Measurement of quality in cluster analysis

Christian Hennig

July 24, 2013

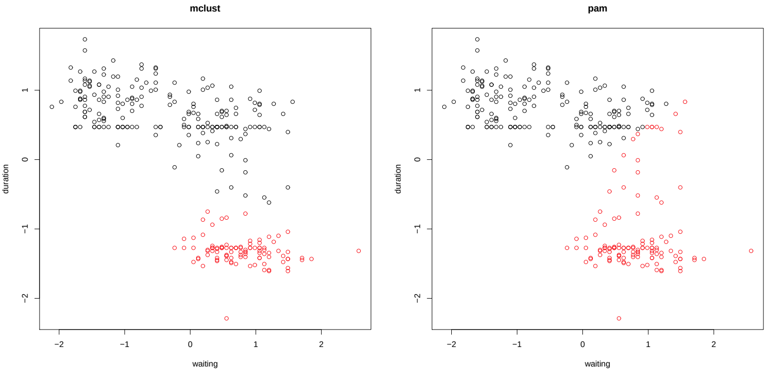
1. Introduction

IFCS task force for **cluster benchmarking**
(Nema Dean, Iven van Mechelen, Fritz Leisch, Doug Steinley, Bernd Bischl, Isabelle Guyon, Christian Hennig)

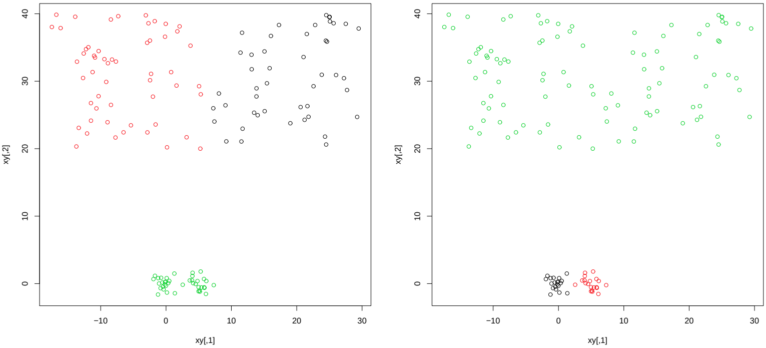
Data repository for systematic comparison of quality of different cluster analysis algorithms

In this presentation: compare quality of clusterings based on clustering and data alone, without reference to known truth.

Which clustering is better?
(Old faithful geyser data)



Which clustering is better?



Which tower is better?

Why datasets without known truth?

Benchmarking approaches:

- Real datasets with known classes
- Simulated datasets from mixture distributions
- Datasets with intuitive classes *by fiat*
- Real datasets without known classes

Why datasets without known truth?

Benchmarking approaches:

- Real datasets with known classes
- Simulated datasets from mixture distributions
- Datasets with intuitive classes *by fiat*
- Real datasets without known classes

Misclassification rates or Rand index
are (more or less) straightforward.

So why use datasets without known truth?

What's wrong with knowing the truth?

Disclaimer: knowing the truth is not evil.
There is definitely a role for datasets
with known truth in cluster benchmarking.

What's wrong with knowing the truth?

Disclaimer: knowing the truth is not evil.
There is definitely a role for datasets
with known truth in cluster benchmarking.

Measuring cluster quality
"ignoring" the truth can be of use
even if truth is known.
(May explain which truths a method can discover.)

What's wrong with knowing the truth?

But. . .

- In datasets with known classes
  clustering is not of real scientific interest.
  (Or one may want to find *different* clusterings.)
  Deviate systematically from real clustering problems.

What's wrong with knowing the truth?

But. . .

- In datasets with known classes
  clustering is not of real scientific interest.
  (Or one may want to find *different* clusterings.)
  Deviate systematically from real clustering problems.
- The fact that we know certain true classes
  doesn't preclude other legitimate/"true" clusterings.

What's wrong with knowing the truth?

But. . .

- In datasets with known classes
  clustering is not of real scientific interest.
  (Or one may want to find *different* clusterings.)
  Deviate systematically from real clustering problems.
- The fact that we know certain true classes
  doesn't preclude other legitimate/"true" clusterings.
- Classes in supervised classification problems
  may not qualify as data analytic clusters.

What's wrong with knowing the truth?

