Some thoughts on cluster benchmarking

Christian Hennig

November 8, 2011

The IFCS cluster benchmarking task force

The IFCS cluster benchmarking task force

Isn’t it what the AG DA-NK is doing anyway?

The IFCS cluster benchmarking task force

Isn’t it what the AG DA-NK is doing anyway?

Is there a “UCI”-approach to clustering?

The IFCS cluster benchmarking task force

Isn't it what the AG DA-NK is doing anyway?

Is there a "UCI"-approach to clustering?

Is the UCI-approach really good for *supervised* classification?

The IFCS cluster benchmarking task force

Isn't it what the AG DA-NK is doing anyway?

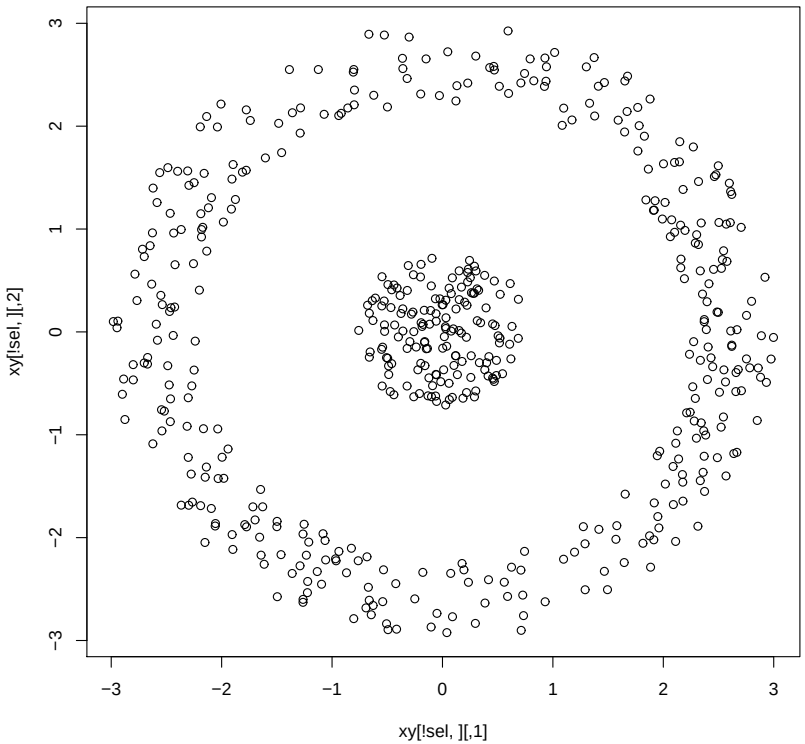
Is there a "UCI"-approach to clustering?

Is the UCI-approach really good for *supervised* classification?

Datasets with known classes deviate *systematically* from real clustering problems.

Stupid cluster benchmarking

"*k*-means cannot do this properly!"



Benchmarking approaches

- Real datasets with known classes
- Simulated datasets (from mixture distributions?)
- Real datasets *without* known classes

Benchmarking approaches

- Real datasets with known classes
- Simulated datasets (from mixture distributions?)
- Real datasets *without* known classes

- Datasets?
- Competitors?
- Quality measurement?

Real datasets with known classes

That’s not the real situation of clustering (except if the dataset had been clustered before the “true” classes became known).

Real datasets with known classes

That's not the real situation of clustering (except if the dataset had been clustered before the "true" classes became known).

The existence of known "true" classes doesn't preclude the existence of other "true" unknown classes.

Real datasets with known classes

That's not the real situation of clustering
(except if the dataset had been clustered
before the "true" classes became known).

The existence of known "true" classes
doesn't preclude the existence
of other "true" unknown classes.

