Cluster validity indexes: calibration, aggregation, further issues

Christian Hennig

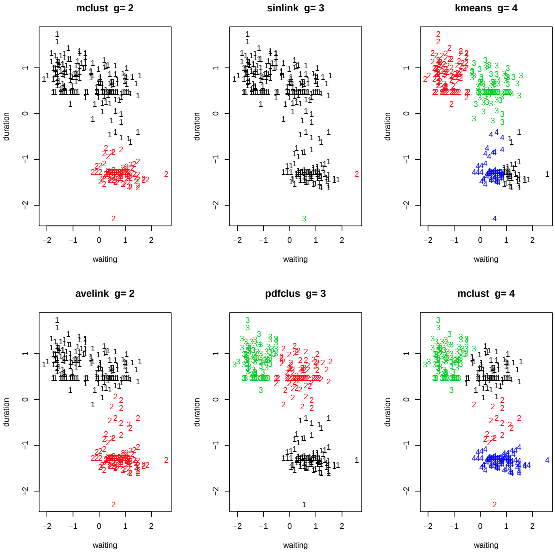
December 14, 2015

1. Introduction

Cluster validation: evaluation of the quality of a clustering, not knowing the truth.

Comparison of different clustering methods
(applied and for benchmarking),
parameters such as the number of clusters,
assessment and interpretation of a clustering.

Cluster validation has many aspects.
Different applications of clustering have different aims,
different aspects may be of interest,
different clusterings on the same data
may be seen as “true”.


mclust g= 2
sinlink g= 3
kmeans g= 4
avelink g= 2
pdfclus g= 3
mclust g= 4
duration
waiting

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
- Low variation of within-cluster densities

Typical clustering aims

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

Typical clustering aims

- Between-cluster separation
- Within-cluster homogeneity (low distances)
- Within-cluster homogeneous distributional shape
- Low variation of within-cluster densities
- 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
- Low variation of within-cluster densities
- 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
- Low variation of within-cluster densities
- 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
- Low variation of within-cluster densities
- 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

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.

2. A collection of validity indexes

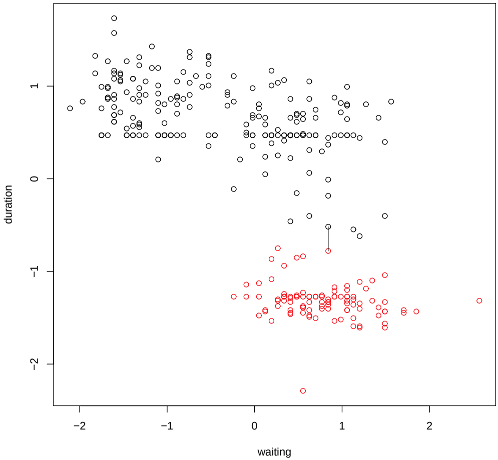
General approach: Measure different aspects of clustering by different statistics to give a *multivariate characterisation* of cluster validity.

Optimal clustering could be found by computing weighted average, according to relative importance of aspects in given application.

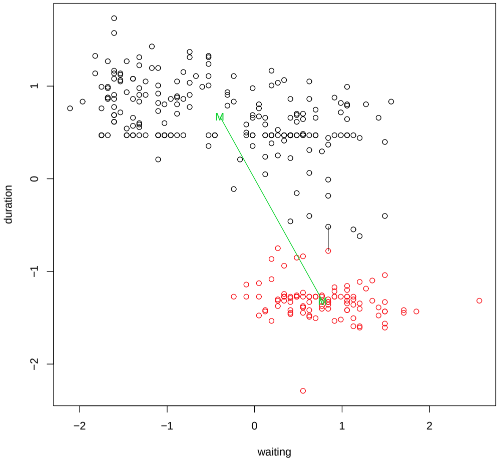
Typical clustering aims

- Between-cluster separation
- Within-cluster homogeneity (low distances)
- Within-cluster homogeneous distributional shape
- Low variation of within-cluster densities
- 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

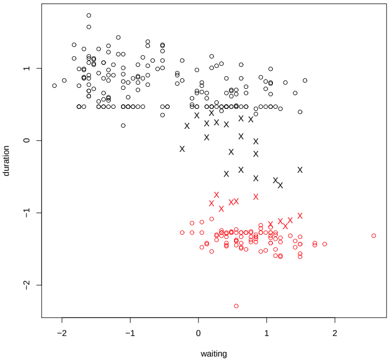
Measuring between-cluster separation



Measuring between-cluster separation



***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.



Measuring “density mountains vs. valleys”

Index that measures whether clusters correspond to “density mountains”, and whether “valleys” are between clusters.

