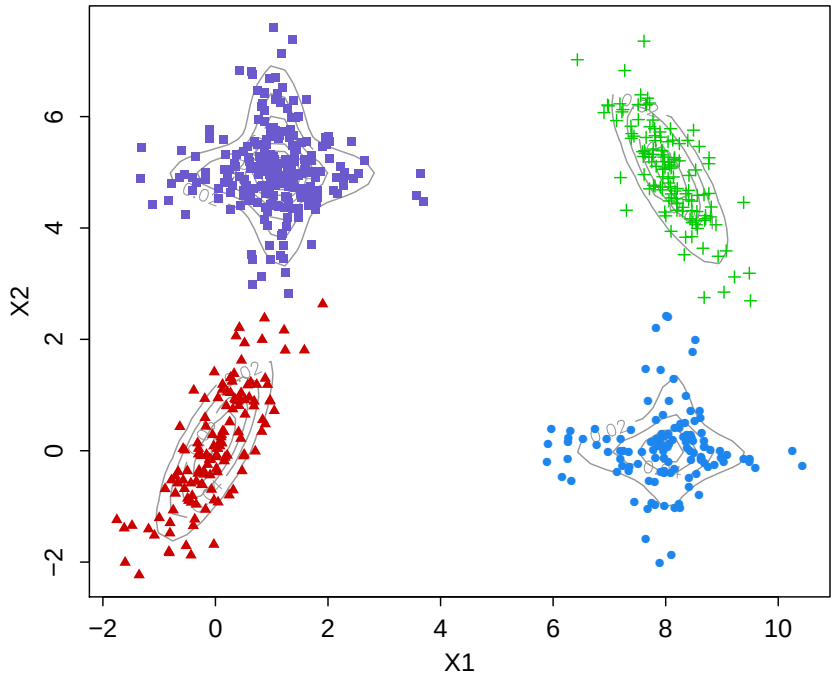


GMM model with the largest BIC is $(\mathtt{VVV}, 6)$

Example: synthetic data with overlapping components



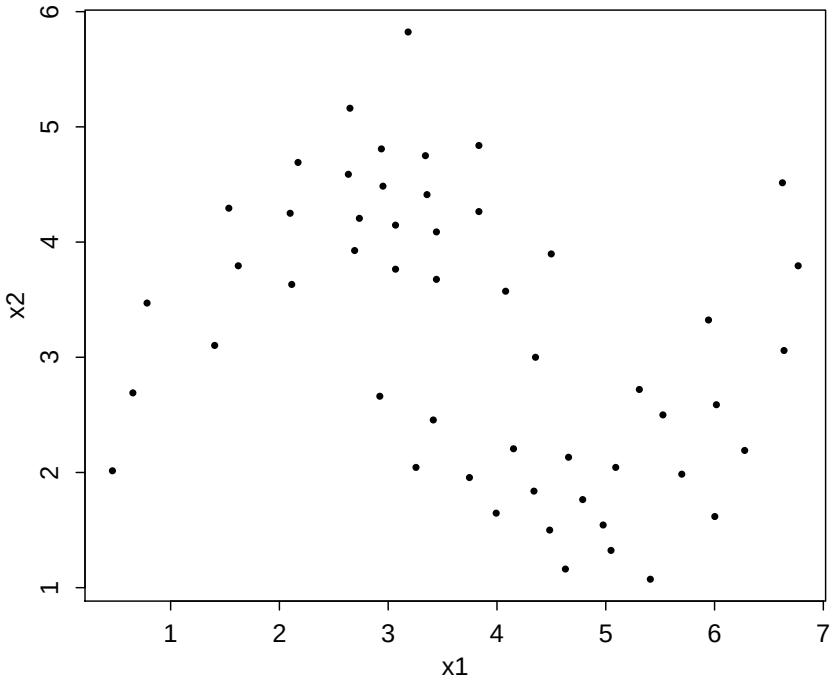
Cluster cores identified by the GMMHD procedure



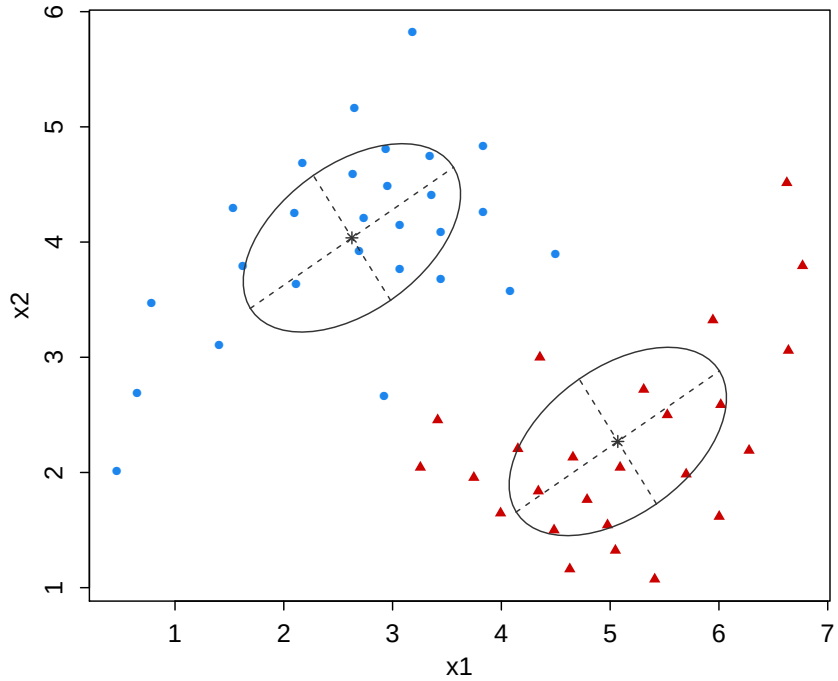
Final GMMHD clustering

Example: two bivariate elongated clusters

Wong & Lane (1983) discussed a data example where the groups are not linearly separable:

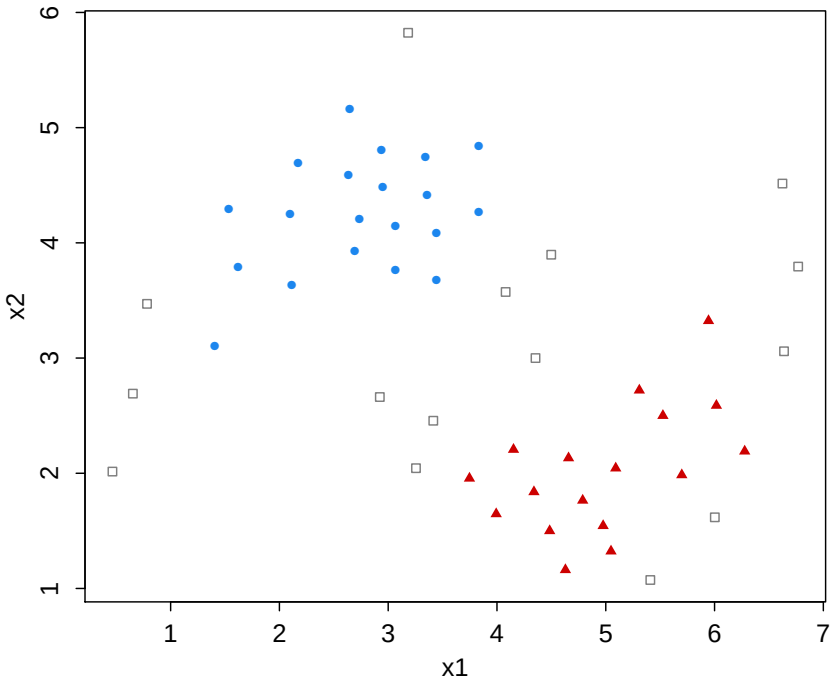


Sample data from two artificial elongated clusters

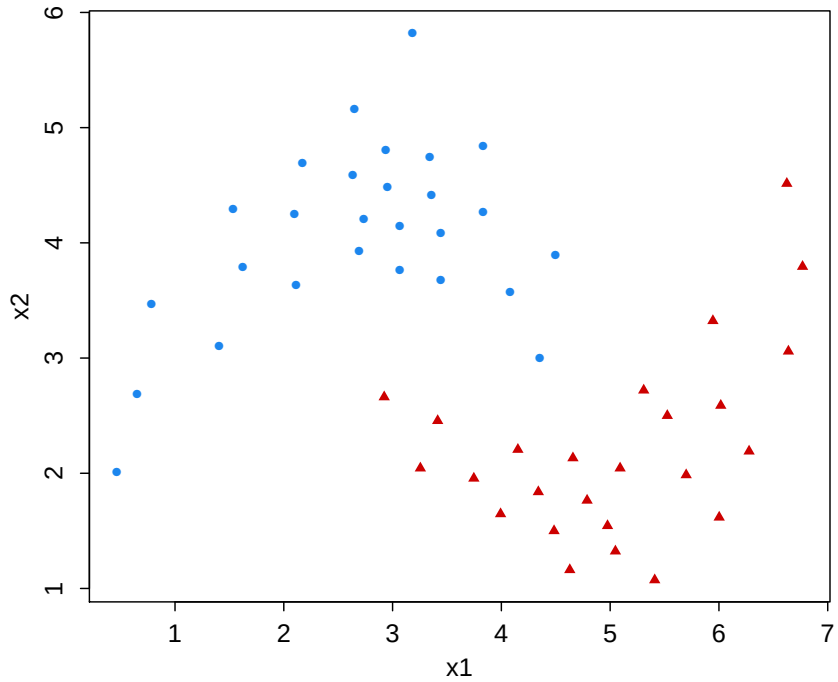


GMM clustering solution (EEE, 2)

Example: two bivariate elongated clusters



Cluster cores identified by the GMMHD procedure



Final GMMHD clustering

High dimensional case

- ✓ The computational complexity of Delaunay triangulation grows exponentially with the dimensionality of data (unfeasible for $p > 5$).
- ✓ The basic idea is to project the data on to a suitable subspace of reduced dimensionality, where connected components can be easily found.
- ✓ GMMDR (Gaussian Mixture Modelling on a Dimension Reduced subspace) is a dimension reduction method which aims at finding the *smallest subspace which captures the clustering information contained in the data* (Scrucca, 2010).
- ✓ The core of the method is to identify those directions where the cluster means $\boldsymbol{\mu}_g$, and the cluster covariances $\boldsymbol{\Sigma}_g$, vary as much as possible, provided that each direction is $\boldsymbol{\Sigma}$-orthogonal to the others.
- ✓ However, here we are more interested in finding those directions which show the maximal separation among clusters (Scrucca, 2014).

Directions estimation

- To recover the directions with the largest separation among clusters, consider the following kernel matrix

$$\boldsymbol{M} = \sum_{k=1}^{G} \pi_k (\boldsymbol{\mu}_k - \boldsymbol{\mu})(\boldsymbol{\mu}_k - \boldsymbol{\mu})^\top$$

- The basis $\boldsymbol{\beta} \in \mathbb{R}^{p \times r}$ of the projection subspace $\mathcal{S}(\boldsymbol{\beta})$ is obtained by solving

$$\arg\max_{\boldsymbol{\beta}} \boldsymbol{\beta}^\top \boldsymbol{M} \boldsymbol{\beta} \qquad \boldsymbol{\beta}^\top \boldsymbol{\Sigma} \boldsymbol{\beta} = \boldsymbol{I}_r \tag{1}$$

- The solution is computed via the generalised eigen-decomposition of $\boldsymbol{M}$ wrt $\boldsymbol{\Sigma}$.
- $\dim(\mathcal{S}(\boldsymbol{\beta})) = r \leq \min(G-1, p)$, and directions are ordered according to the corresponding eigenvalues.
- Using $\widehat{\boldsymbol{M}}$ computed from the estimates obtained fitting a GMM, and the sample covariance $\widehat{\boldsymbol{\Sigma}}$, the solution of (1) gives $\widehat{\boldsymbol{\beta}}$.
- The data are then projected as $\boldsymbol{Z} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}$.

Pruning directions

- ✓ Because some directions are associated with small eigenvalues, we would like to discard them because they provide little or no clustering information.
- ✓ A subset selection procedure discussed in Scrucca (2010), and based on the proposal of Raftery & Dean (2006), is adopted.
- ✓ The basic idea is to use BIC to evaluate the inclusion/exclusion of a feature from the set of active features in a stepwise greedy search algorithm.
- ✓ Once the relevant GMMDR directions have been obtained, the GMMHD algorithm can be applied on the selected features.

