Decisions that are needed when using cluster analysis, and research that helps with making them

Christian Hennig

Supported by EPSRC Grant EP/K033972/1

1. Introduction

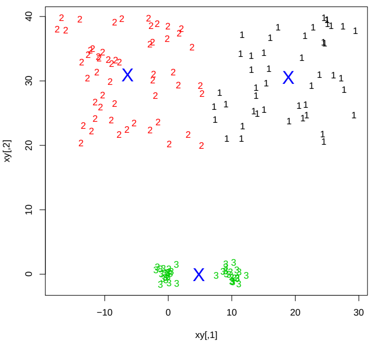
How to do clustering?
How to represent information?
What's the best method?

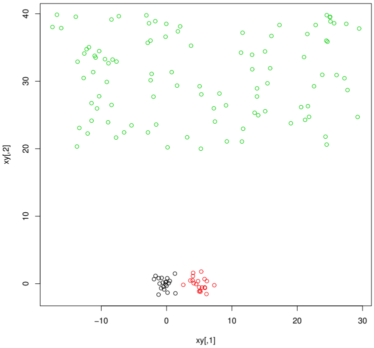
1. Introduction

How to do clustering?
How to represent information?
What's the best method?

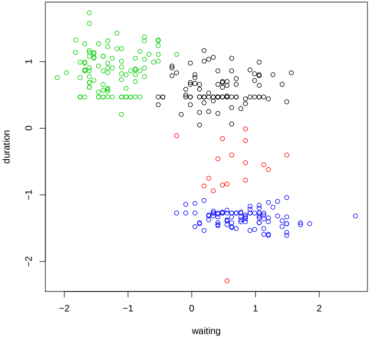
What do we want from clustering?
Various things…
that aren't necessarily served by the same clustering.

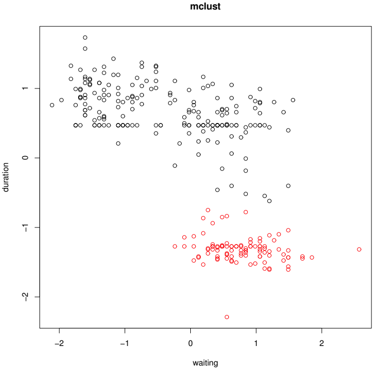
Optimal representation vs. pattern recognition





Granularity?





Clustering is applied in various fields
with various aims that have different requirements

E.g., object recognition in images
requires separation on suitable features,

E.g., object recognition in images
requires separation on suitable features,

clustering for information reduction requires
good representation by centroids,

E.g., object recognition in images
requires separation on suitable features,

clustering for information reduction requires
good representation by centroids,

clustering regions for mapping requires
similar (and close?) regions within clusters,

E.g., object recognition in images
requires separation on suitable features,

clustering for information reduction requires
good representation by centroids,

clustering regions for mapping requires
similar (and close?) regions within clusters,

noisy observations from heterogeneous sources
require telling apart homogeneous probability models.

How to do clustering
depends on decisions that we have to make.

How to do clustering
depends on decisions that we have to make.

Speaking of “the natural clusters” is misleading;
it suggests that the data on their own
can know the best clustering,
and decisions can/should be made automatically.

How to do clustering
depends on decisions that we have to make.

Speaking of “the natural clusters” is misleading;
it suggests that the data on their own
can know the best clustering,
and decisions can/should be made automatically.

Not so. . .
but how can cluster analysis research help deciding?

2. The data used for clustering

Decision how to represent information in data

- Definition, choice and selection of variables
- Dissimilarity definition
- Transformation and standardisation
- Dealing with missing information etc.

These often make a huge difference.

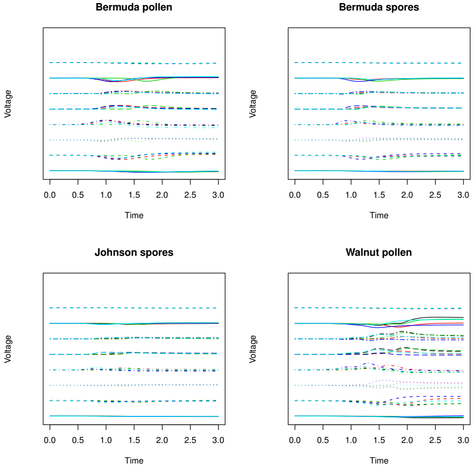
2.1 Standardisation

. . . is reweighting of variables
in order to give all variables same influence
(depending on method).

2.1 Standardisation

. . . is reweighting of variables
in order to give all variables same influence
(depending on method).