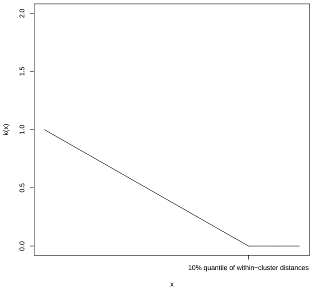
Prefer distance-based non-parametric index able to deal with any data format that allows distances to be computed.

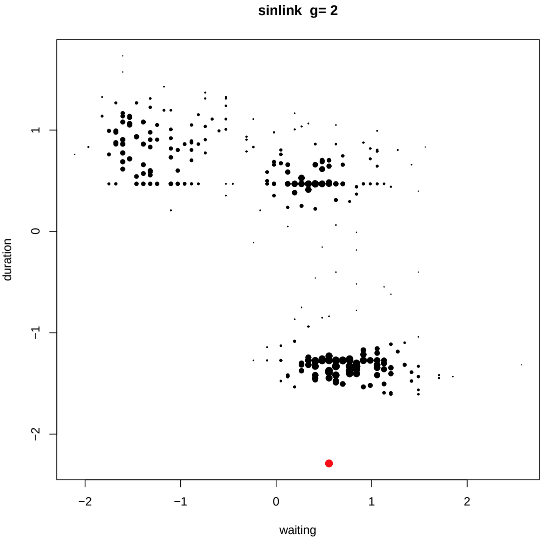
Two aspects:

(a) Density goes down from mode;
no gaps and valleys within clusters.

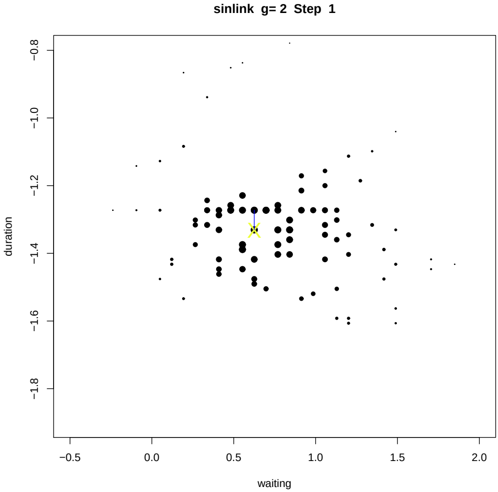
(b) Cluster borders are valleys;
they don't run through mountains.

Distance-based kernel density:

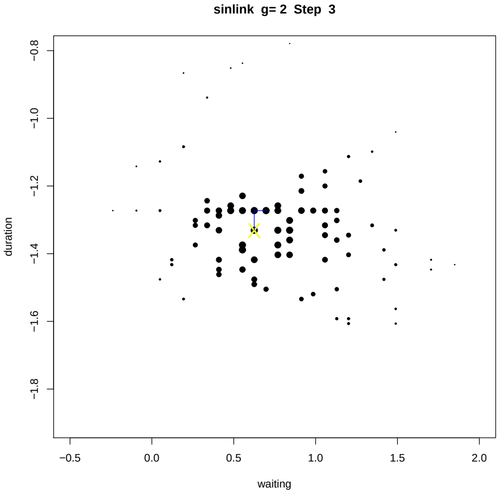



sinlink g= 2
duration
waiting
−2
−1
0
1
2
−2
−1
0
1

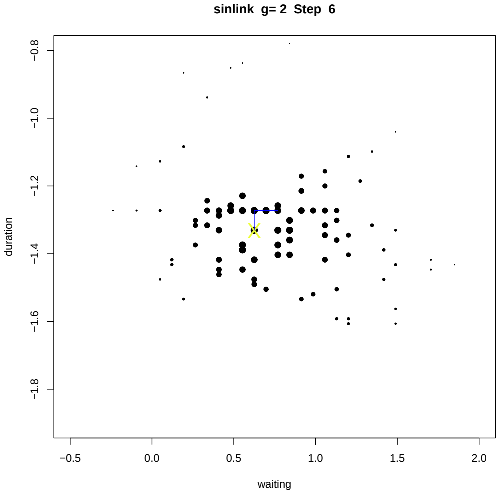
Start from cluster modes



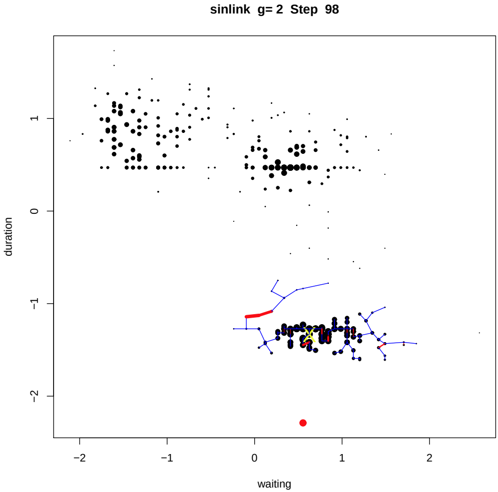
Connect closest point to cluster

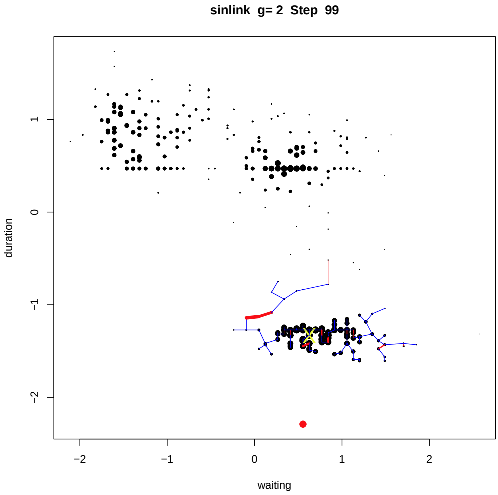


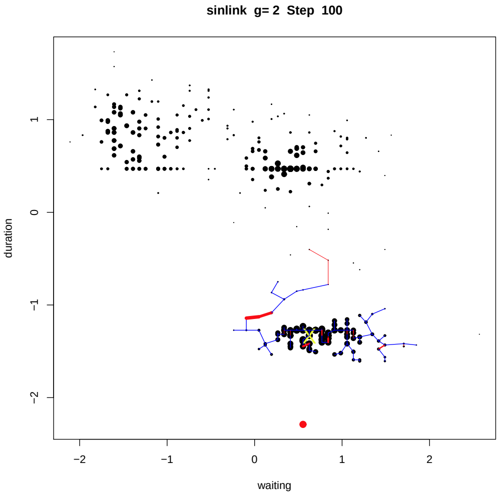
As long as density goes down, no penalty

