UCL

Discussion - Session on Visualisation

Christian Hennig

September 8, 2010

Overview

1. Philosophy of visualisation
2. Dimension reduction as visualisation tool
3. Large p/small n situations
4. Heatmaps
5. Some remarks on cluster validation

1. Philosophy of visualisation

Visualisation is about showing what is “interesting” about the data in a way that the viewer can make sense of it.

Many methods rely on model assumptions; people tend to ask: “Are the assumptions fulfilled?”

However, more relevant questions are:
“How does the method define *relevant, interesting information*?
How does it make sure that this is made visible?
What is not showed, and under which circumstances may this be relevant as well?”

Model assumptions may help, but don’t address it directly.

2. Dimension reduction as visualisation tool

Obviously dimension reduction means loss of information.

Do we lose relevant information?

Principal components show maximum variance directions,
i.e., relevant information is lost
if it is not visible along maximum variance directions.

Are PCs a good projection to visualise a clustering on?

Discriminant coordinates optimise
variation between group means.
Show “interesting” information in the sense of group separation.
Good if separation between groups
is same as separation between group means.

Overestimate discriminative power,
but that’s not really a problem for visualisation.

Main idea in Hennig (2005) is to use a 2-d plot for every single cluster, not to try to show all separation in 2-d.

Are PPFC/PFC/PIRE intended to be visualisation tools?

Are PPFC/PFC/PIRE intended to be visualisation tools?

How can it be described what they take as “interesting relevant” information?
How can it be described what they ignore, and under which circumstances this may be relevant?
Are there essential differences in this respect between these methods?

3. Large *p*/small *n* situations

It seems to me that in such situations we tend to make model assumptions that essentially cannot be checked.
"Large *p*" means complex assumptions, e.g., $\epsilon \sim \mathcal{N}(0, \Delta)$.
"Small *n*" means "not much information to check them".

3. Large *p*/small *n* situations

It seems to me that in such situations we tend to make model assumptions that essentially cannot be checked.
"Large *p*" means complex assumptions, e.g., $\epsilon \sim \mathcal{N}(0, \Delta)$.
"Small *n*" means "not much information to check them".

What is the role of these assumptions?
We cannot be expected to believe them.
They are "convenience assumptions" in some sense.
But should we bother about their appropriateness? How?

4. Heatmaps

Decisions to make for a heatmap:

- Order/clustering method for rows and columns
- Standardisation row/column-wise, and how?
- Discretisation, choice of colours
  (why is red/green the standard?)

Any comments how, why, what impact?