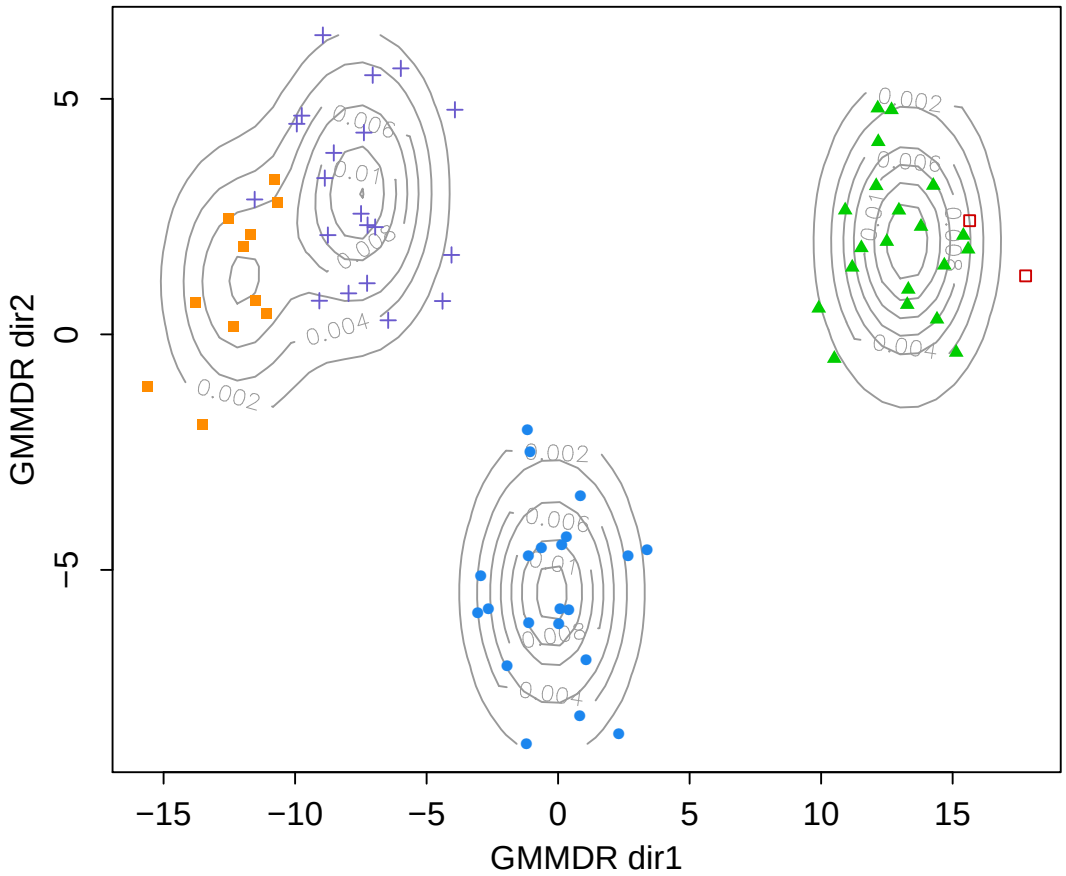
Flea beetles data

- ✓ Data on 6 physical measurements on three species of flea beetles (Ch. concinna, Ch. heptapotamica, and Ch. heikertingeri) are measured on 74 observations.

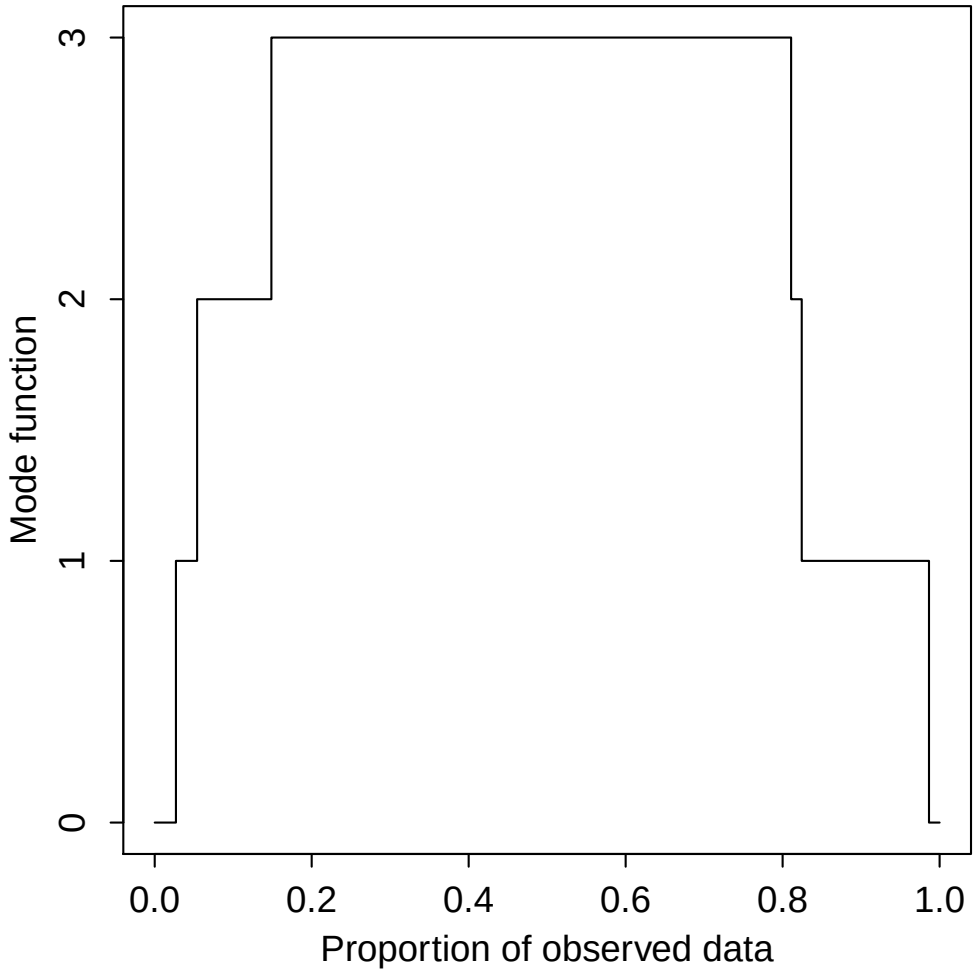
```
Mclust EEE model with 5 components:

 log.likelihood  n df       BIC       ICL
      -1292.308 74 55 -2821.339 -2825.769

            cluster
group         1  2  3  4  5
  Concinna   21  0  0  0  0
  Heikert.    0  0  0 20 11
  Heptapot.   0  2 20  0  0

AdjRandIndex = 0.7676
```