Unions and parts of known classes may qualify (i.e., there are "communities", "species", "subspecies" in biology)

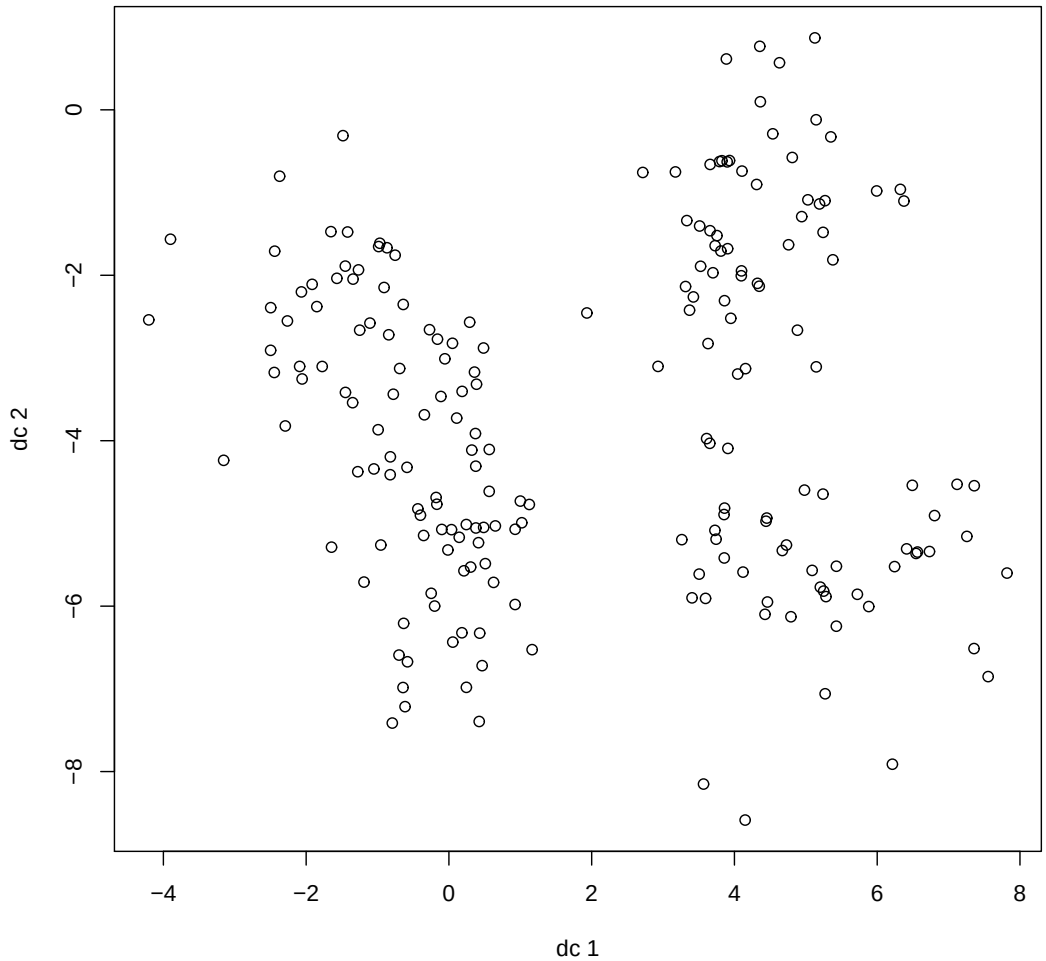
Real datasets with known classes

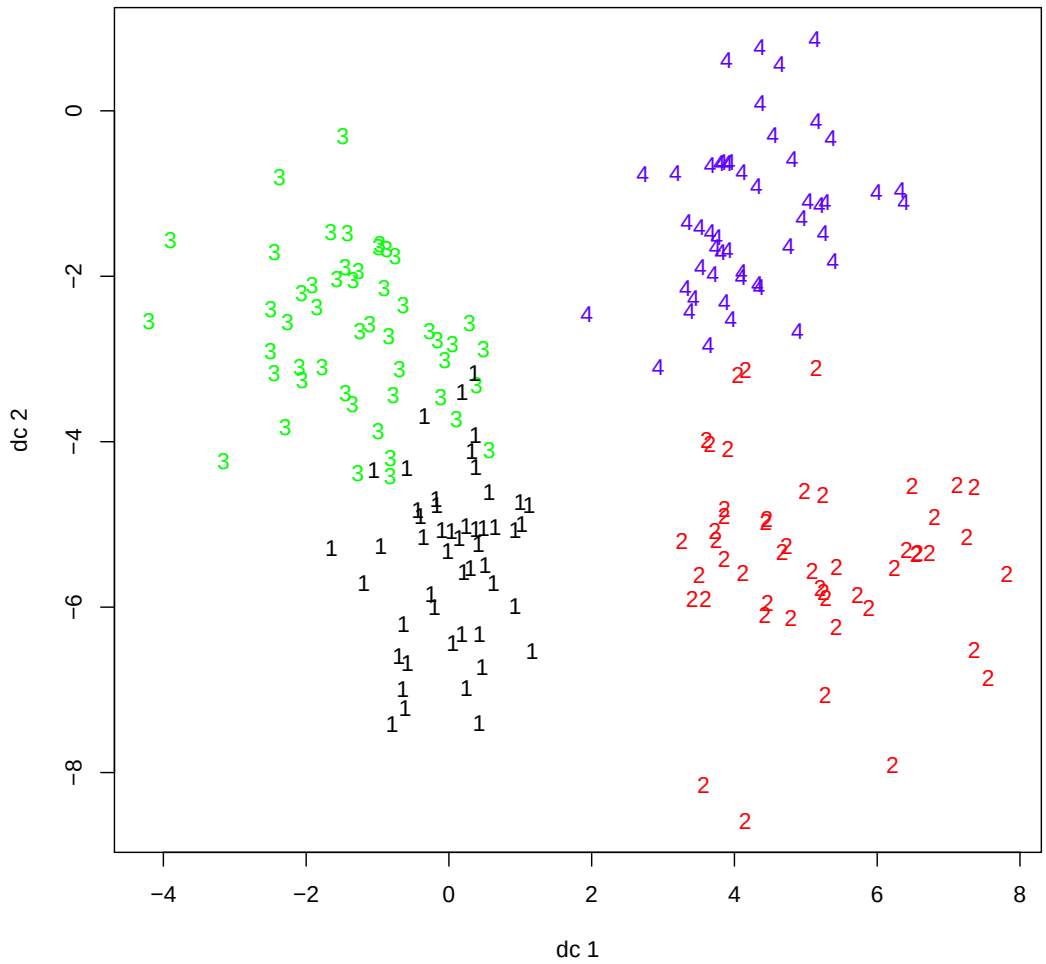
That's not the real situation of clustering
(except if the dataset had been clustered
before the "true" classes became known).

The existence of known "true" classes
doesn't preclude the existence
of other "true" unknown classes.

Unions and parts of known classes may qualify (i.e., there are "communities", "species", "subspecies" in biology)

Known classes may *not* cluster in any data analytic sense.


dc 2
0
−2
−4
−6
−8
−4
−2
0
2
4
6
8
dc 1


dc 2
dc 1
0
−2
−4
−6
−8
−4
−2
0
2
4
6
8

Just checking misclassification rates doesn’t teach us much about the characteristics of the method.

Just checking misclassification rates doesn't teach us much about the characteristics of the method.

Performance in datasets with known classes
still *is* legitimate as a quality indicator,
but it would be helpful to connect success or failure to the characteristics of the data and classes,
so that something can be learned.

Just checking misclassification rates doesn't teach us much about the characteristics of the method.

Performance in datasets with known classes
still *is* legitimate as a quality indicator,
but it would be helpful to connect success or failure to the characteristics of the data and classes,
so that something can be learned.

Therefore, for benchmarking not only datasets,
but also data analytic characteristics of classes should be presented.

Everybody who comes up with a new method
should be able to come up with an example
where one could think that the method could be applied,
and where it fails (with given reasons).

If you cannot do this, you don’t understand your own method.

Simulated datasets (from mixture distributions)

This is less straightforward than most people think.