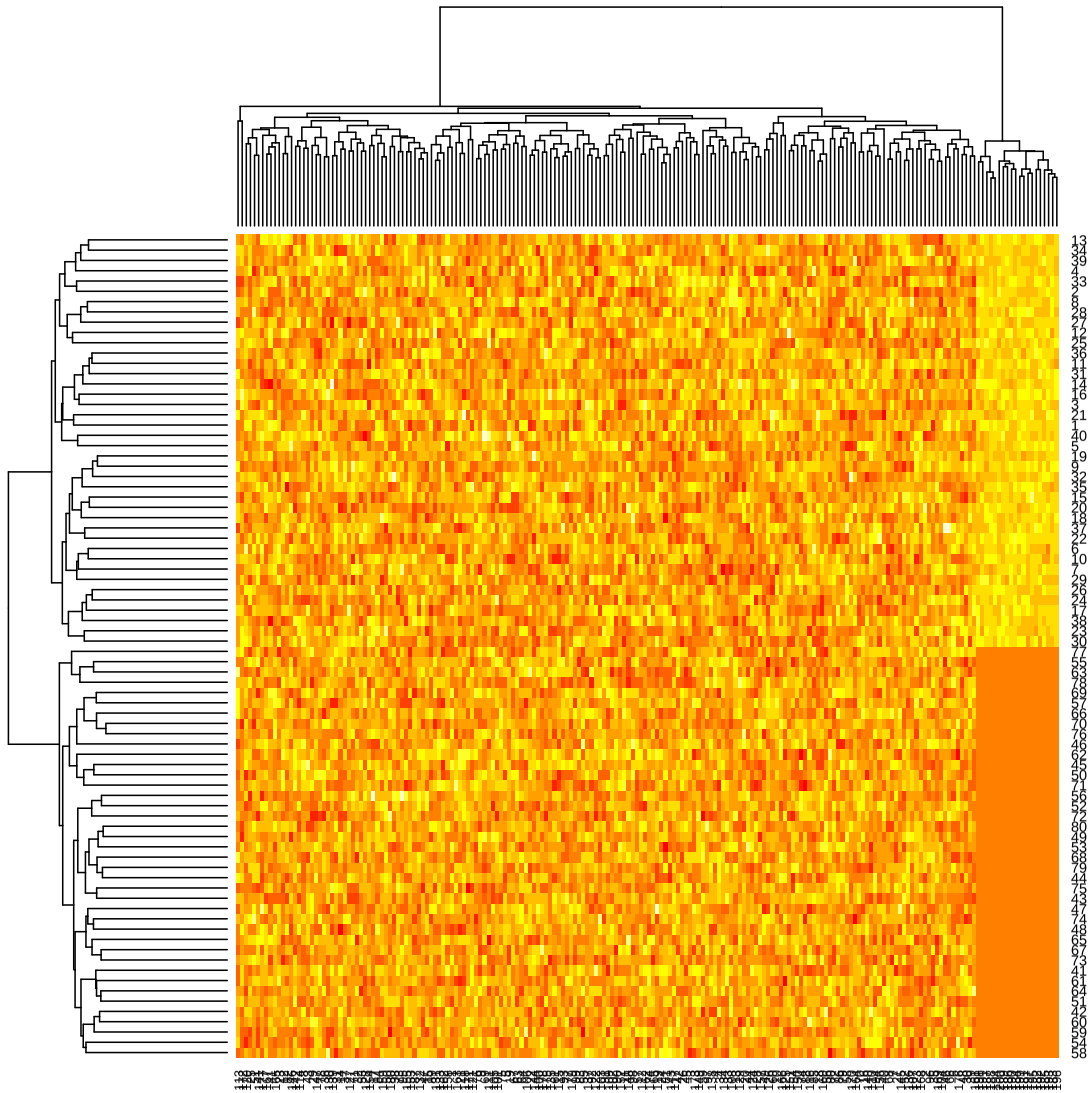
4. Heatmaps

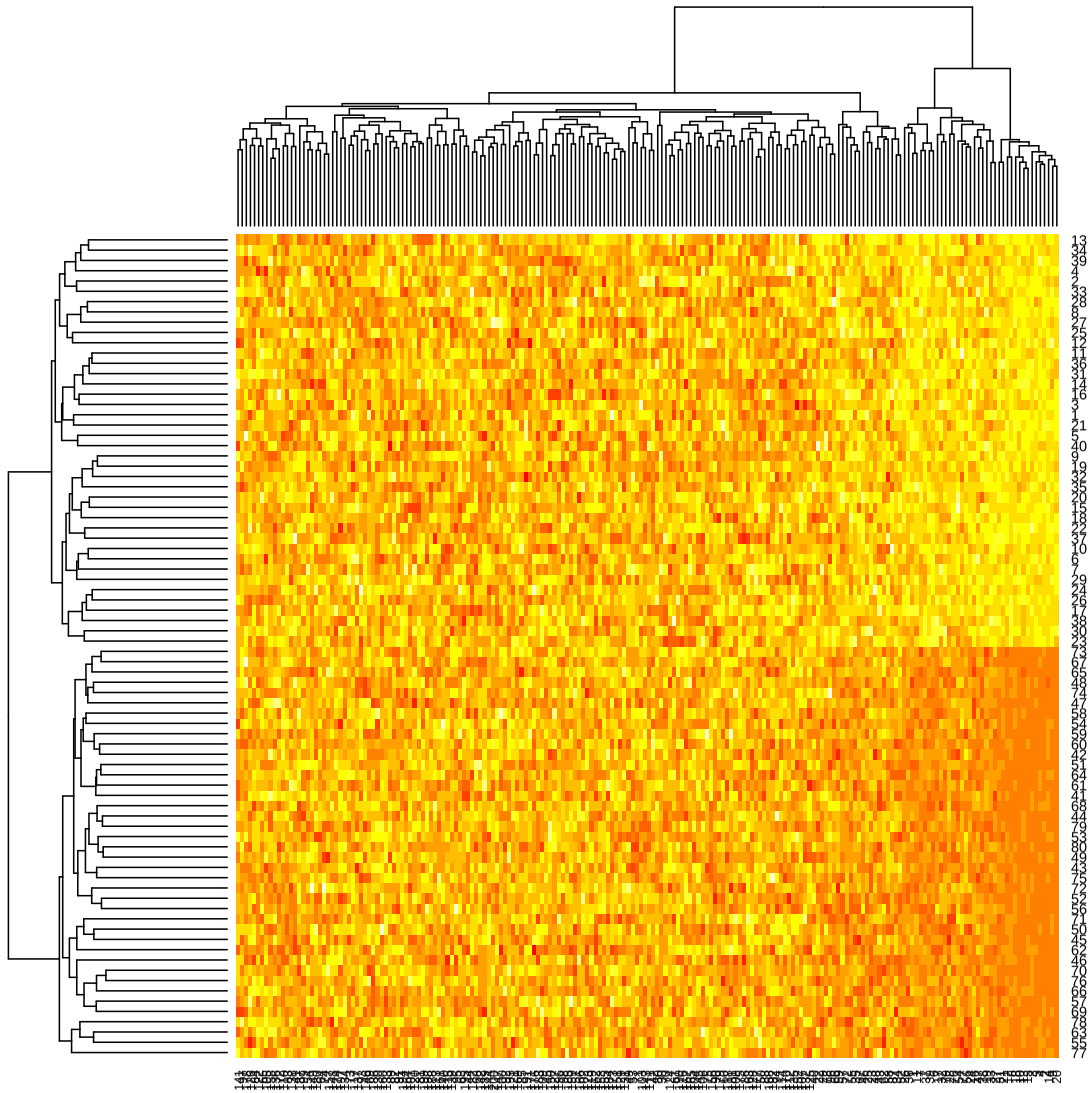
Decisions to make for a heatmap:

- Order/clustering method for rows and columns
- Standardisation row/column-wise, and how?
- Discretisation, choice of colours
  (why is red/green the standard?)