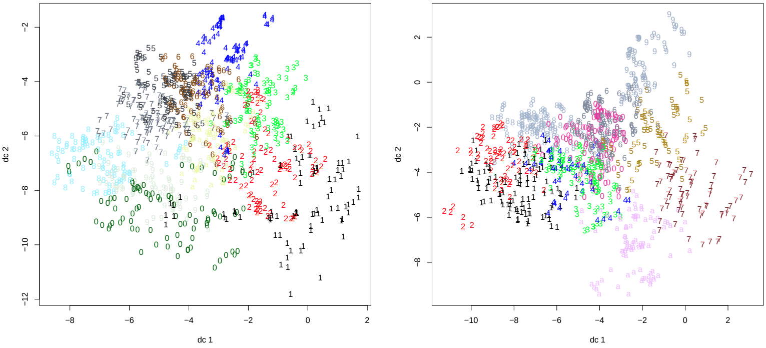
But. . .

- In datasets with known classes
  clustering is not of real scientific interest.
  (Or one may want to find *different* clusterings.)
  Deviate systematically from real clustering problems.
- The fact that we know certain true classes
  doesn't preclude other legitimate/"true" clusterings.
- Classes in supervised classification problems
  may not qualify as data analytic clusters.

So there could be better truths than the known one.

Which clustering is better?

(10-d vowel data; Hastie, Tibshirani and Friedman ESL)



What's wrong with knowing the truth?

- Identification mixture components/clusters is problematic.
- Mixture of two components may be unimodal.

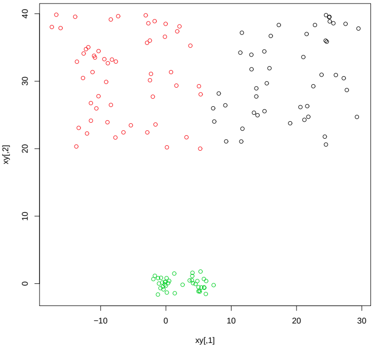
What's wrong with knowing the truth?

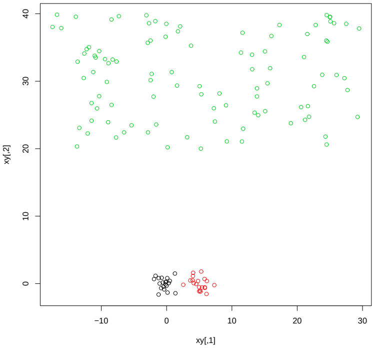
- Identification mixture components/clusters is problematic.
- Mixture of two components may be unimodal.
- Observations in tails may rather be outliers than cluster members (t-distributions).

What's wrong with knowing the truth?

- Identification mixture components/clusters is problematic.
- Mixture of two components may be unimodal.
- Observations in tails may rather be outliers than cluster members (t-distributions).
- Clustering aims may deviate from finding intuitive clusters or mixture components.

Which clustering is better?





2. Basic thoughts

There is a range of **cluster validation indexes** measuring clustering quality, such as
Average silhouette width (ASW)
(Kaufman and Rouseeuw 1990)

$sw(i, \mathcal{C}) = \frac{b(i,\mathcal{C}) - a(i,\mathcal{C})}{\max(a(i,\mathcal{C}), b(i,\mathcal{C}))}$,

$$a(i, \mathcal{C}) = \frac{1}{|C_j| - 1} \sum_{x \in C_j} d(x_i, x), \ b(i, \mathcal{C}) = \min_{x_i \notin C_l} \frac{1}{|C_l|} \sum_{x \in C_l} d(x_i, x).$$

Maximum average $sw \Rightarrow$ good $\mathcal{C}$.

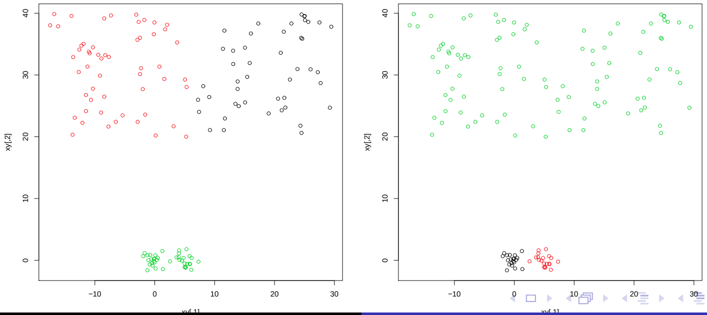
Most such indexes balance within-cluster homogeneity against between-cluster separation.

“One size fits it all”-approach.

Most such indexes balance within-cluster homogeneity against between-cluster separation.

