Determining the number of clusters - discussion

Christian Hennig

July 17, 2013

1. Introduction

Mirkin's useful overview:

- pre processing
- at processing
- post processing

Density level sets: at processing
SICL: post processing.
Pre processing hardly exists; iK-means fascinating.
Can apply to other clustering methods?

Level sets and mixture/ICL not covered by Mirkin;
k-means type methods,
implicitly for spherical clusters (mostly).

Different cluster concepts:

- dense but arbitrarily shaped,
- elliptical but with potentially different cov-matrices,
- there are many more.

Level sets and mixture/ICL not covered by Mirkin;
k-means type methods,
implicitly for spherical clusters (mostly).

Different cluster concepts:

- dense but arbitrarily shaped,
- elliptical but with potentially different cov-matrices,
- there are many more.

What are “natural” clusters?

Is the term “natural” helpful in statistical arguments?

It refers to something that we cannot (yet) make precise,
it refers to human intuition about *abstract* structures, which is a result of social processes.

That not what people think of when using the term “natural”.

2. Comparing methods

See benchmarking session yesterday.

- What data?
- Methods to compare
- Quality measurement (k/ARI)

There is no single “best” method.
“True clusters” have different characteristics, methods have different implicit concepts.

Mirkin’s results need qualification.

We better find out which method is best
for which kinds of clusters (separation, shape etc.).

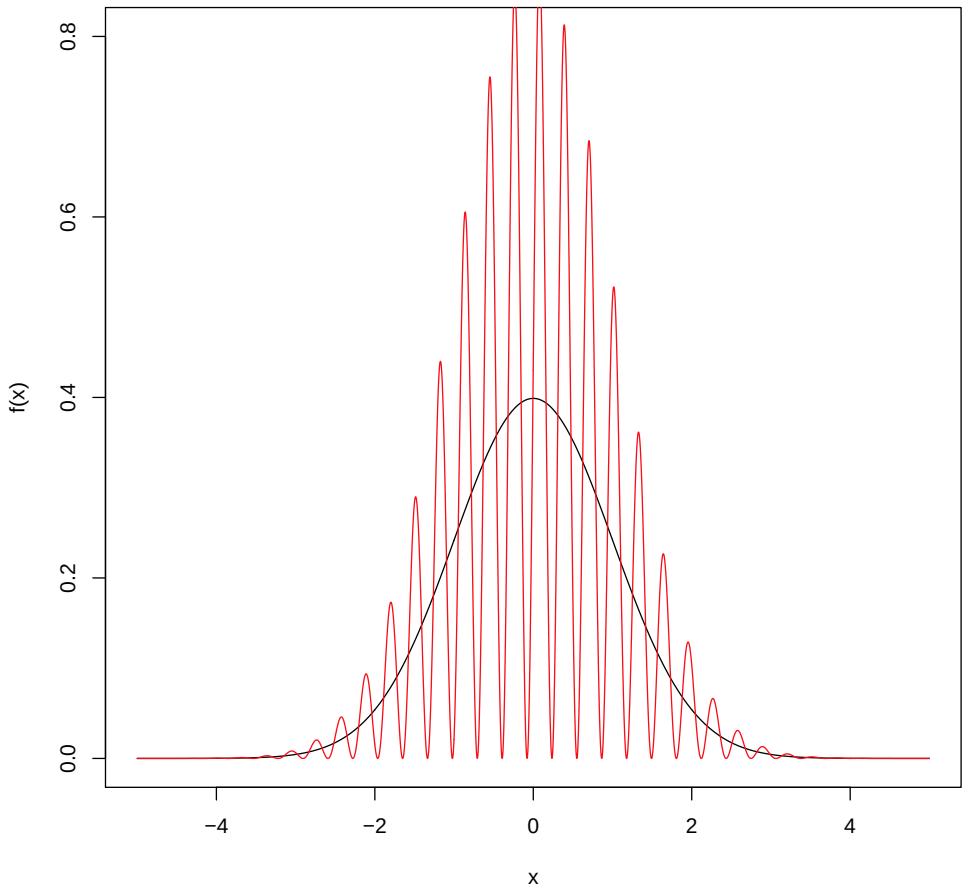
For example, *average silhouette width*
has won another comparative study (Arbelaitz et al. 2012),
will emphasize gaps more,
and will tolerate more flexible cluster shapes
than criteria based on k-means objective
(H-index, Calinski & Harabasz etc.);
latter are better for spherical normal clusters.