If relevant information content
is proportional to variation
it's a reason *not* to standardise
(and not to use anything scale invariant).


Bermuda pollen
Bermuda spores
Johnson spores
Walnut pollen
Voltage
Time
0.0
0.5
1.0
1.5
2.0
2.5
3.0

2.2 Variable selection (and dimension reduction)

There's much literature
on automatic variable selection for clustering.

Principles:

- Models assuming some variables "noise"
(eg, Law, Figueiredo & Jain 2004)

2.2 Variable selection (and dimension reduction)

There's much literature
on automatic variable selection for clustering.

Principles:

- Models assuming some variables "noise" (eg, Law, Figueiredo & Jain 2004)
- Select variables to find "most clustered clustering" (eg, Montanari & Lizzani 2001)

2.2 Variable selection (and dimension reduction)

There's much literature
on automatic variable selection for clustering.

Principles:

- Models assuming some variables "noise"
  (eg, Law, Figueiredo & Jain 2004)
- Select variables to find "most clustered clustering"
  (eg, Montanari & Lizzani 2001)
- Eliminate redundant (dependent) variables
  (eg, Fraiman, Justel & Svarc 2008)

Obviously this is helpful for exploratory DA. . .

Obviously this is helpful for exploratory DA. . .

. . . but the selected variables define the meaning of the clustering!

Is this for the data to decide?

Reasons against automatic variable selection

- Don't eliminate variables essential for research aim (prescribed cluster meaning; or use for external objective)

Reasons against automatic variable selection

- Don't eliminate variables essential for research aim (prescribed cluster meaning; or use for external objective)
- Statistically dependent variables
may be different in meaning;
dependence information is not redundant

Reasons against automatic variable selection

- Don’t eliminate variables essential for research aim (prescribed cluster meaning; or use for external objective)
- Statistically dependent variables may be different in meaning; dependence information is not redundant
- “Most clustered clustering” may mean very little due to degrees of freedom in finding it.

Reasons against automatic variable selection

- Don't eliminate variables essential for research aim (prescribed cluster meaning; or use for external objective)
- Statistically dependent variables may be different in meaning; dependence information is not redundant
- "Most clustered clustering" may mean very little due to degrees of freedom in finding it.

Could create meaningful summary variables rather than reducing dimension automatically.

3. Characterising clustering methods

For choosing a suitable clustering method, need understanding how what they do relates to clustering aim.

Every clustering method comes with an implicit "cluster concept" with certain characteristics.

Understanding by "direct interpretation" and theoretical investigation of characteristics.

E.g., k-means:

$$\sum_{i=1}^{n} \|\mathbf{x}_i - \mathbf{m}_{c(i)}\|^2 = \min!$$

- Formalises representation of objects by centroids

E.g., k-means:

$$\sum_{i=1}^{n} \|\mathbf{x}_i - \mathbf{m}_{c(i)}\|^2 = \min!$$

- Formalises representation of objects by centroids
- Distances in all directions from centroid count the same (standardisation will have impact)

E.g., *k*-means:

$$\sum_{i=1}^{n} \|\mathbf{x}_i - \mathbf{m}_{c(i)}\|^2 = \min!$$

- Formalises representation of objects by centroids
- Distances in all directions from centroid count the same (standardisation will have impact)
- No enforcement of between-cluster separation

E.g., k-means:

$$\sum_{i=1}^{n} \|\mathbf{x}_i - \mathbf{m}_{c(i)}\|^2 = \min!$$

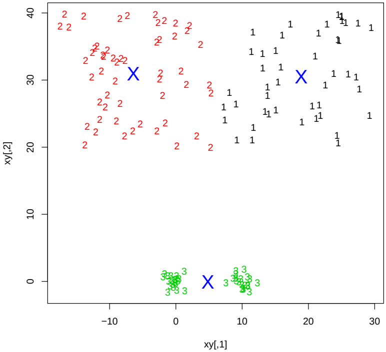
- Formalises representation of objects by centroids
- Distances in all directions from centroid count the same (standardisation will have impact)
- No enforcement of between-cluster separation
- Squared distances $\Rightarrow$
  unforgiving against large within-cluster distances

E.g., k-means:

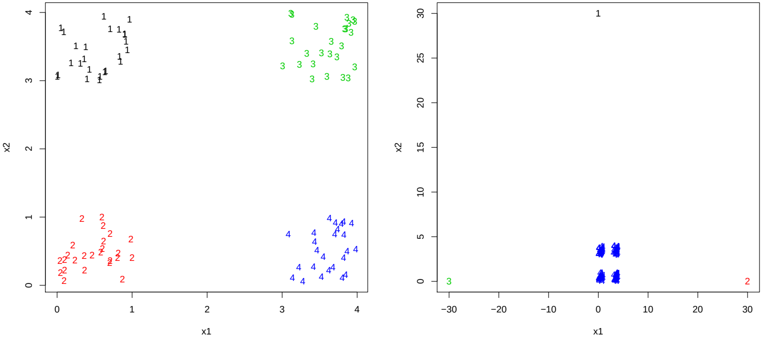
$$\sum_{i=1}^{n} \|\mathbf{x}_i - \mathbf{m}_{c(i)}\|^2 = \min!$$

- Formalises representation of objects by centroids
- Distances in all directions from centroid count the same (standardisation will have impact)
- No enforcement of between-cluster separation
- Squared distances $\Rightarrow$ unforgiving against large within-cluster distances
- Maximum likelihood for partition model of spherical Gaussians *(characteristic, not requirement!)*

Representation not separation



Unforgiving against large within-cluster distances



Theory:
Axiomatic characterisation of clustering methods
Jardine and Sibson (1971), Fisher and van Ness (1971)

Theory:
Axiomatic characterisation of clustering methods
Jardine and Sibson (1971), Fisher and van Ness (1971)

Impossibility Theorem (Kleinberg, 2002)

There's no clustering method that fulfills
scale invariance, richness, "consistency"

Consistency: if all within-cluster distances
are made smaller and all between-cluster distances
made larger, we get the same clustering.

Theory:
Axiomatic characterisation of clustering methods
Jardine and Sibson (1971), Fisher and van Ness (1971)

Impossibility Theorem (Kleinberg, 2002)

There's no clustering method that fulfills
scale invariance, richness, "consistency"

Consistency: if all within-cluster distances
are made smaller and all between-cluster distances
made larger, we get the same clustering.

This is not usually desirable!

Until that point: focus on supposedly *generally desirable* characteristics, “admissibility”.

Until that point: focus on supposedly
generally desirable characteristics, “admissibility”.

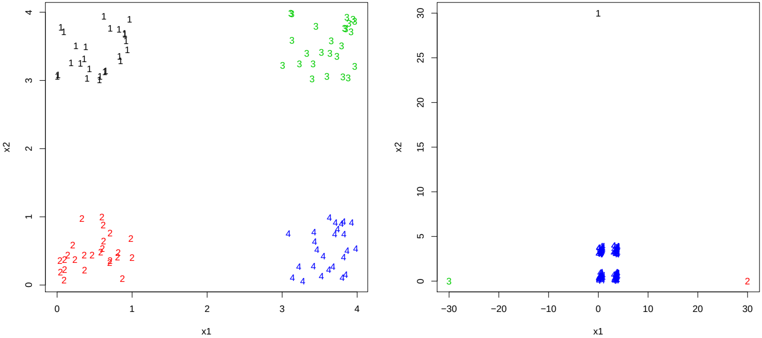
More useful for helping with decisions:
Characteristics that *distinguish*
different kinds of clustering methods, e.g.,

order invariance: order-preserving transformation
of distances doesn’t change clustering
(Ackerman et al. 2010; Fisher & van Ness 1971)

Attempts to formalise robustness
against adding observations:
Hennig (2008) cluster-wise "dissolution point"
Ackerman et al. (2012) "δ-robustness"

How dissimilar can a clustering become
when adding g points?

k-means' fixed *k* dissolution point is minimum.



Theoretical characterisation of clustering methods
relevant for practice by
connecting characteristics to clustering aims
is an underdeveloped,
important field for further research!

4. **Cluster Benchmarking** - for help with choosing a method

ifcs

CLUSTER BENCHMARK
DATA REPOSITORY

HTTP://IFCS.BOKU.AC.AT/REPOSITORY

(With Iven van Mechelen, Nema Dean, Fritz Leisch, Rainer Dangl, Anne-Laure Boulesteix, Isabelle Guyon, Doug Steinley)

4.1 Approaches to cluster benchmarking

Comparison of clustering methods on

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

4.1 Approaches to cluster benchmarking

Comparison of clustering methods on

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

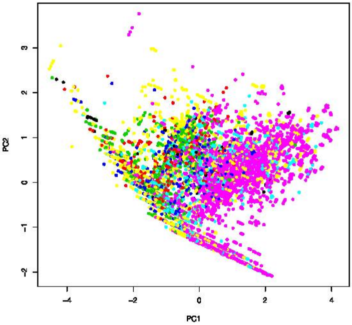
- Datasets?
- Competitors?
- Quality measurement?