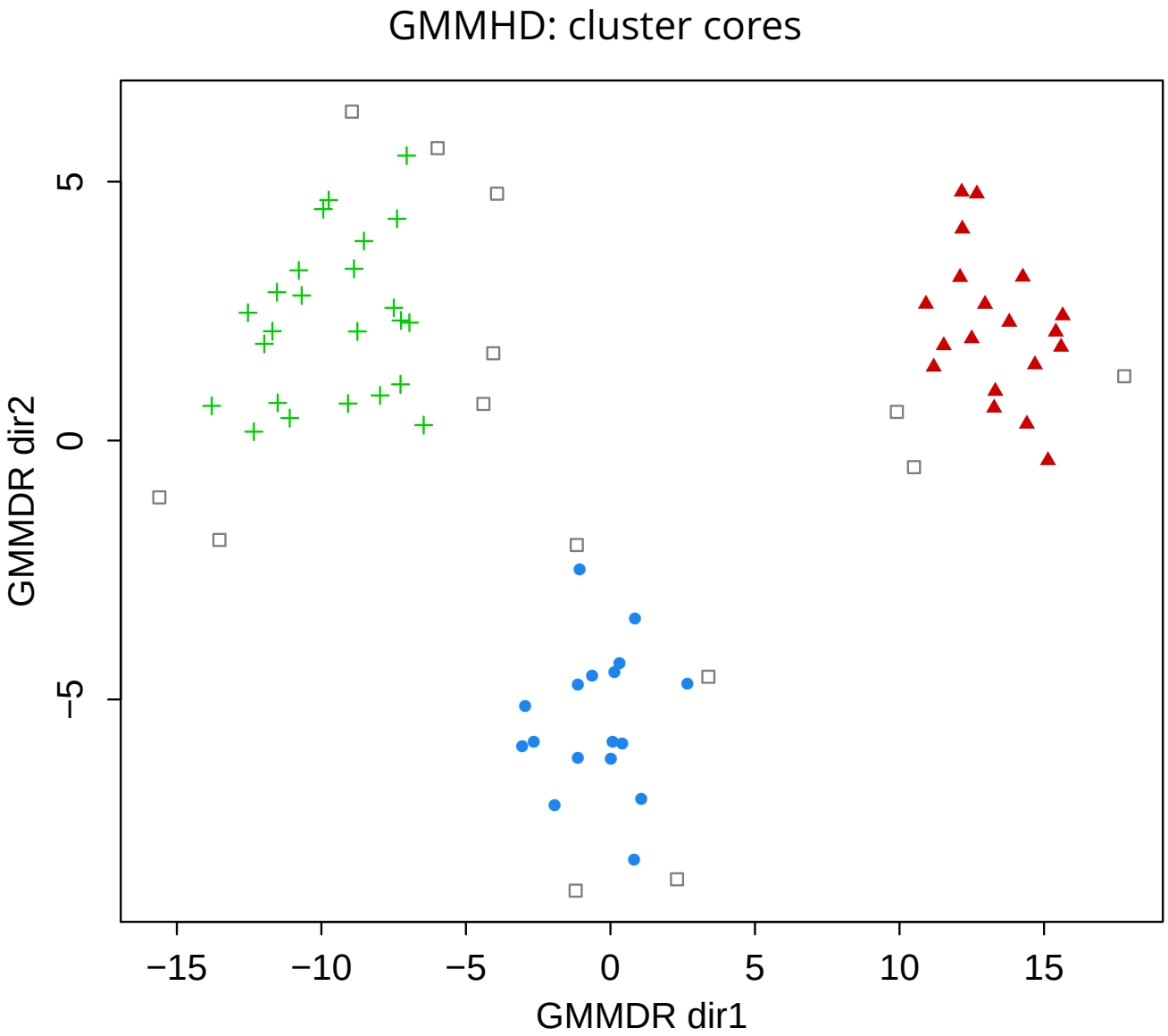
```
Estimated basis vectors:
           Dir1      Dir2
tars1 -0.229559 -0.019778
tars2  0.142747 -0.082415
head   0.422500  0.452652
aede1  0.010746 -0.361974
aede2 -0.861267 -0.809869
aede3  0.080766 -0.031781

                 Dir1    Dir2
Eigenvalues    1.8604   1.346
Cum. %        58.0226 100.000
```

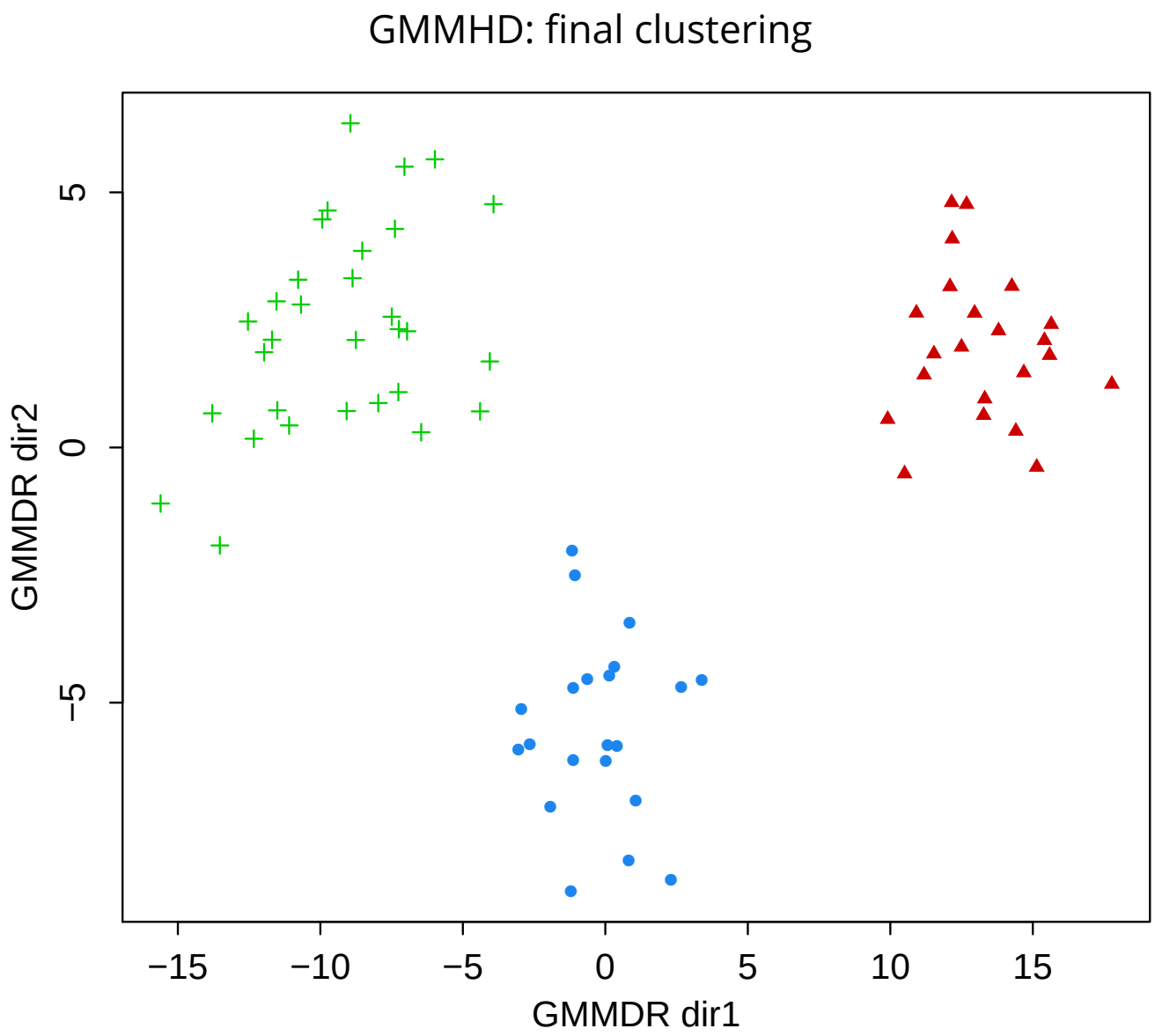
Mclust solution projected along the first two GMMDR directions



```
GMMHD

Initial cluster cores:
   1    2    3 <NA>
  17   19   24   14

Final clustering:
 1  2  3
21 22 31

             cluster
group          1  2  3
  Concinna    21  0  0
  Heikert.     0  0 31
  Heptapot.    0 22  0

AdjRandIndex = 1.00
```