- It must be “mixtures of cluster-shaped distributions” - in general, mixture/non-mixture is not a proper classification of distributions.

Simulated datasets (from mixture distributions)

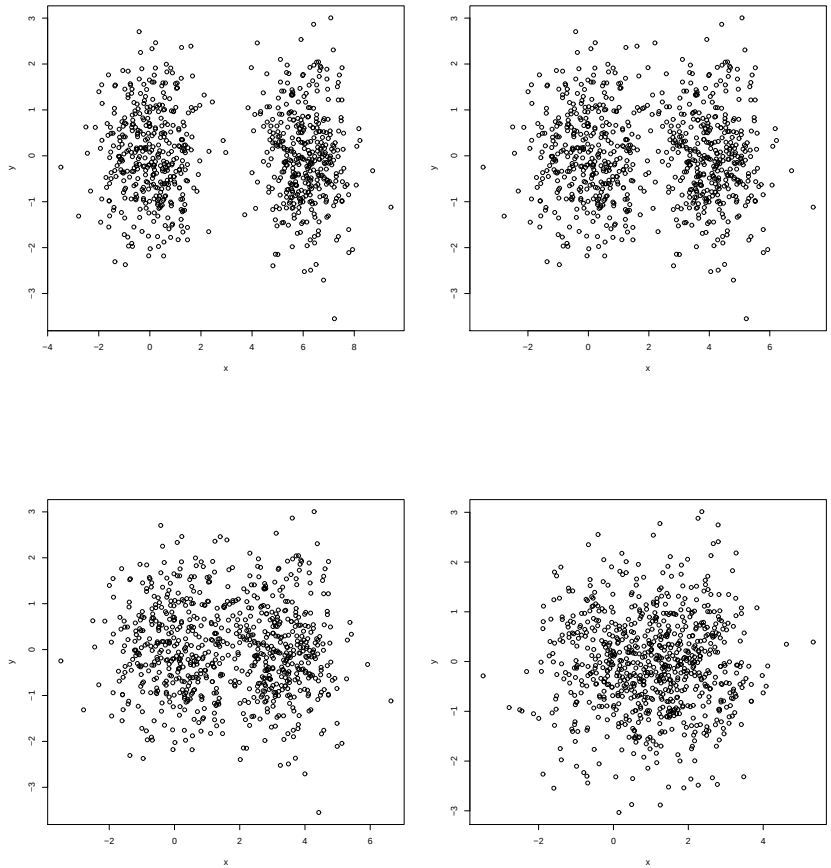
This is less straightforward than most people think.

- It must be “mixtures of cluster-shaped distributions” - in general, mixture/non-mixture is not a proper classification of distributions.
- A mixture of several Gaussians can be a single cluster (depending on what is meant by “a cluster”).

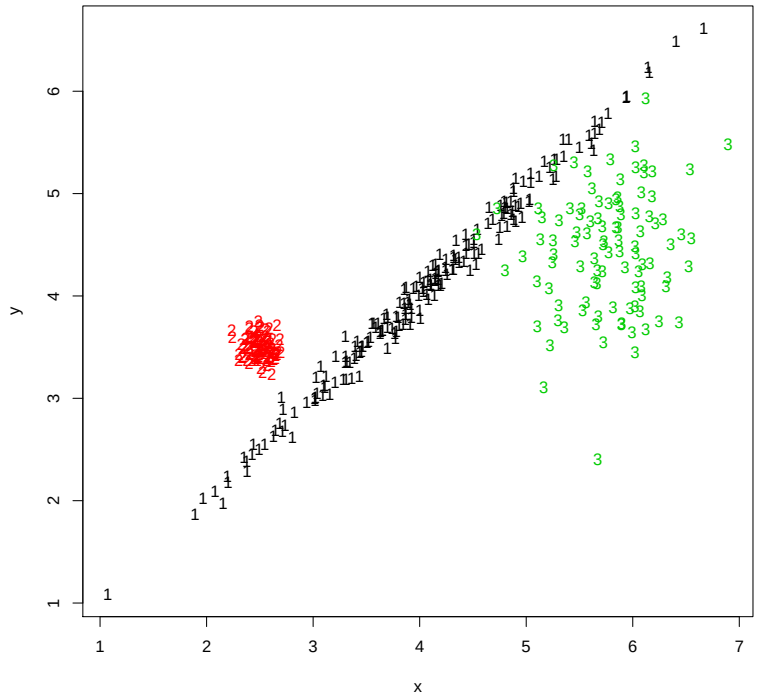
Simulated datasets (from mixture distributions)