“One size fits it all”-approach.

Homogeneity will always dominate here:



General philosophy

There are various different aims of clustering.
Depending on application,
these aims carry different weights.

General philosophy

There are various different aims of clustering.
Depending on application,
these aims carry different weights.

Measure them separately to characterise
what a method does best,
instead of producing a single ranking.

General philosophy

There are various different aims of clustering.
Depending on application,
these aims carry different weights.

Measure them separately to characterise
what a method does best,
instead of producing a single ranking.

Can piece together overall quality
as weighted mean of separate statistics.

Typical clustering aims

- Between-cluster separation

Typical clustering aims

- Between-cluster separation
- Within-cluster homogeneity (low distances)

Typical clustering aims

- Between-cluster separation
- Within-cluster homogeneity (low distances)
- Within-cluster homogeneous distributional shape

Typical clustering aims

- Between-cluster separation
- Within-cluster homogeneity (low distances)
- Within-cluster homogeneous distributional shape
- Good representation of data by centroids

Typical clustering aims

- Between-cluster separation
- Within-cluster homogeneity (low distances)
- Within-cluster homogeneous distributional shape
- Good representation of data by centroids
- Good representation of dissimilarity
  by clustering-induced metric

Typical clustering aims

- Between-cluster separation
- Within-cluster homogeneity (low distances)
- Within-cluster homogeneous distributional shape
- Good representation of data by centroids
- Good representation of dissimilarity
  by clustering-induced metric
- Clusters are regions of high density
  without within-cluster gaps

Typical clustering aims

- Between-cluster separation
- Within-cluster homogeneity (low distances)
- Within-cluster homogeneous distributional shape
- Good representation of data by centroids
- Good representation of dissimilarity
  by clustering-induced metric
- Clusters are regions of high density
  without within-cluster gaps
- Uniform cluster sizes

Typical clustering aims

- Between-cluster separation
- Within-cluster homogeneity (low distances)
- Within-cluster homogeneous distributional shape
- Good representation of data by centroids
- Good representation of dissimilarity by clustering-induced metric
- Clusters are regions of high density without within-cluster gaps
- Uniform cluster sizes
- Stability (requires knowledge of method)

E.g., pattern recognition in images requires separation,

E.g., pattern recognition in images
requires separation,

clustering for information reduction requires
good representation by centroids,

E.g., pattern recognition in images
requires separation,

clustering for information reduction requires
good representation by centroids,

groups in social network analysis shouldn't have
large within-cluster gaps,

E.g., pattern recognition in images
requires separation,

clustering for information reduction requires
good representation by centroids,

groups in social network analysis shouldn't have
large within-cluster gaps,

underlying "true" classes (biological species)
may cause homogeneous distributional shapes.

3. Cluster quality statistics

Aim: measure all that's of interest
by statistics in $[0, 1]$ (1 is good)
(so that different statistics are comparable
and weighted means make sense).

3. Cluster quality statistics

Aim: measure all that's of interest
by statistics in $[0, 1]$ (1 is good)
(so that different statistics are comparable
and weighted means make sense).

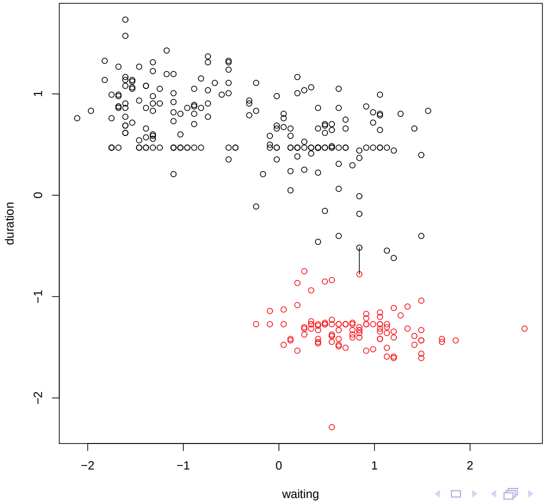
Principle of direct interpretation:
Aim at *translating* requirements directly into formulae;
that's not optimisation, not estimation of any "truth".