GMMHD: cluster cores
GMMDR dir2
GMMDR dir1
5
0
−5
−15
−10
−5
0
5
10
15


GMMHD: final clustering
GMMDR dir2
GMMDR dir1
5
0
−5
−15
−10
−5
0
5
10
15

Yeast data

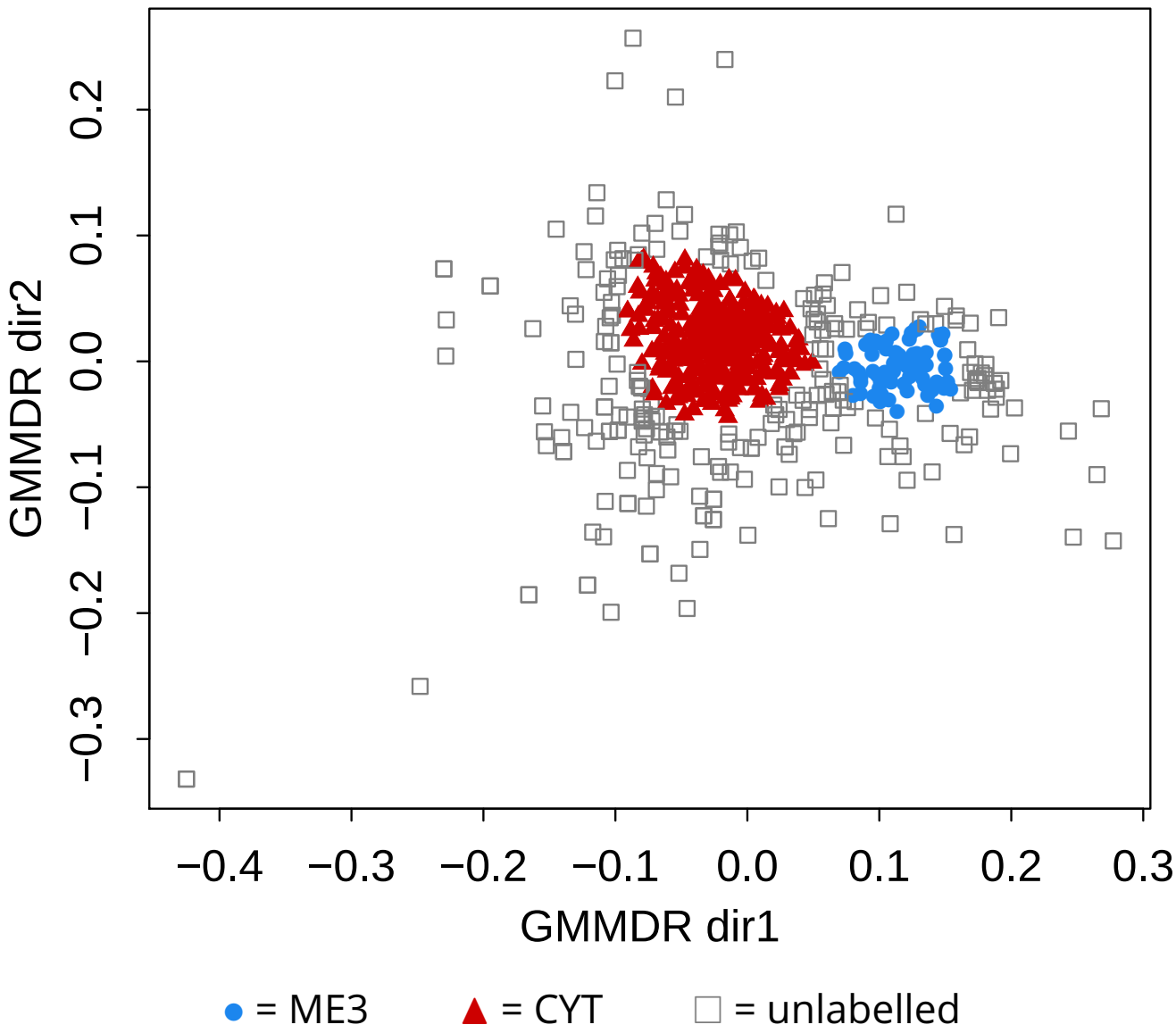
- ✓ Franczak et al. (2013) analysed a dataset with 1,484 proteins in two cellular localisation sites (CYT = cytosolic or cytoskeletal, ME3 = membrane protein, no N-terminal signal) and three variables for clustering: McGeoch's method for signal sequence recognition (mcg), the score of the ALOM membrane spanning region prediction program (alm), and the score for the discriminant analysis of the amino acid content of vacuolar and extracellular proteins (vac).
- ✓ They fitted a mixture of shifted asymmetric Laplace (SAL) distributions for clustering purposes, which gave favourable results (ARI $= 0.8134$).
- ✓ The GMM with the largest BIC is model EEI with 8 components (ARI $= 0.4972$), where a large number of components is required to account for the asymmetry in the data.

