Clusternomics

Stephen Coleman

17/10/2019

Introduction

Principal aim: to model both **global** (across dataset) and **local** (dataset-specific) clustering structure. Specifically, Clusternomics (Gabasova, Reid, and Wernisch 2017) can allow for clusters merging and separating in different datasets (also referred to as **contexts**). A motivating example is shown below for two 1D datasets.

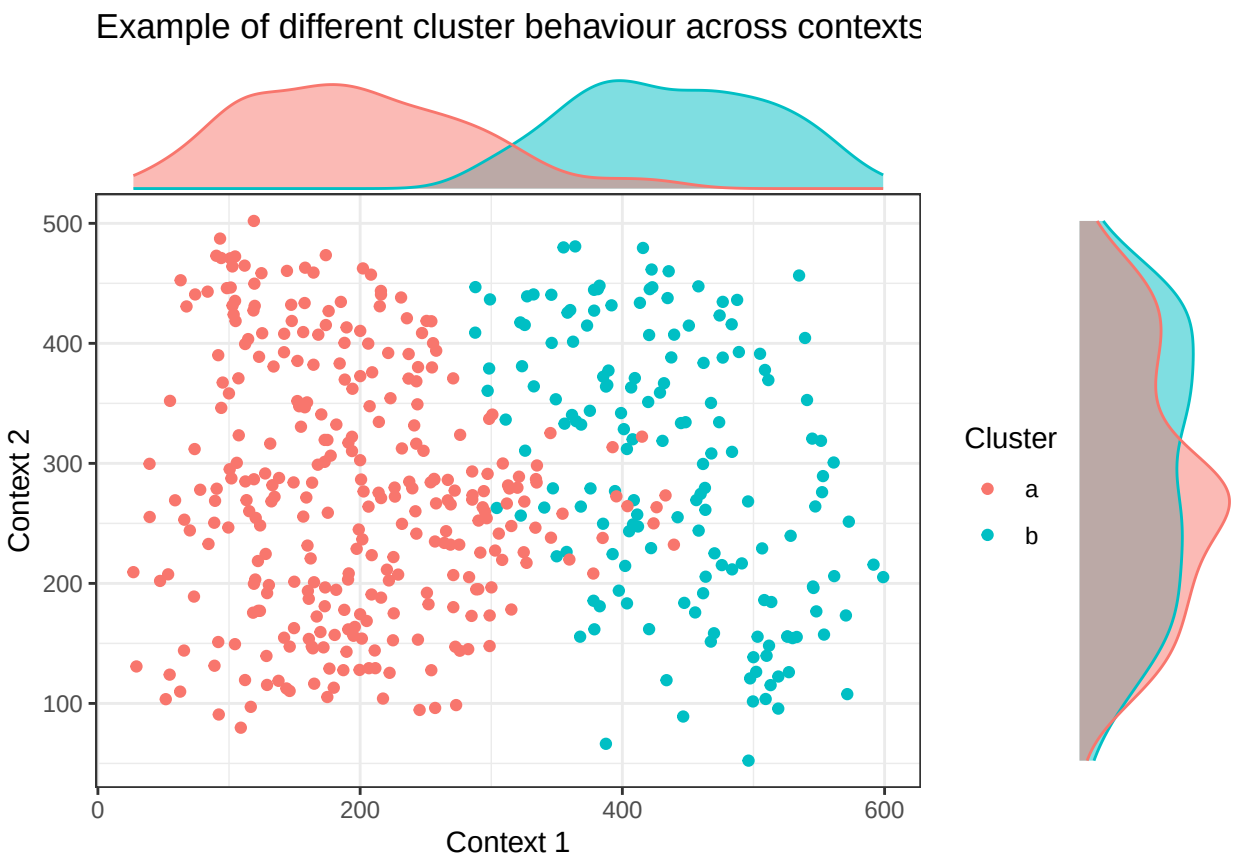
```
library(ggplot2) # ubiquitous
library(ggExtra) # for ggMarginal

# Personal preference
theme_set(theme_bw())

# Data generated at http://drawdata.xyz/
my_data <- read.csv("./example_data.csv")

# Plot the data
p1 <- ggplot(data = my_data, mapping = aes(x = x, y = y, colour = z)) +
  geom_point() +
  labs(
    title = "Example of different cluster behaviour across contexts",
    x = "Context 1",
    y = "Context 2",
    colour = "Cluster"
  )

# Add marginal density plots by grouping
ggMarginal(p1, groupColour = T, groupFill = T)
```



Here we have two separate subpopulations. In Context 1 the sub-populations are discernible as two local clusters. However this not the case in Contex 2. Clusternomics allows for such disagreement in local models while conveying the complexity of clustering structure (in this case that there really is two sub-populations present) to the global model.

The model

Clusternomics, which is embedded in the Bayesian clustering framework, uses a hierarchical Dirichlet mixture model to identify structure on both the local and the global level. The model is fit to the data via Gibbs sampling. The model does not assume that cluster behaviour will be consistent across heterogeneous datasets. This is not to assume that the clustering structure uncovered in one dataset should not inform the clustering in another dataset. This can be summarised as so:

1. Clustering structure in one dataset should inform the clustering in another dataset. If two points are clustered together in one context they should be more inclined to cluster together in other contexts.
2. Different degrees of dependence should be allowed between clusters across contexts. The model should work when datasets have the same underlying structure and also when each dataset is effecitvely independent of all others. Fundamental to this is allowing datasets to have different numbers of clusters.

To enable these modelling aims, Clusternomics explicitly represents the local clusters (i.e. dataset specific) and the global structure that emerges when considering the combination of the datasets. The global clusters are defined by combinations of local clusters. Consider the case where 3 clusters emerge in Context 1 (denoted by labels $\{1, 2, 3\}$) and 2 clusters emerge in Context 2 (denoted by labels $\{A, B\}$). In this case our global structure has the possibile form:

$$\{(1, A), (2, A), (3, A), (1, B), (2, B), (3, B)\}$$

Thus if a point is assigned a label of 1 in Context 1 and a label of A in Context 2 it increases the probability of cluster $(1, A)$ becoming populated at the global level. However, it is possible that some of the possible global clusters described above are not realised as some local clusters overlap across datasets. Consider the case that labellings 1 and 2 from the first context are captured entirely by label A in the second context with a albel of 3 corresponding perfectly to label B. In this case our global structure would take the form:

$$\{(1, A), (2, A), (3, B)\}$$

In this way the local structure informs the global structure.

References

Gabasova, Evelina, John Reid, and Lorenz Wernisch. 2017. "Clusternomics: Integrative Context-Dependent Clustering for Heterogeneous Datasets." *PLoS Computational Biology* 13 (10). Public Library of Science: e1005781.