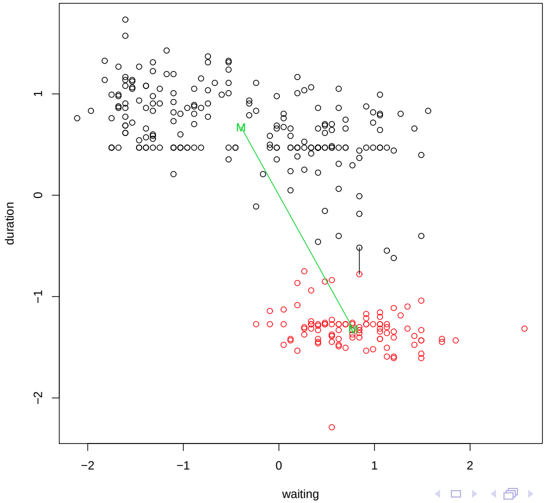
Warning: requires bold subjective tuning decisions.

And it's work in progress.

Measuring between-cluster separation

$\exists$ several ways measuring separation (as for other aims).
Straightforward: min distance between any two clusters, or distance between centroids (e.g., k-means).


duration
waiting
−2
−1
0
1
2
−2
−1
0
1

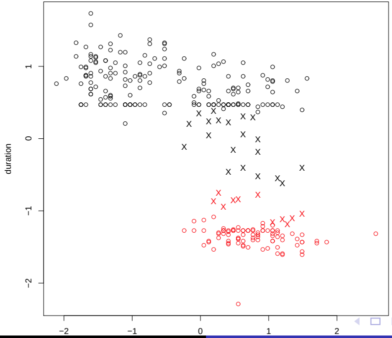

duration
waiting
M
M
−2
−1
0
1
2

Measuring between-cluster separation

∃ several ways measuring separation (as for other aims).
Straightforward: min distance between any two clusters,
or distance between centroids (e.g., k-means).

These measure quite different concepts of separation.
(min distance relies on only two points;
centroid distance ignores what goes on at border.)

***p*-separation index**:
More stable version of "min distance":
Average distance to nearest point in different cluster for
$p = 10\%$ "border" points in any cluster.
(ASW averages 100% to all in neighbouring cluster.)



p-**stability index**:
Average distance to nearest point in different cluster for $p = 10\%$ “border” points in any cluster.

Problems: choice of p, standardisation.

***p*-stability index**:
Average distance to nearest point in different cluster for
$p = 10\%$ "border" points in any cluster.

Problems: choice of p, standardisation.

May standardise by maximum distance;
range then is $[0, 1]$, but values may be very small,
max distance may be outlying,
implicit downweighting if used in
overall quality weighted mean.

Could use average/median distance etc. and bound by 1 (“separation larger ave distance $\Rightarrow$ perfect”).

Could use average/median distance etc. and bound by 1 (“separation larger ave distance $\Rightarrow$ perfect”).

Probably not fully satisfactory.

May use nonlinear transformation to $[0, 1]$
pronouncing differences between lower values,
taking into account whether
Max distance $>>$ ave/median distance.

Could use average/median distance etc. and bound by 1 ("separation larger ave distance $\Rightarrow$ perfect").

Probably not fully satisfactory.

May use nonlinear transformation to $[0, 1]$
pronouncing differences between lower values,
taking into account whether
Max distance $>>$ ave/median distance.

Stick to max distance standardisation here.

Could use average/median distance etc. and bound by 1 ("separation larger ave distance $\Rightarrow$ perfect").

Probably not fully satisfactory.

May use nonlinear transformation to $[0, 1]$
pronouncing differences between lower values,
taking into account whether
Max distance $>>$ ave/median distance.

Stick to max distance standardisation here.

$p = 0.1$ intuitive; sensitivity?

Alternative concept:
Distance-based knn density index

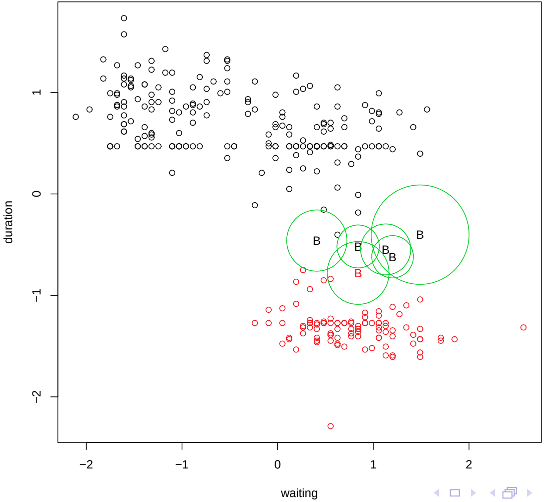
Measures whether border points have lowest density, highest density is within clusters i.