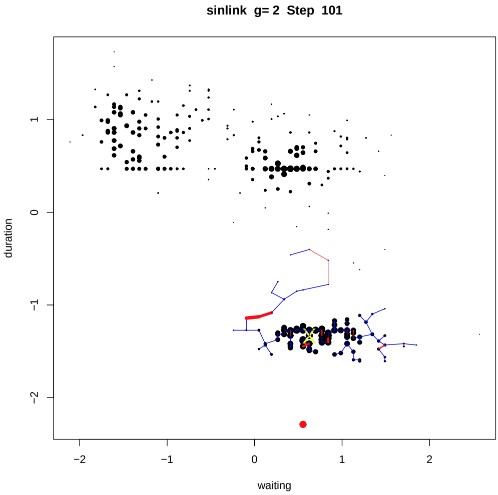


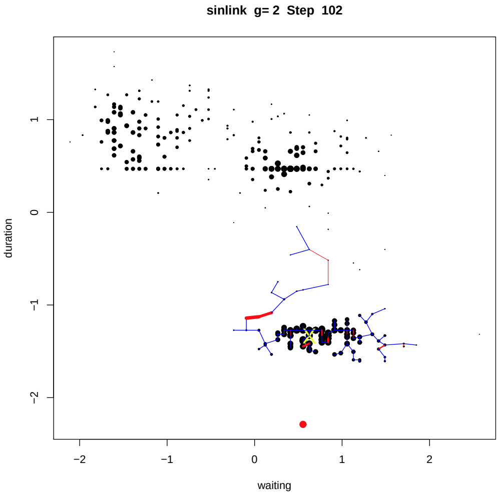
Penalty for density increase

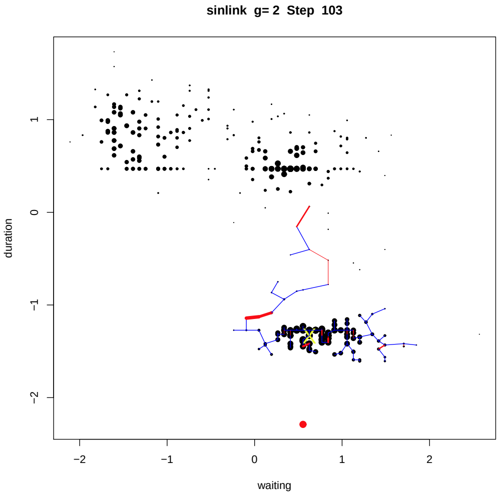


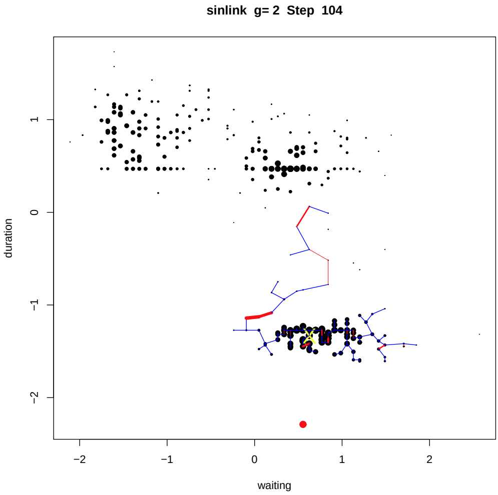

sinlink g= 2 Step 99
duration
waiting
−2
−1
0
1
2
−2
−1
0
1

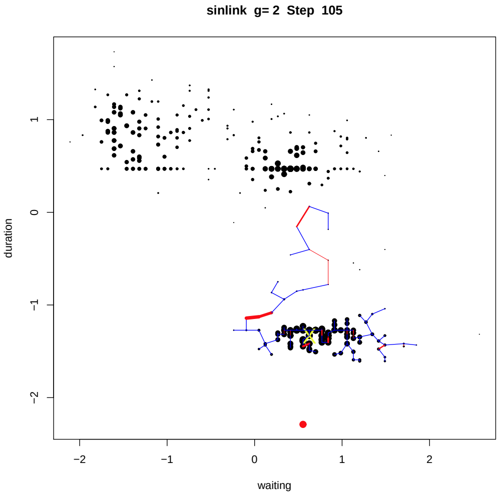

sinlink g= 2 Step 100
duration
waiting
−2
−1
0
1
2

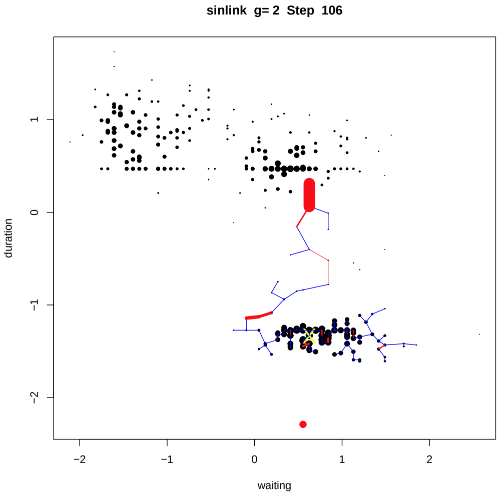

sinlink g= 2 Step 101
duration
waiting
−2
−1
0
1
2

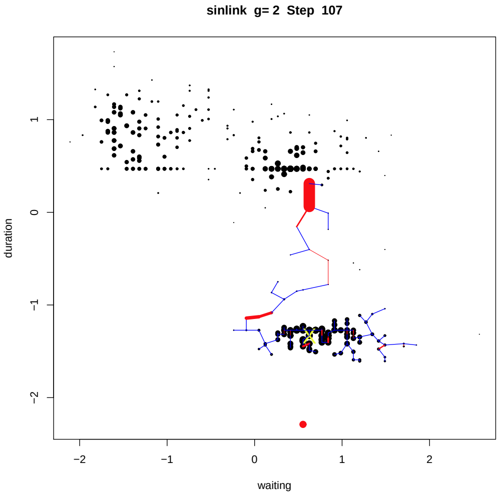

sinlink g= 2 Step 102
duration
waiting
−2
−1
0
1
2

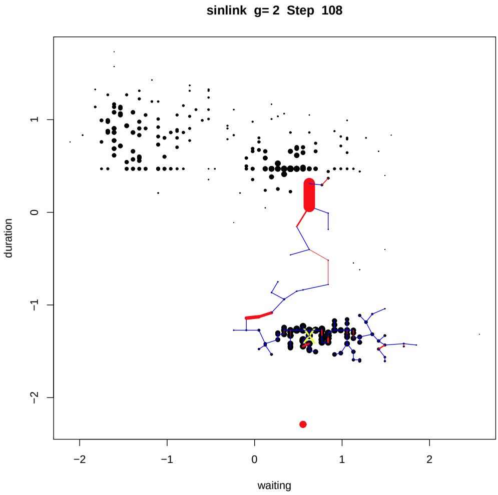

sinlink g= 2 Step 103
duration
waiting
−2
−1
0
1
2
−2
−1
0
1

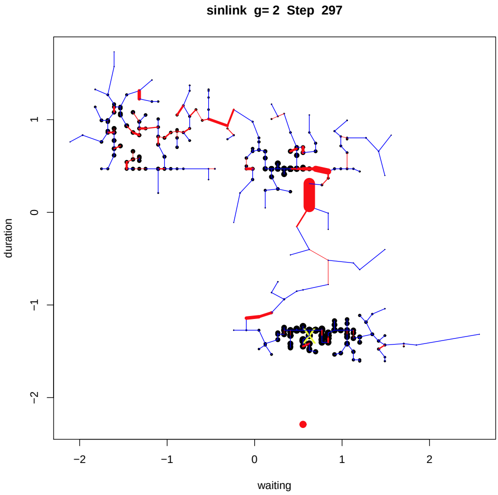

sinlink g= 2 Step 104
duration
waiting
1
0
−1
−2
−2
−1
0
1
2