This is less straightforward than most people think.

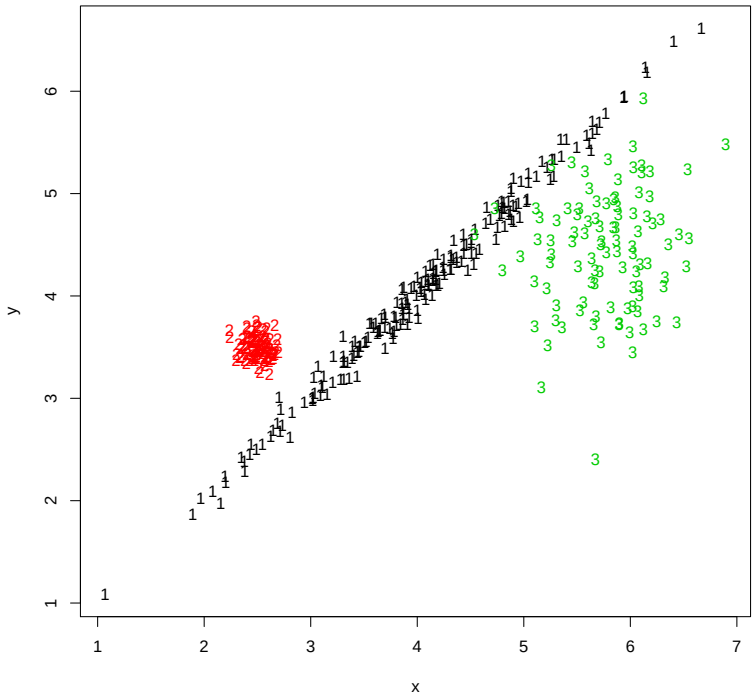
- It must be “mixtures of cluster-shaped distributions” - in general, mixture/non-mixture is not a proper classification of distributions.
- A mixture of several Gaussians can be a single cluster (depending on what is meant by “a cluster”).
- Does a t-distribution qualify as “cluster shaped”? An exponential?