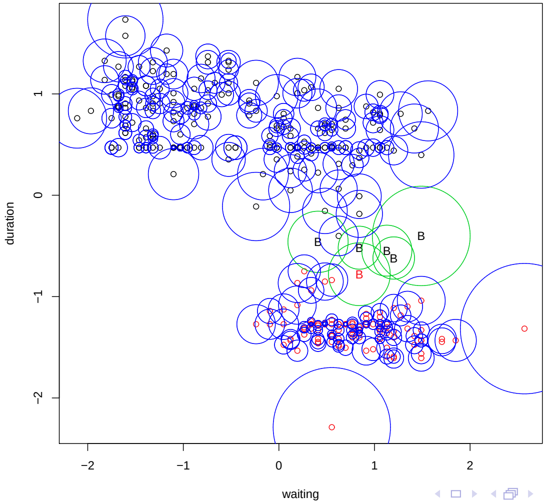
Alternative concept:
Distance-based knn density index

Measures whether border points have lowest density, highest density is within clusters i.

"Border points" here: n_i^B points that have points from other clusters among $k = 4$-nearest neighbours, n_i^I interior points.

Pointwise density: $k/(2*$mean distance to k-nn).


duration
waiting
1
0
−1
−2
−2
−1
0
1
2
B
B
B
B
B
B


duration
waiting
1
0
−1
−2
−2
−1
0
1
2
B
B
B
B
B
B

Clusterwise density index r_i^*:
(mean border density)/(mean interior density),
0 if $n_i^B = 0$, 1 if $n_i^I = 0$.

Aggregation ($\in [0, 1]$):

$$I_D = 1 - ((1-q)r_1 + qr_2)1((1-q)r_1 + qr_2 \leq 1),$$
$$r_1 = \sum w_i r_i^*,\ r_2 = \frac{\bar{b}}{\bar{i}},$$
$$q = 0.5 - \left|\frac{n - \sum n_i^I}{n} - 0.5\right|,\ w_i = \frac{n_i^I}{\sum n_i^I},$$

Overall: $\bar{b}$ mean border density, $\bar{i}$ mean interior density.

Aggregation ($\in [0,1]$):

$$I_D = 1 - ((1-q)r_1 + qr_2)1((1-q)r_1 + qr_2 \leq 1),$$
$$r_1 = \sum w_i r_i^*,\ r_2 = \frac{\bar{b}}{\bar{\bar{i}}},$$
$$q = 0.5 - \left|\frac{n - \sum n_i^l}{n} - 0.5\right|,\ w_i = \frac{n_i^l}{\sum n_i^l},$$

Idea: r_1 measures "cluster-relative" density ratio,
r_2 overall.

Both of interest, but for $\sum n_i^l \approx n$ or 0,
one side of r_2 relies on very weak information.

Aggregation ($\in [0,1]$):

$$I_D = 1 - ((1-q)r_1 + qr_2)1((1-q)r_1 + qr_2 \leq 1),$$
$$r_1 = \sum w_i r_i^*, \; r_2 = \frac{\bar{b}}{\bar{\bar{i}}},$$
$$q = 0.5 - \left|\frac{n - \sum n_i^l}{n} - 0.5\right|, \; w_i = \frac{n_i^l}{\sum n_i^l},$$

Idea: r_1 measures "cluster-relative" density ratio,
r_2 overall.

Both of interest, but for $\sum n_i^l \approx n$ or 0,
one side of r_2 relies on very weak information.

Although r_1 downweights clusters with n_i^l small,
outlier one-point clusters still produce too good I_D.

Other statistics

- Within-cluster average distance

Other statistics

- Within-cluster average distance
- Aggregated within-cluster similarity (Kolmogorov etc.) to normal/uniform

Other statistics

- Within-cluster average distance
- Aggregated within-cluster similarity (Kolmogorov etc.) to normal/uniform
- Within-cluster (squared) distance to centroid

Other statistics

- Within-cluster average distance
- Aggregated within-cluster similarity (Kolmogorov etc.) to normal/uniform
- Within-cluster (squared) distance to centroid
- ρ(distance, cluster induced distance) (Hubert's Γ)

Other statistics

- Within-cluster average distance
- Aggregated within-cluster similarity (Kolmogorov etc.) to normal/uniform
- Within-cluster (squared) distance to centroid
- ρ(distance, cluster induced distance) (Hubert's Γ)
- Entropy of cluster sizes