sinlink g= 2 Step 105
duration
waiting
−2
−1
0
1
2

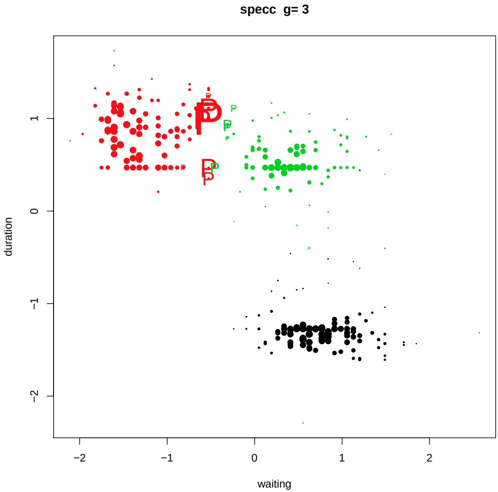

sinlink g= 2 Step 106
duration
waiting
1
0
−1
−2
−2
−1
0
1
2


sinlink g= 2 Step 107
duration
waiting
−2
−1
0
1
2


sinlink g= 2 Step 108
duration
waiting
−2
−1
0
1
2


sinlink g= 2 Step 297
duration
waiting
1
0
−1
−2
−2
−1
0
1
2

Add penalty density$*$density from other clusters

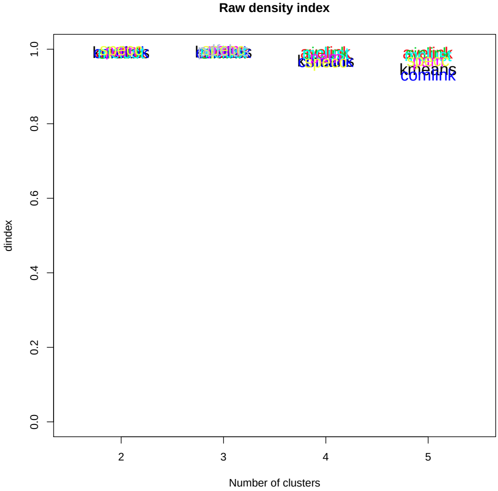


3. Index calibration and aggregation

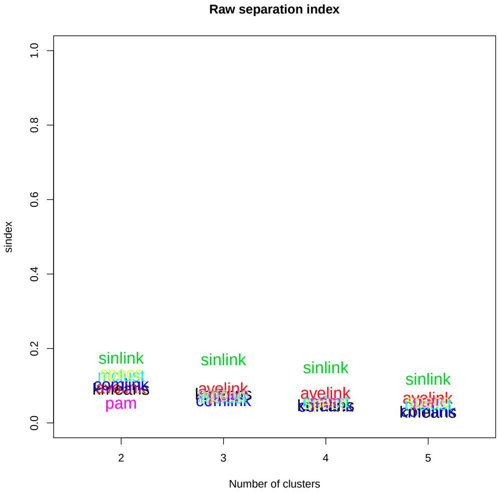
May want to aggregate indexes,
and to know whether differences are small or big.

Standardise all indexes to maximum 1 and “large is good”.
But this is not enough calibration.

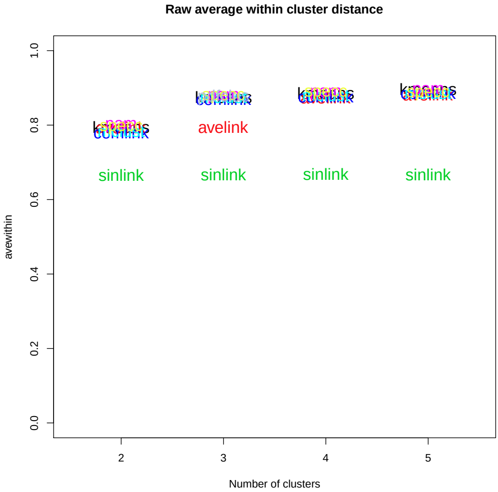
Are differences between methods meaningful?