Why not just use data with known true classes?

Why not just use data with known true classes?

- There may be more than one legitimate clustering in a dataset.
- Real datasets with known classes will tell us nothing about generalisability.
- “True classes” may not be data analytic clusters.

7 standard occupation classes such as “manual workers”, “managerials and professionals”, “not working”



4.2 Quality measurement with unknown truth

(current research, Hennig 2017, arxiv)

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
- Good representation of data by centroids
- Good representation of dissimilarity by clustering
- Clusters are regions of high density without within-cluster gaps
- Uniform cluster sizes
- Clusters easily characterisable by few variables
- Clusters well related to external information
- Stability

Measuring within-cluster homogeneity
by average within-cluster dissimilarity.

Measuring within-cluster homogeneity
by average within-cluster dissimilarity.

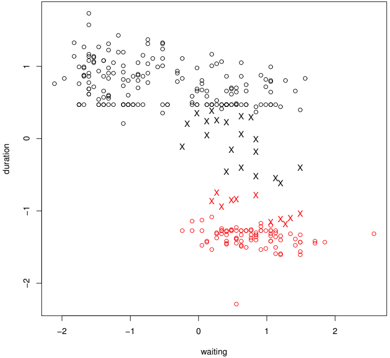
Measuring representation of dissimilarities
by "Pearson-Γ" (Hubert and Schultz 1976, Halkidi et al. 2016):
Correlation between vector of dissimilarities
and 0-1 vector for "in same/different cluster".

Measuring within-cluster homogeneity
by average within-cluster dissimilarity.

Measuring representation of dissimilarities
by "Pearson-Γ" (Hubert and Schultz 1976, Halkidi et al. 2016):
Correlation between vector of dissimilarities
and 0-1 vector for "in same/different cluster".

Representation of objects by centroids
- eg, *k*-means criterion.

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



Clusters corresponding to high density regions
- tough to measure (work in progress);

Many open problems:
theoretical characterisation of indexes,
relations between indexes,
non-distance based criteria,
big data computation/approximation.

4.3 Data with known clusters used in different ways

Benchmarking with data with known clusters
is surely relevant;
discovering hidden structure
is a major aim of clustering,
and can only be tested this way.

4.3 Data with known clusters used in different ways

Benchmarking with data with known clusters
is surely relevant;
discovering hidden structure
is a major aim of clustering,
and can only be tested this way.

Unfortunately, benchmarking with known clusters
is often used for 1-d quality ranking
with little attention to generalisation and
what can be learnt for clustering in practice.

Data with known clusters used in different ways:

- On what characteristics of real clustering problem does a method's performance depend?

Data with known clusters used in different ways:

- On what characteristics of real clustering problem does a method's performance depend?
- On what features/characteristics of the data does a method's performance depend?

Data with known clusters used in different ways:

- On what characteristics of real clustering problem does a method's performance depend?
- On what features/characteristics of the data does a method's performance depend?
- How are the different characteristics/criteria and methods related on real data?

5. Cluster validation

Assessing the quality of a clustering on a dataset,
incl. comparing clusterings, parameters
(such as number of clusters)

Elements of cluster validation

- Internal validation indexes
- Stability assessment
- Use of external information
- Testing for clustering structure
- Visual exploration
- Comparison of different clusterings on same data/ sensitivity analysis

5.1 Visual exploration

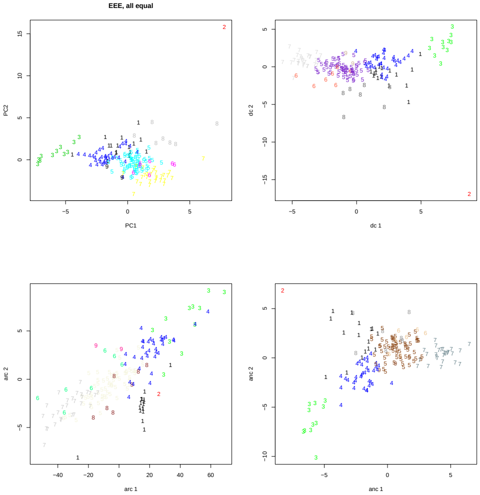
Visualisation is *central* for cluster validation.
I'd never trust and interpret a clustering
without visual evaluation.

5.1 Visual exploration

Visualisation is *central* for cluster validation.
I'd never trust and interpret a clustering
without visual evaluation.

(If at all possible, I wouldn't leave number of clusters
to automatic criterion, rather decide from graphs
and background information.)