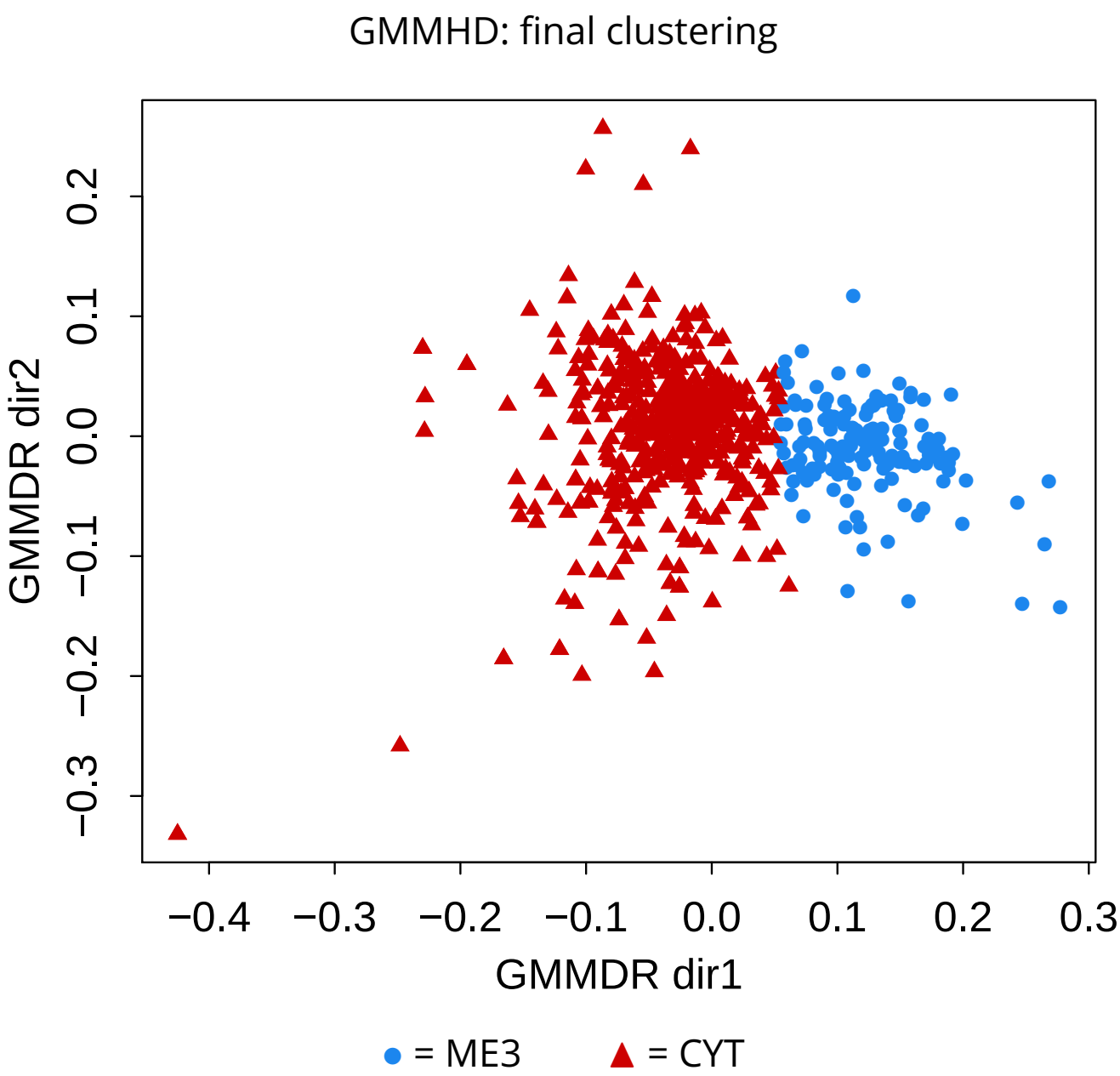
GMMDR

```
Estimated basis vectors:
         Dir1     Dir2
mcg -0.089080 0.066101
alm -0.993285 0.112436
vac -0.073824 0.991458

                Dir1     Dir2
Eigenvalues   0.5737   0.0283
Cum. %       95.2976 100.0000
```

GMMHD

```
Initial cluster cores:
   1    2 <NA>
 311   76  239

Final clustering:
  1   2
475 151


     cluster
group   1   2
  CYT 457   6
  ME3  18 145

AdjRandIndex = 0.8427
```


GMMHD: cluster cores
GMMDR dir2
GMMDR dir1
0.2
0.1
0.0
−0.1
−0.2
−0.3
−0.4
−0.3
−0.2
−0.1
0.0
0.1
0.2
0.3
= ME3
= CYT
= unlabelled


GMMHD: final clustering
GMMDR dir2
GMMDR dir1
0.2
0.1
0.0
−0.1
−0.2
−0.3
−0.4
−0.3
−0.2
−0.1
0.0
0.1
0.2
0.3
= ME3
= CYT

Italian wines data

- ✓ Data on 13 chemical and physical properties of three types of wine (Barolo, Grignolino, and Barbera) measured on 178 observations.
- ✓ We perform the analysis on standardized scale.

```
Mclust VEI (diagonal, equal shape) model with 8 components:

 log.likelihood   n  df       BIC       ICL
      -2392.975 178 131 -5464.765 -5478.056

Clustering table:
 1  2  3  4  5  6  7  8
40 18 22 22  4 27 18 27

             cluster
group          1  2  3  4  5  6  7  8
  Barbera      0  0  0  0  4  0 17 27
  Barolo      40 18  1  0  0  0  0  0
  Grignolino   0  0 21 22  0 27  1  0

AdjRandIndex = 0.4808
```

GMMDR

Estimated basis vectors:

	Dir1	Dir2
Alcohol	0.15030	-0.34141
Malic	-0.01957	-0.15750
Ash	0.03459	-0.26951
Alcalinity	-0.20998	0.26606
Magnesium	0.00034	0.08521
Phenols	-0.04041	0.13265
Flavanoids	0.67830	0.09616
Nonflavanoid	0.01072	-0.00663
Proanthocyanins	-0.02429	0.11065
Intensity	-0.46219	-0.44011
Hue	0.05599	0.09705
OD280	0.23485	0.13693
Proline	0.44429	-0.66855