Are differences between methods meaningful?



Are differences between methods meaningful?



Relate values to variability in clusterings
on given fixed data.

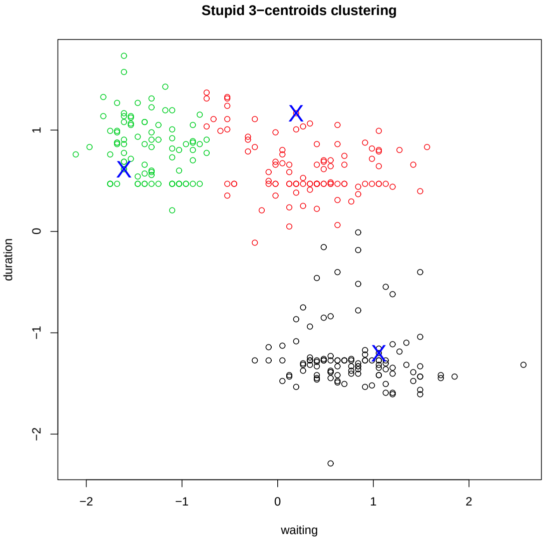
(Note difference to usual probability modelling.)

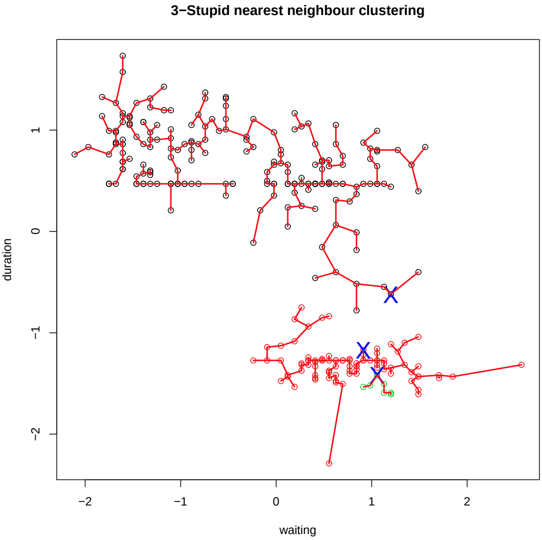
How to generate “random clusterings” on fixed data?

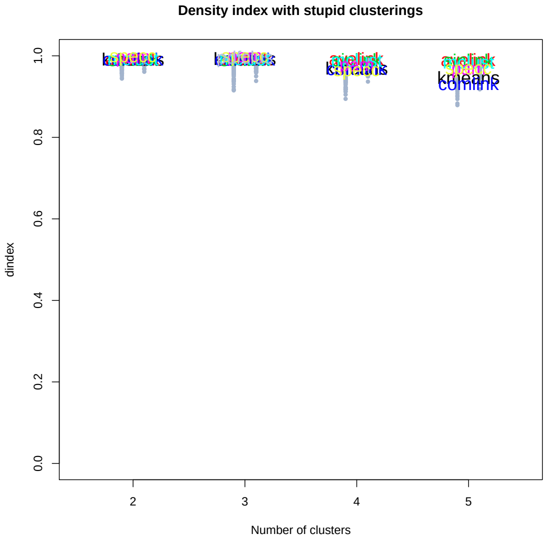
Generate random
"stupid k-centroids"
and *"stupid nearest neighbour"*-clusterings.

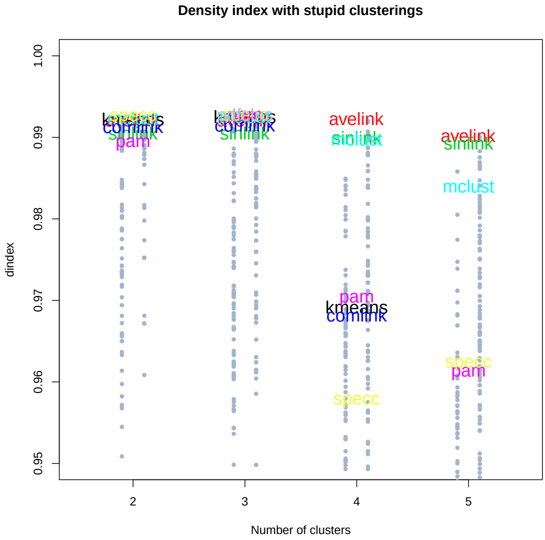
"k-centroid" methods produce "compact" clusters;
nearest neighbour/single linkage produces clusters
with flexible shapes.