Other statistics

- Within-cluster average distance
- Aggregated within-cluster similarity (Kolmogorov etc.) to normal/uniform
- Within-cluster (squared) distance to centroid
- ρ(distance, cluster induced distance) (Hubert's Γ)
- Entropy of cluster sizes
- Within-cluster nearest neighbour distances coefficient of variation

Other statistics

- Within-cluster average distance
- Aggregated within-cluster similarity (Kolmogorov etc.) to normal/uniform
- Within-cluster (squared) distance to centroid
- ρ(distance, cluster induced distance) (Hubert's Γ)
- Entropy of cluster sizes
- Within-cluster nearest neighbour distances coefficient of variation
- Average largest within-cluster gap

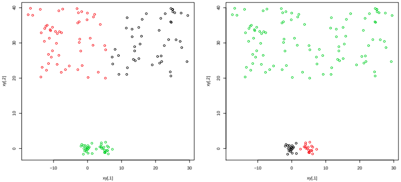
All need standardisation/transformation.

Most are dissimilarity-based,
allow flexible use with non-Euclidean data,
given meaningful distance measure.

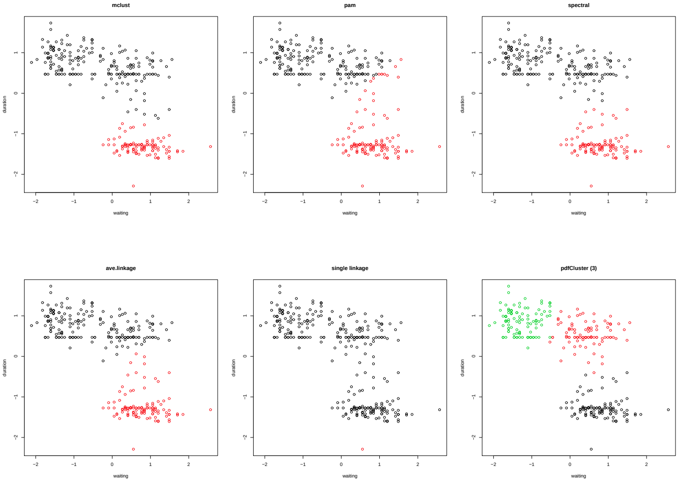
Data set submission to benchmarking repository requires filling in questionnaire, e.g.

- Should clusters be similar or dissimilar in size?
- Are there requirements on what should be the unifying/common ground for elements to belong to the same cluster? Small within-cluster dissimilarities (and, if yes, in which respect)?
- Are there requirements on what should be the discriminating ground for elements to belong to different clusters? Large between-cluster dissimilarities (and, if yes, in which respect)? Separation (and, if yes, of which kind)? Other (and, if yes, what form do these requirements take)?
- Are there requirements on the between-cluster heterogeneity, that is, the structure of between-cluster differences (e.g., should lie in low-dimensional space, other)?
- (Some other; not all yet formalised by indexes)
- Please indicate the importance of those criteria selected by filling in a numerical weight.

4. Examples



	3-means	mclust-3
ave within	0.811	0.643
sep index	0.163	0.306
density index	0.460	0.876
within gap	0.927	0.949


mclust
pam
spectral
ave.linkage
single linkage
pdfCluster (3)
duration
waiting
−2
−1
0
1
2

	mclust	pam	spect	ave.l	sing.l	comp.l	pdf3
ave within	0.783	0.797	0.792	0.794	0.666	0.779	0.875
sep index	0.127	0.045	0.127	0.096	0.175	0.103	0.065
density	0.910	0.733	0.864	0.903	0.969	0.874	0.719
gap	0.888	0.891	0.891	0.891	0.929	0.891	0.906
coef var	0.541	0.567	0.554	0.564	0.573	0.545	0.554
gamma	0.679	0.708	0.709	0.711	0.064	0.664	0.767
normality	0.880	0.838	0.854	0.841	0.786	0.882	0.856
entropy	0.923	0.974	0.941	0.952	0.023	0.913	0.999

Weighthed mean:
full weight: ave within, sep index
0.8 weight: entropy
half weight: within nn cov, gap, min separation,
density index, hubert gamma, normality, uniformity

pdf3	spect	ave.l	mclust	kmeans	comp.l	pam	sing.l
0.624	0.622	0.622	0.619	0.618	0.610	0.601	0.460

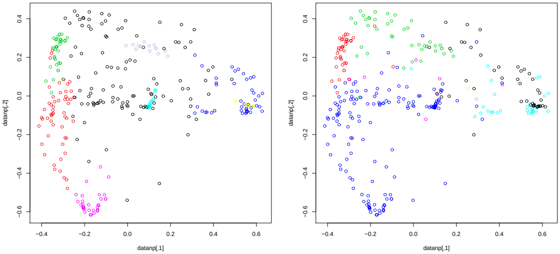
- Problem with pam captured.