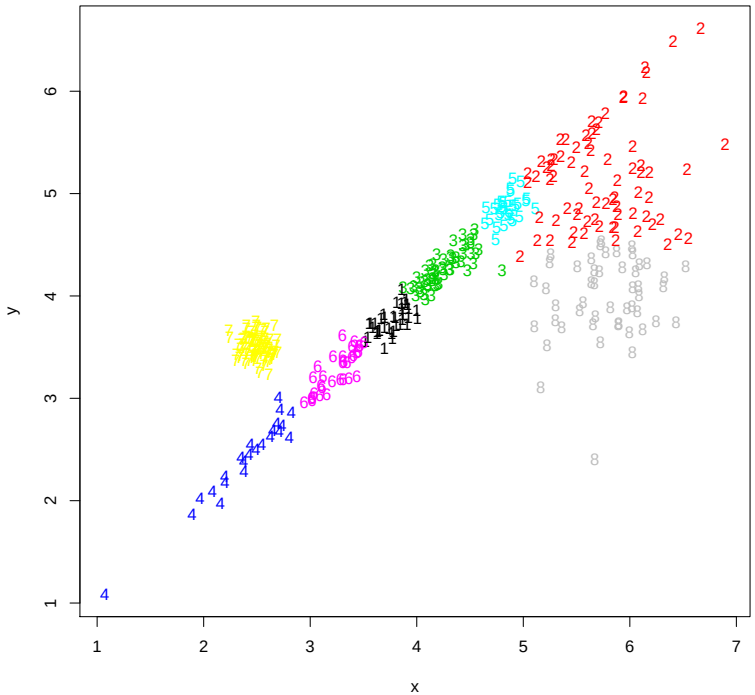

x
y

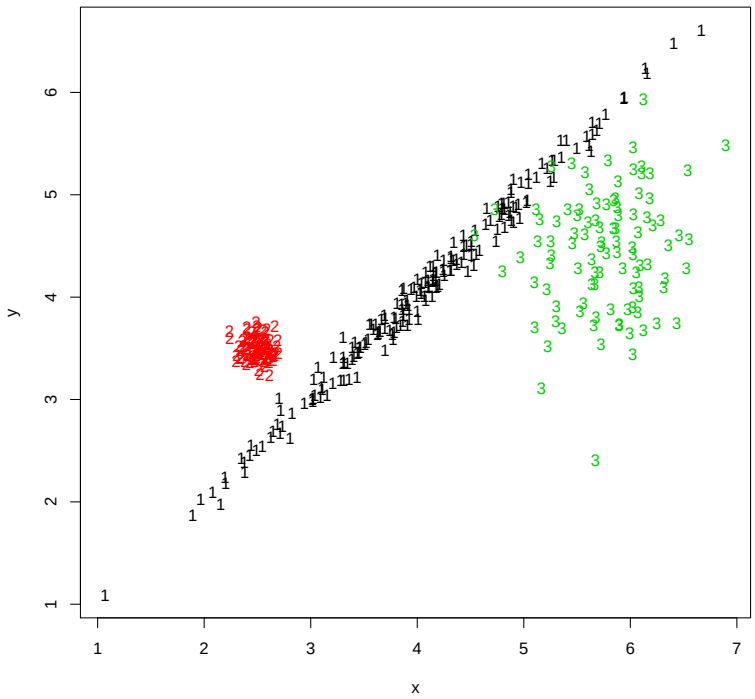
$$f(\mathbf{x}) = \sum_{i=1}^{k} \pi_i \varphi_{\mathbf{a}_i, \Sigma_i}(\mathbf{x})$$

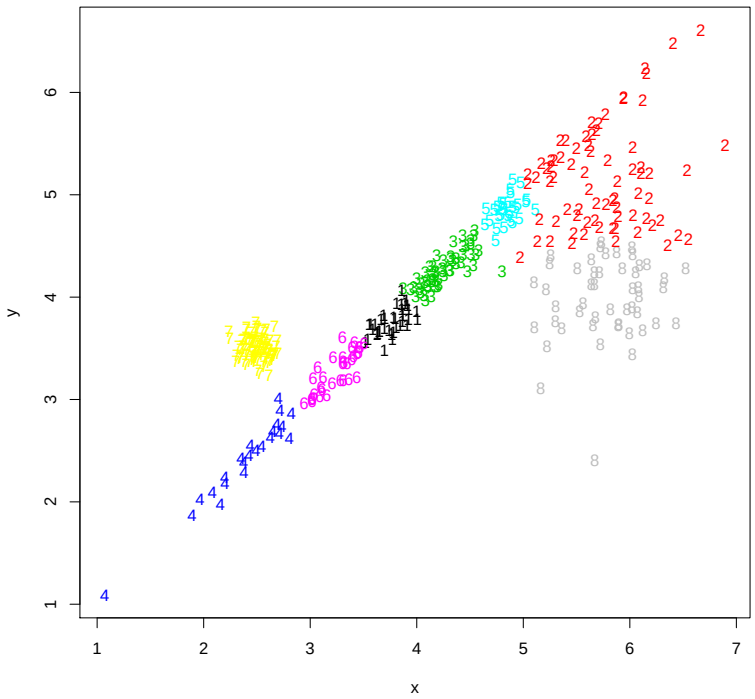


$f(\mathbf{x}) = \sum_{i=1}^{k} \pi_i \varphi_{\mathbf{a}_i, \Sigma_i}(\mathbf{x})$ - Σ_i diagonal?