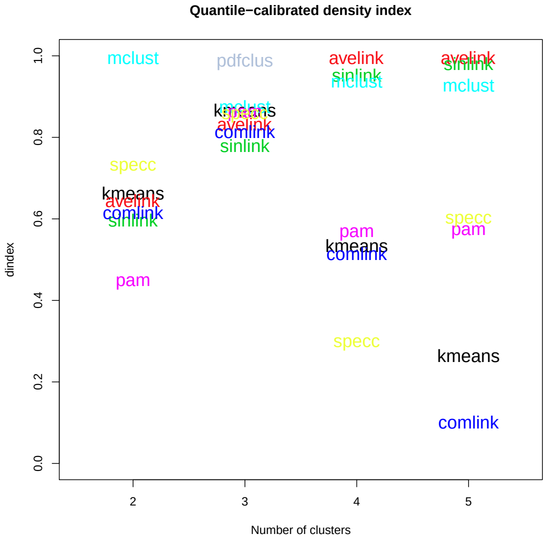
Use both to explore variability in clustering.

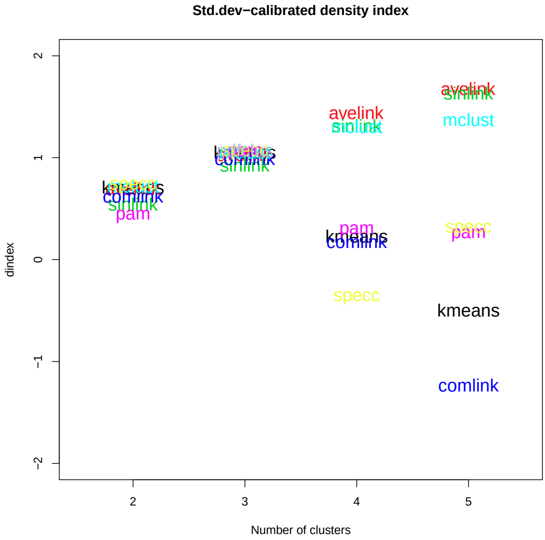

Stupid 3−centroids clustering
X
X
X
duration
1
0
−1
−2
−2
−1
0
1
2
waiting


3−Stupid nearest neighbour clustering
duration
waiting
−2
−1
0
1
2
X

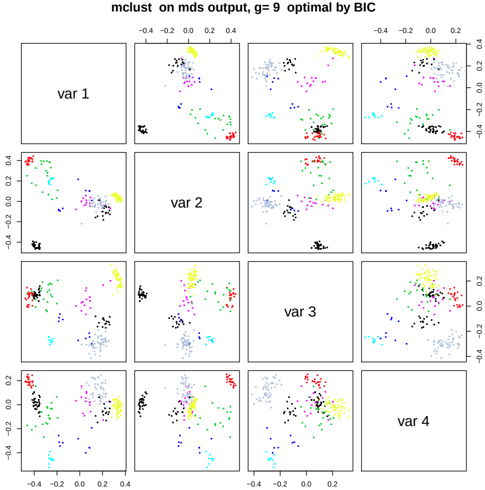

Density index with stupid clusterings
dindex
Number of clusters
0.0
0.2
0.4
0.6
0.8
1.0
2
3
4
5
kmeans
complink
averlink


Density index with stupid clusterings
dindex
1.00
0.99
0.98
0.97
0.96
0.95
avelink
sinlink
mclust
pam
kmeans
complink
specc
2
3
4
5
Number of clusters


Quantile−calibrated density index
dindex
Number of clusters
0.0
0.2
0.4
0.6
0.8
1.0
2
3
4
5
mclust
specc
kmeans
avelink
comlink
sinlink
pam
pdfclus
mclust
kmeans
specc
avelink
comlink
sinlink
avelink
sinlink
mclust
pam
kmeans
comlink
specc
avelink
sinlink
mclust
specc
pam
kmeans
comlink


Std.dev−calibrated density index
dindex
Number of clusters
2
3
4
5
−2
−1
0
1
2
avelink
sinlink
mclust
avelink
sinlink
mclust
kmeans
comlink
sinlink
pam
pam
kmeans
comlink
specc
pam
specc
kmeans
comlink

Trigona bees example for species delimitation



```
          st.dev    quantile
avewithin 2.447696  0.9951456 # best
variation 2.813598  0.9951456 # best
diameter  1.639204  0.8592233
gap       0.3729861 0.5048544
sindex    0.9757806 0.7912621
dindex    1.331795  0.9660194
pg        1.765785  0.9757282
withinss  2.061715  0.9902913
```

Aggregate index
useful for species delimitation from

pg	Individuals in same cluster if distance low.
gap	Within-species gaps shouldn't exist.
sindex	Species should be separated.
avewithin	Limited variation within species.