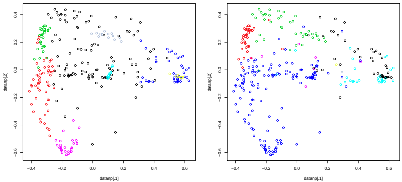
- Problem with pam captured.
- Single linkage: useless (entropy 0.023; one-point cluster), good values in some indexes (careful!), bad in others.

- Problem with pam captured.
- Single linkage: useless (entropy 0.023; one-point cluster), good values in some indexes (careful!), bad in others.
- Comparison 2-cluster vs. 3-cluster (pdfCluster):
  individual indexes unfair;
  ave within better, separation worse with larger k (etc.)
  Depends on proper weighting.
  Could add parsimony index.

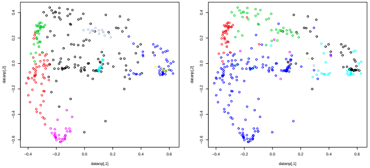
- Problem with pam captured.
- Single linkage: useless (entropy 0.023; one-point cluster), good values in some indexes (careful!), bad in others.
- Comparison 2-cluster vs. 3-cluster (pdfCluster):
  individual indexes unfair;
  ave within better, separation worse with larger k (etc.)
  Depends on proper weighting.
  Could add parsimony index.
- mclust not best in normality,
  ave.l not best in ave within!
  Individual indexes may favour certain methods,
  but not as obvious as it seems.

European land snails data (Hausdorf, Hennig 2003,2006)
Presence-absence (0-1) data for species in regions;
“geographical Kulczynski dissimilarity”;
clustering for “biotic elements” (natural history),
originally clustered with mclust after MDS.
Can compare with distance-based.

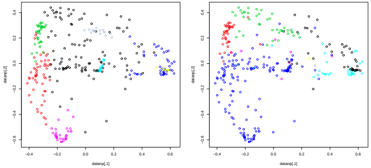




	mclust	ave.l
ave within	0.619	0.645
gap	0.766	0.771
density index	0.503	0.852
sep index	0.055	0.126
entropy	0.929	0.717
normality (MDS)	0.805	0.781
uniformity (MDS)	0.393	0.302

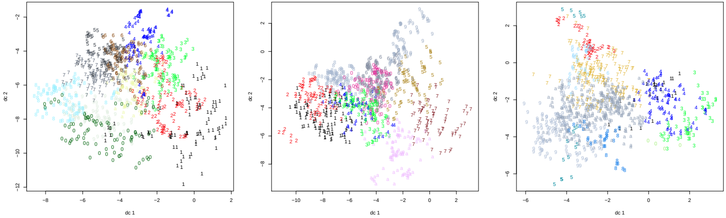


mclust only better in entropy
and MDS-distribution based indexes.



mclust only better in entropy
and MDS-distribution based indexes.

Perturbed by “noise cluster”;
better cluster with “noise component”.
How to use indexes with unclustered data?
Could just ignore them but that gives
clustering with “noise” unfair advantage.



	true	11-means	spectral
ave within	0.691	0.734	0.692
sep index	0.069	0.093	0.130
hubert gamma	0.224	0.411	0.400
entropy	1.000	0.983	0.739
ARI	1.000	0.205	0.142

	true	11-means	spectral
ave within	0.691	0.734	0.692
sep index	0.069	0.093	0.130
hubert gamma	0.224	0.411	0.400
entropy	1.000	0.983	0.739
ARI	1.000	0.205	0.142

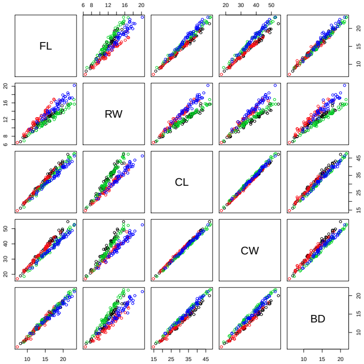
"true" no good clustering.
Ave within, entropy are *only* indexes positively correlated with ARI!