Visualisation for cluster validation (Rao 1952, Hennig 2004)

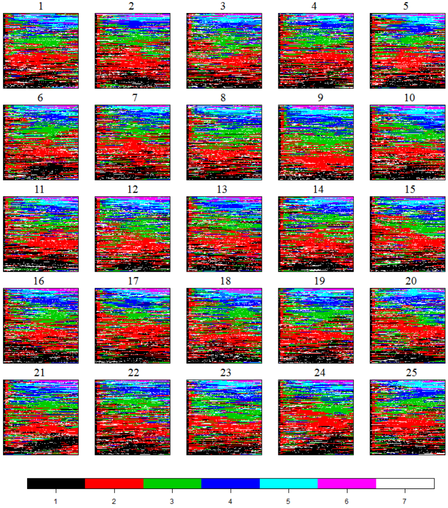


5.2 Parametric bootstrap

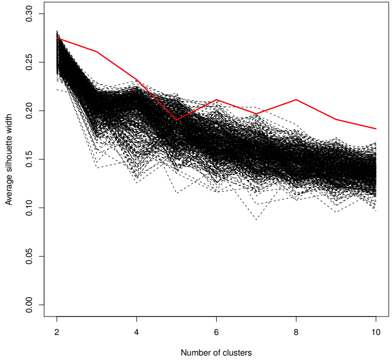
Parametric bootstrap (Efron 1979):
fitting a model to data, sampling from fitted model.

Uses in cluster validation:
Compare real data
with homogeneity model for "no clustering" (Hennig & Lin 2015)
capturing all non-clustering structure.

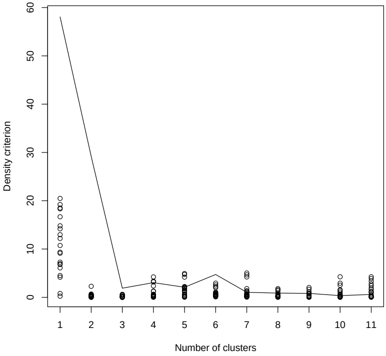
Buja et al. (2009) - visual testing:



Comparing ASW with values under null model



Comparing quality index with fitted models for all n.o.c. (work in progress)



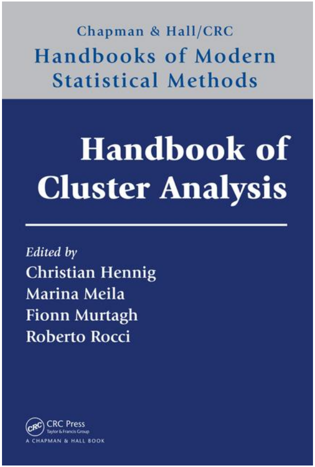
Conclusion

Too much cluster analysis research
is about automatising decisions.

Too little cluster analysis research
acknowledges what researchers should decide,
and tries to help them making the decisions.

There are endless opportunities for
this kind of research.

Some marketing:



. . . and some more:

- C. Hennig (2004) Asymmetric linear dimension reduction for classification. Journal of Computational and Graphical Statistics 13, 930-945
- C. Hennig (2008) Dissolution point and isolation robustness: robustness criteria for general cluster analysis methods. Journal of Multivariate Analysis 99, 1154-1176.
- C. Hennig and C.-J. Lin (2015) Flexible parametric bootstrap for testing homogeneity against clustering and assessing the number of clusters. Statistics and Computing 25, 821-833
  Open access
- C. Hennig (2015) Clustering strategy and method selection. In Hennig, C., M. Meila, F. Murtagh, and R. Rocci (Eds.). Handbook of Cluster Analysis. Chapman and Hall/CRC
  Free version on arxiv
- C. Hennig (2017) Cluster validation by measurement of clustering characteristics relevant to the user. Submitted.
  Free version on arxiv

Other authors (selection):

- M. Ackerman, S. Ben-David, and D. Loker (2010) Towards Property-Based Classification of Clustering Paradigms. NIPS 2010.
- A. Buja, D. Cook, H. Hofmann, M. Lawrence, E.-K. Lee, D. Swayne, H. Wickham (2009) Statistical inference for exploratory data analysis and model diagnostics. Philosophical Transactions of The Royal Society, A.
- L. Fisher and J. W. Van Ness (1971) Admissible Clustering Procedures. Biometrika 58, 91-104.
- J. Kleinberg (2002) An Impossibility Theorem for Clustering. NIPS 2002.
- C. R. Rao (1952) Advanced Statistical Methods in Biometric Research. Wiley, New York.