	Dir1	Dir2
Eigenvalues	1.598	1.1764
Cum. %	57.598	100.0000

GMMHD

Initial cluster cores:

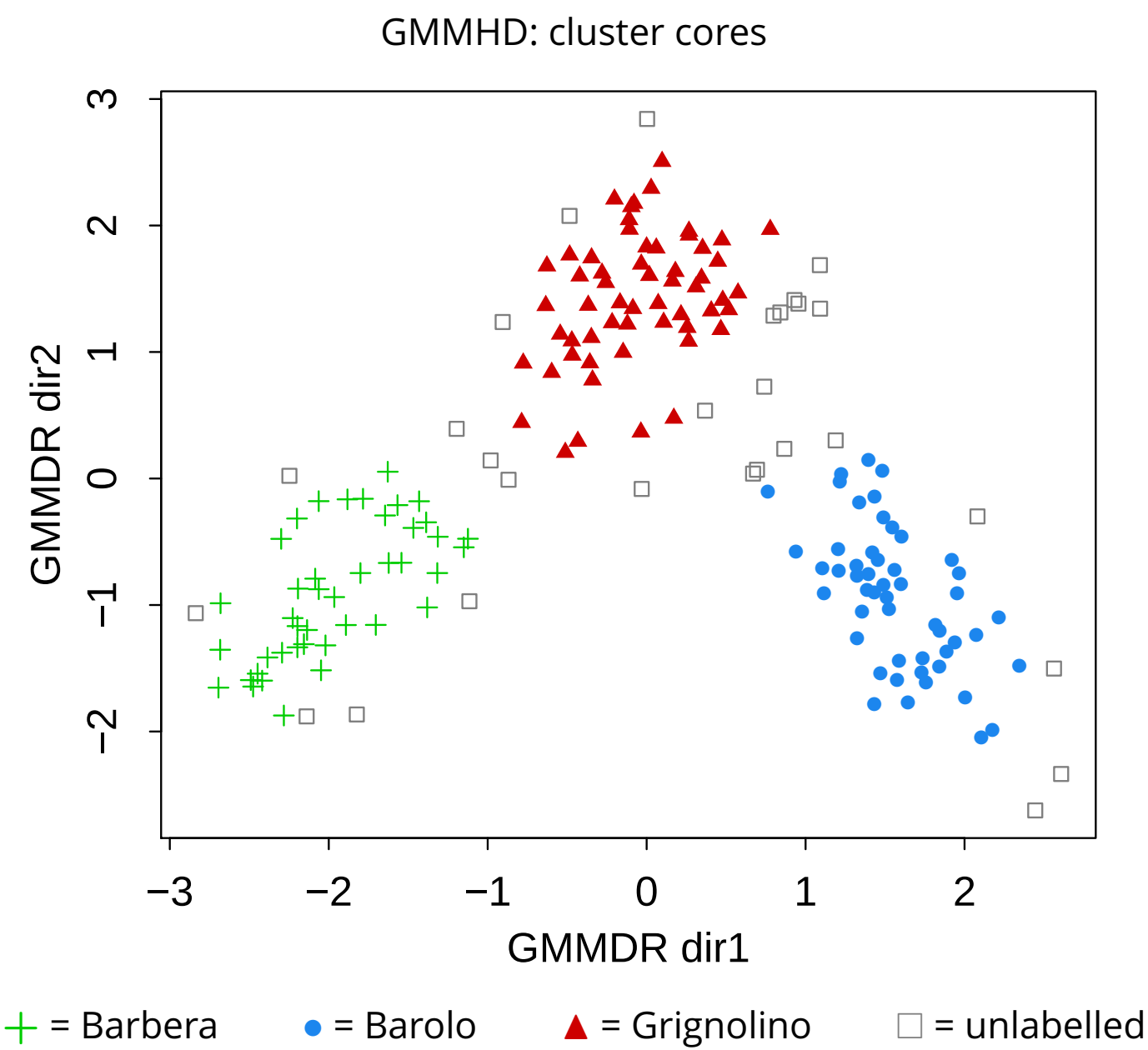
1	2	3	<NA>
51	57	42	28

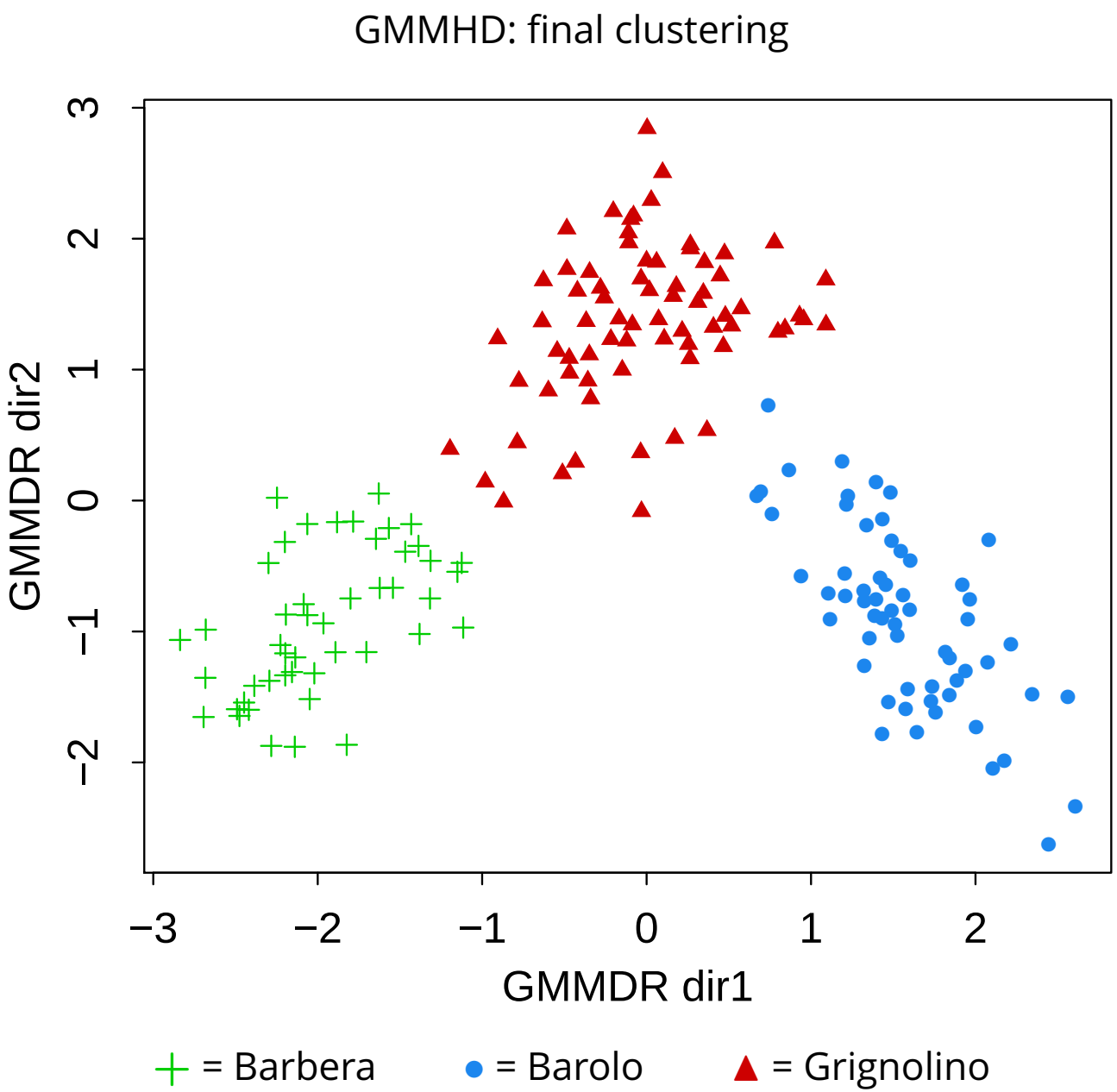
Final clustering:

1	2	3
60	71	47

group	cluster 1	2	3
Barbera	0	1	47
Barolo	59	0	0
Grignolino	1	70	0

AdjRandIndex = 0.9651


GMMHD: cluster cores
GMMDR dir2
GMMDR dir1
3
2
1
0
−1
−2
−3
−2
−1
0
1
2
= Barbera
= Barolo
= Grignolino
= unlabelled


GMMHD: final clustering
GMMDR dir2
GMMDR dir1
3
2
1
0
−1
−2
−3
−2
−1
0
1
2
= Barbera
= Barolo
= Grignolino

Unimodal skewed data

- ✓ Reasonable clustering methods should not only be able to recognise the presence of homogeneous groups in the data, but also to detect situations where there is no evidence of clusters.
- ✓ Number of clusters selected in 100 samples of size $n = 200$ from p-dimensional independent $\chi^2(10)$ distributions

	$p = 2$				$p = 5$				$p = 10$			
Number of clusters	1	2	3	4+	1	2	3	4+	1	2	3	4+
GMM	5	65	30	0	22	76	1	1	68	31	1	0
GMMHD	97	3	0	0	99	1	0	0	96	4	0	0

- ✓ Number of clusters selected in 100 samples of size $n = 200$ from p-dimensional skew-t unimodal distributions

	$p = 2$				$p = 5$				$p = 10$			
Number of clusters	1	2	3	4+	1	2	3	4+	1	2	3	4+
GMM	0	50	47	3	0	23	72	5	0	5	83	12
GMMHD	98	2	0	0	97	3	0	0	99	1	0	0

Some conclusions

The proposed approach appears to:

1. improve the identification of non-Gaussian clusters;
2. be able to identify clusters which cannot be obtained by combining mixture components (Baudry et al., 2010; Hennig, 2010);
3. improve over the approach based on nonparametric density estimation as the dimensionality increases.

Future works

- ✓ improve the computational requirements, in particular when n and/or p are large;
- ✓ investigate the case $n \ll p$;
- ✓ extend the approach to non-Gaussian model-based clustering (i.e. mixture of skew-normal, mixture of skew-t, ...);
- ✓ investigate how to deal with missing values.

The GMMHD methodology will soon be available in the R package `MCLUST`.

References

Azzalini, A. & Torelli, N. (2007). Clustering via nonparametric density estimation. *Statistics and Computing*, 17(1), 71–80.

Banfield, J. & Raftery, A. E. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics*, 49, 803–821.

Baudry, J. P., Raftery, A. E., Celeux, G., Lo, K., & Gottardo, R. (2010). Combining mixture components for clustering. *Journal of Computational and Graphical Statistics*, 19(2), 332–353.

Celeux, G. & Govaert, G. (1995). Gaussian parsimonious clustering models. *Pattern Recognition*, 28, 781–793.

Fraley, C. & Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97(458), 611–631.

Franczak, B. C., Browne, R. P., & McNicholas, P. D. (2013). Mixtures of shifted asymmetric laplace distributions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (PrePrints), 1.

Fukunaga, K. & Hostetler, L. (1975). The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Transactions on Information Theory*, 21(1), 32–40.

Hartigan, J. A. (1975). *Clustering Algorithms*. New York: John Wiley & Sons.

Hennig, C. (2010). Methods for merging gaussian mixture components. *Advances in Data Analysis and Classification*, 4(1), 3–34.

McLachlan, G. & Peel, D. (2000). *Finite Mixture Models*. New York: Wiley.

Menardi, G. & Azzalini, A. (2014). An advancement in clustering via nonparametric density estimation. *Statistics and Computing*, 24(5), 753–767.

Raftery, A. E. & Dean, N. (2006). Variable selection for model-based clustering. *Journal of the American Statistical Association*, 101(473), 168–178.

Schwartz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6, 31–38.

Scrucca, L. (2010). Dimension reduction for model-based clustering. *Statistics and Computing*, 20(4), 471–484.

Scrucca, L. (2014). Graphical tools for model-based mixture discriminant analysis. *Advances in Data Analysis and Classification*, 8(2), 147–165.

Stuetzle, W. (2003). Estimating the cluster tree of a density by analyzing the minimal spanning tree of a sample. *Journal of Classification*, 20(1), 25–47.

Stuetzle, W. & Nugent, R. (2010). A generalized single linkage method for estimating the cluster tree of a density. *Journal of Computational and Graphical Statistics*, 19(2), 397–418.

Wong, M. & Lane, T. (1983). A kth nearest neighbour clustering procedure. *Journal of the Royal Statistical Society. Series B*, 45, 362–368.

Zhu, X. & Goldberg, A. B. (2009). Introduction to semi-supervised learning. *Synthesis lectures on artificial intelligence and machine learning*, 3(1), 1–130.