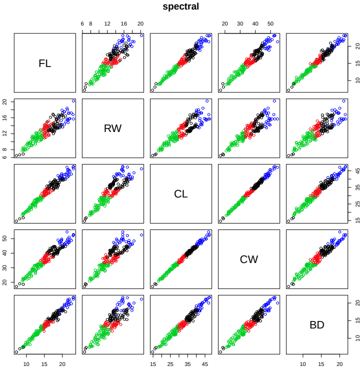
	true	11-means	spectral
ave within	0.691	0.734	0.692
sep index	0.069	0.093	0.130
hubert gamma	0.224	0.411	0.400
entropy	1.000	0.983	0.739
ARI	1.000	0.205	0.142

"true" no good clustering.
Ave within, entropy are *only* indexes
positively correlated with ARI!

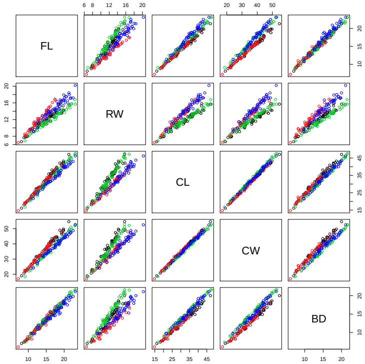
Good ARI needs good ave within and nothing else here.
Use to explain results in data with known classes.

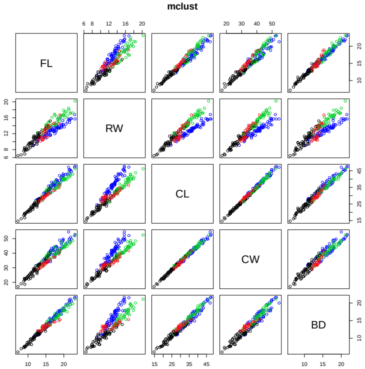
Crabs data (2 species, m/f):





Crabs data (2 species, m/f):





	true	mclust	spectral
ave within	0.761	0.828	0.908
density	0.167	0.027	0.246
hubert gamma	0.060	0.291	0.591
ARI	1.000	0.316	0.023

- "true" is worst according to most indexes.
  But there *is* a visible pattern!
- *All* indexes (except entropy) are
  negatively correlated with ARI.
- mclust has best ARI out of 8 methods
  but quite bad index values.

Indexes fail to capture what goes on here:-(

5. Discussion

- Clustering quality is multidimensional.

5. Discussion

- Clustering quality is multidimensional.
- Provide multidimensional evaluation, characterising a method's behaviour.

5. Discussion

- Clustering quality is multidimensional.
- Provide multidimensional evaluation, characterising a method's behaviour.
- Can aggregate criteria by weighted mean given well justified weights.

5. Discussion

- Clustering quality is multidimensional.
- Provide multidimensional evaluation, characterising a method's behaviour.
- Can aggregate criteria by weighted mean given well justified weights.
- Can use to explain performance in data with known truth

5. Discussion

- Clustering quality is multidimensional.
- Provide multidimensional evaluation, characterising a method's behaviour.
- Can aggregate criteria by weighted mean given well justified weights.
- Can use to explain performance in data with known truth
- Designers of new methods should specify what aspects of clustering they aim at, so that it can be tested.

Open problems

- Dependence on subjective tuning hurts (weights, k-nn, percentage, standardisation).

Open problems

- Dependence on subjective tuning hurts (weights, k-nn, percentage, standardisation).
- But it's honest; such decisions are needed to define a good clustering in practice.

Open problems

- Dependence on subjective tuning hurts (weights, k-nn, percentage, standardisation).
- But it's honest; such decisions are needed to define a good clustering in practice.
- Proper behaviour of criteria (standardisation, transformation, different numbers of clusters) for fair aggregation and comparability?

Open problems

- Dependence on subjective tuning hurts (weights, k-nn, percentage, standardisation).
- But it's honest; such decisions are needed to define a good clustering in practice.
- Proper behaviour of criteria (standardisation, transformation, different numbers of clusters) for fair aggregation and comparability?
- Index choice vs. method definition (average linkage not always optimal for ave. distance)

Open problems

- Dependence on subjective tuning hurts (weights, k-nn, percentage, standardisation).
- But it's honest; such decisions are needed to define a good clustering in practice.
- Proper behaviour of criteria (standardisation, transformation, different numbers of clusters) for fair aggregation and comparability?
- Index choice vs. method definition (average linkage not always optimal for ave. distance)
- With given weights, optimise quality?