Any comments how, why, what impact?

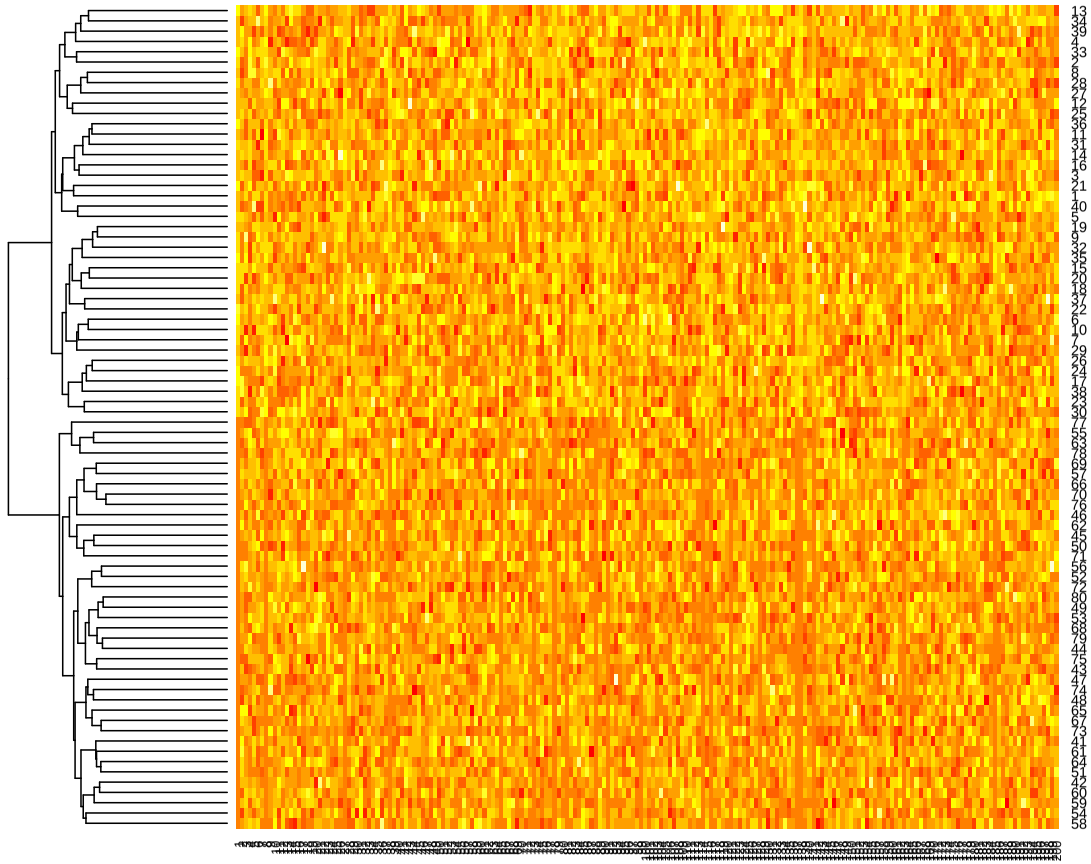
Are heatmaps OK to see interactions between variables and clusterings that are not well visible in marginal distributions?
I tried to produce counterexample but failed.





Special about heatmaps is that reordering,
although it doesn't ignore information,
hides information from human eye.

Need not to discuss "what is there",
but what impact on perceived information
the presentation of information has.



5. Some remarks on cluster validation

Gribov's confusion plots look nice
but raise question how to interpret them for cluster validation.

If two clustering methods give "about the same result",
to what extent does this confirm the clustering?
Depends on the chosen methods, and whether they are similar (like k-means and Ward) or not.

Cluster validation is often seen as a synonym for *estimating the number of clusters* (by optimising some validation criterion).

In my view, they are essentially different because a clustering with the number of clusters already estimated still needs validation.

The result of optimising a validity statistic (i.e., stability) is not necessarily itself valid (stable).

Cluster validation is often seen as a synonym for *estimating the number of clusters* (by optimising some validation criterion).

In my view, they are essentially different because a clustering with the number of clusters already estimated still needs validation.

The result of optimising a validity statistic (i.e., stability) is not necessarily itself valid (stable).

Using bootstrap in clustering can produce artifacts (multiple points; cluster separation can only increase.)