Ran single , average, complete linkage,
pam and mclust/mds on 2 to 12 clusters
$\Rightarrow$ 55 clusterings.

Ranking:

1. Complete linkage/12 - 6.61
2. Average linkage/11 - 6.13
3. Average linkage/10 - 6.08
4. Average linkage/12 - 5.95

12. mclust/MDS/9 - 5.56

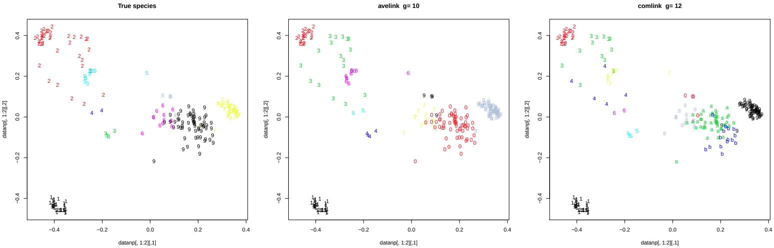
Truth is known here (9 species).
Adjusted Rand indexes:

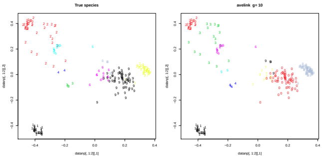
1. Complete linkage/12 - 6.61 - 0.851 (4)
2. Average linkage/11 - 6.13 - 0.944 (2)
3. Average linkage/10 - 6.08 - 0.951 (1)
4. Average linkage/12 - 5.95 - 0.940 (3)

12. mclust/MDS/9 - 5.56 - 0.832 (8)

4. Cluster-wise diagnosis

Many statistics give cluster-wise information.
Can use this to assess individual clusters.





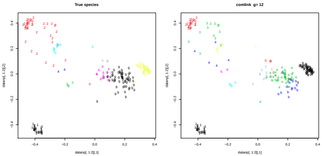
Cluster 3 is worst w.r.t. many statistics.

```
$diameter
 [1] 0.5000000 0.4090909 0.8181818 0.3846154 0.3750000 0.6250000 0.5769231
 [8] 0.6250000 0.5769231 0.7692308

$average.distance
 [1] 0.2907563 0.2648221 0.5805110 0.3461538 0.3750000 0.3679752 0.3979290
 [8] 0.2203210 0.3333333 0.4378032

$separation
 [1] 0.6666667 0.5000000 0.5000000 0.7083333 0.7083333 0.5833333 0.5000000
 [8] 0.2777778 0.4615385 0.2777778

$cwidegap
 [1] 0.2083333 0.2272727 0.5000000 0.3461538 0.3750000 0.4166667 0.4615385
 [8] 0.3333333 0.5000000 0.3461538
```



Cluster 11/12 are not separated; 3/4 not homogeneous.

```
$average.distance
 [1] 0.2907563 0.2648221 0.5553613 0.4772727 0.3461538 0.3750000 0.3679752
 [8] 0.3979290 0.2203210 0.3333333 0.3872155 0.3157814

$separation
 [1] 0.6666667 0.5000000 0.4545455 0.4545455 0.7083333 0.7083333 0.5833333
 [8] 0.5000000 0.2777778 0.4615385 0.3076923 0.2777778

$cwidegap
 [1] 0.2083333 0.2272727 0.5000000 0.5000000 0.3461538 0.3750000 0.4166667
 [8] 0.4615385 0.3333333 0.5000000 0.3461538 0.3076923

$dpenalty
 [1] 0.29268293 0.34482759 0.00000000 0.00000000 0.00000000 0.00000000
 [7] 0.00000000 0.39285714 0.57956209 0.00000000 0.53599131 0.08569501

$npenalty
 [1] 0.000000000 0.000000000 0.000000000 0.000000000 0.000000000 0.000000000
 [7] 0.000000000 0.000000000 0.008193015 0.000000000 0.042865962 0.087788505
```

Soon to come:
IFCS Cluster Benchmarking Repository
(Iven Van Mechelen, Nema Dean, Isabelle Guyon,
Anne-Laure Boulesteix, Doug Steinley, Friedrich Leisch,
Christian Hennig, Rainer Dangl)

This work is supported by EPSRC Grant EP/K033972/1.