Even if the true model yields visible clusters,
these are not necessarily the clusters of interest.

However, to know the performance of methods on simulated data is certainly useful for their characterisation and understanding.

Real datasets without known classes

Most clearly connected to real clustering tasks.

Real datasets without known classes

Most clearly connected to real clustering tasks.

Major issue: measurement of quality.
Misclassification rate won't work.

Information about clustering aim would be good,
but this is often difficult to connect to quality measurement.

Compute validation indexes?

But if these were important, why not optimise them directly; why bother to use other clustering methods?

Compute validation indexes?

But if these were important, why not optimise them directly; why bother to use other clustering methods?

Validation indexes measure certain aspects of clustering, and this is relevant if somebody claims that her new method fulfills this aspect.

Compute validation indexes?

But if these were important, why not optimise them directly; why bother to use other clustering methods?

Validation indexes measure certain aspects of clustering, and this is relevant if somebody claims that her new method fulfills this aspect.

Need to be connected to clustering aims.

Benchmarking should not be ranking!

Methods need to be characterised in multidimensional ways:

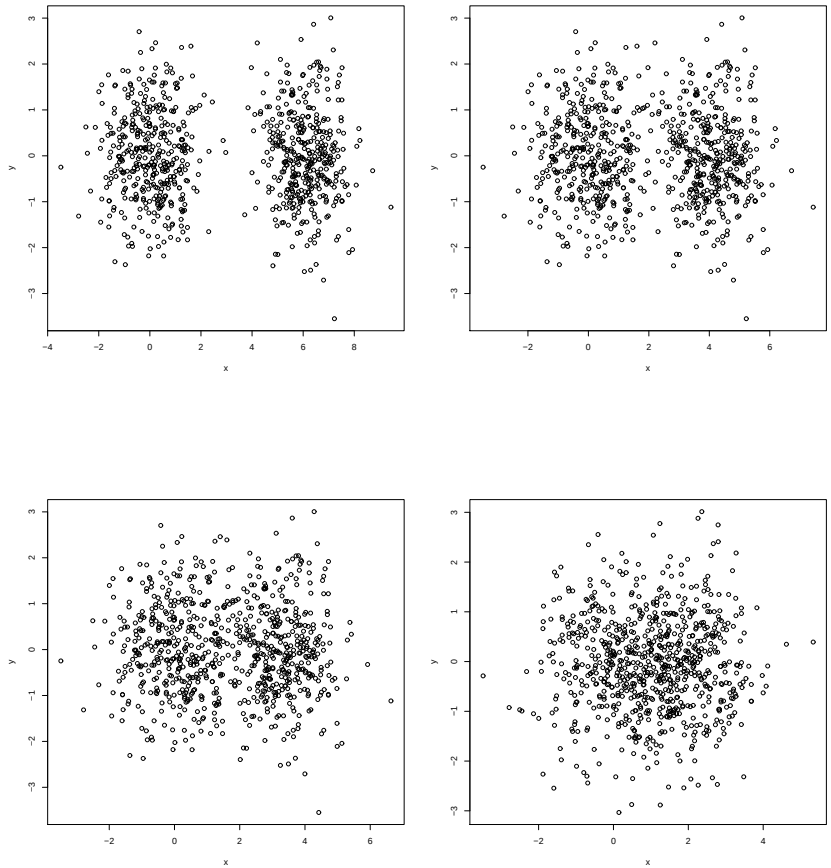
- Separation vs. homogeneity
- Cluster shapes (spherical/elliptical/skew/nonlinear)
- Small subgroups/outliers
- Stability/consistency
- Implicit similarity concept

New methods should come with specific (nontrivial) claims, and should be tested against these.

Competitors should be chosen suitable for the same aims.

Benchmarking datasets need to be documented so that it is clear for what specific claims they are relevant.

Results may not be interpreted on 1-d quality scale.



Find three clusters!