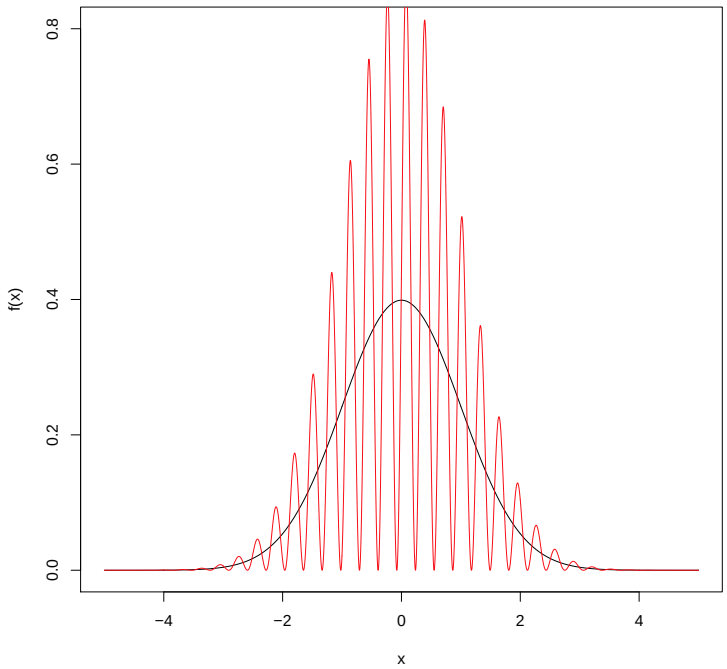
3. Level sets

Certainly intuitive, flexible and well worked out.

Conceptual issues:
(a) Choice of bandwidth and level vs. choice of k

(b) The pathological discontinuity of densities as functionals of distributions (Davies 1995, 2008).





Arbitrarily close to distribution with density
are distributions without density
or with infinitely many density modes,
. . . and this happens for real data.

It may be better to interpret this as estimating *smoothed density*,

$$f_h^*(x) = \int \frac{1}{h} K\left(\frac{x-y}{h}\right) dF(y),$$

i.e. bandwidth is part of *problem definition*.

4. Supervised ICL

Use of external information: properly treated here
and rarely elsewhere.

Role of “illustrative variables”?
To what extent should they *influence* the clustering?
If used for number of clusters,
they do that to some extent (may be intended),
but not as strong as if used as clustering variable.

How to check whether it works well?

Choose true clustering as u: too nice.
Choose u unrelated to true clustering: too nasty".

More interesting: u with some but weak connection to truth, could explore this by varying strength of connection.

How to check whether it works well?

Choose true clustering as u: too nice.
Choose u unrelated to true clustering: too nasty".

More interesting: u with some but weak connection to truth, could explore this by varying strength of connection.

In real example: illustrative variables shouldn't be *only* quality assessment criterion
(then they should help clustering).
Unrelated solution perhaps not so "useless".

Assumption: y and u independent given z:

Good for theory but probably unrealistic.
Would expect u to be somehow connected to y, not totally determined by z.
(E.g. self-assessment ecological awareness vs. energy consumption).

Assumption: y and u independent given z:

Good for theory but probably unrealistic.
Would expect u to be somehow connected to y,
not totally determined by z.
(E.g. self-assessment ecological awareness vs.
energy consumption).

May still be useful in practice though.
(Could check by simulations
where assumption is slightly violated.)

Assumption: y and u independent given z:

What does it *do*?
Tries to enforce to find a z that explains
all connection between y and u.

That's pretty strong.

Assumption: y and u independent given z:

What does it *do*?
Tries to enforce to find a z that explains
all connection between y and u.

That's pretty strong.

But then it *only* can act
at level of comparing different number of clusters,
so impact is limited.
(I think it's only good
because the impact is limited...)