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.

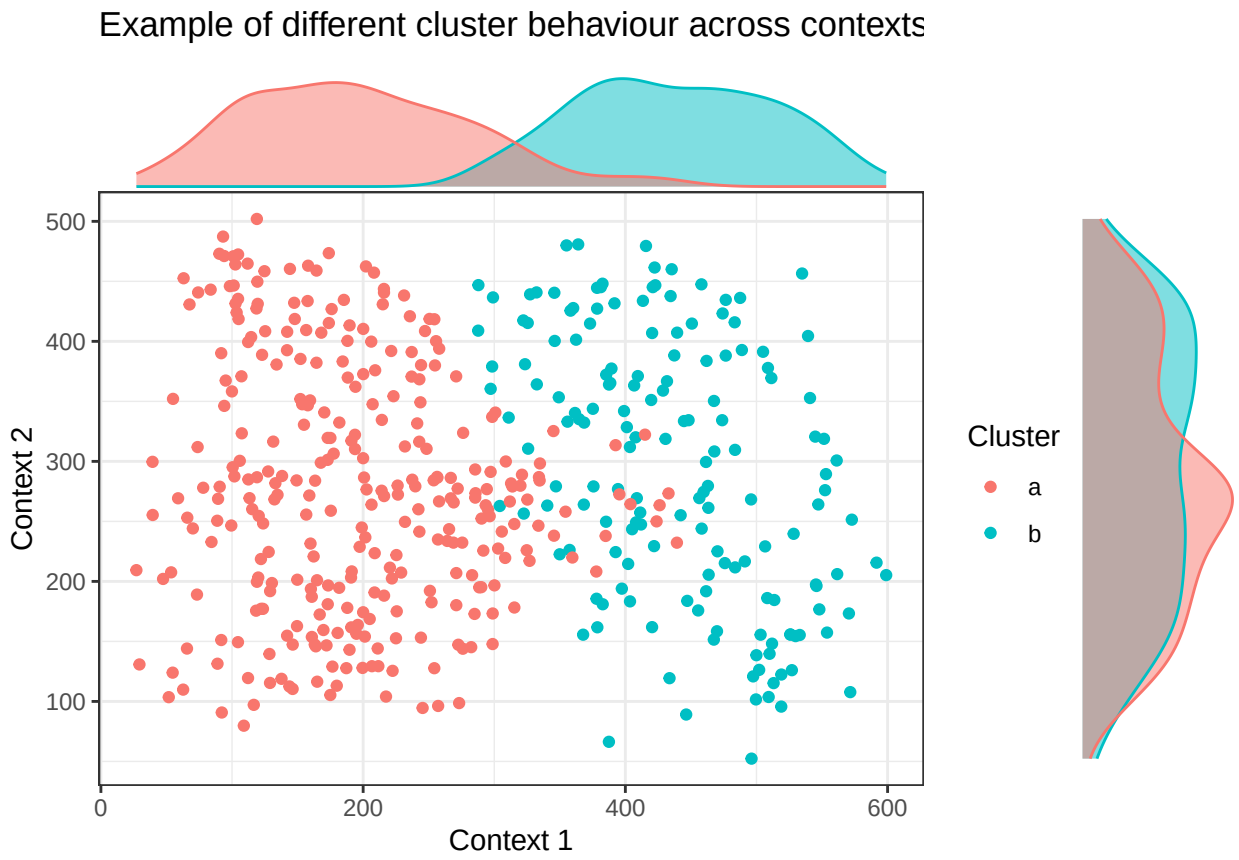
```
library(ggplot2) # ubiquitous
library(ggExtra) # for ggMarginal

# 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.

The original paper introduces two models that are "asymptotically equivalent". The first is easier to develop an intuition of, but it is the second that is implmented as it is more computationally efficient.

Notation

Let us denote the number of datasets by L and all observed data by X. Let

$$X = (X_1, \ldots, X_L),$$
$$X_l = (X_{l1}, \ldots, X_{ln})$$

where X_l is the data of the lth context.

It is assumed that it is the same n individuals in each dataset in the same order. Therefore we have L membership vectors denoting cluster membership:

$$C = (C_1, \ldots, C_L),$$
$$C_l = (c_{l1}, \ldots, c_{ln}).$$

Basic model

The basis of the integrative model is a finite approximation of a Dirichlet process known as a Dirichlet-Multinomial Allocation mixture model (Green and Richardson 2001). There is a nice explanation of this model in Savage et al. (2013).

In this case we model the latent structure in the lth dataset using a mixture of K_l components. This means that the full model density is the weighted sum of the probability density functions associated with each component where the weights, π_{lk}, are the proportion of the total population assigned to the kth components:

$$p(X_{li}|c_{li} = k) = \pi_{lk} f(X_{li}|\theta_{lk}),$$
$$p(X_{li}) = \sum_{k=1}^{K_l} \pi_{lk} f_l(X_{li}|\theta_{lk}).$$

Here K_l is the total number of clusters present and θ_{lk} are the parameters defining the kth distribution in the lth dataset.

The weights, $\pi_l = (\pi_{l1}, \ldots, \pi_{lK_l})$, follow a Dirichlet distribution with concentration parameter α_0.

The distributions in the mixture model for each dataset are:

$$
\begin{aligned}
\pi_l &\sim \text{Dirichlet}\left(\frac{\alpha_0}{K_l}, \ldots, \frac{\alpha_0}{K_l}\right) \\
c_{li} &\sim \text{Categorical}(\pi_l) \\
\theta_{lk} &\sim h_l \\
X_{li}|c_{li} = k &\sim f_l(X_{li}|\theta_{lk})
\end{aligned}
$$

where H_l is some prior distribution for parameteters for each mixture component; F_l is a probability distribution for samples given the parameters θ_{lk}. Note that each context may have different distributions and hyperparameters.

First formulation: the intuitive model

For ease of understanding, let $L = 2$. Then each context has its own mixture weights with symmetric Dirichlet priors:

$$
\begin{aligned}
\pi_1 &\sim \text{Dirichlet}\left(\frac{\alpha_1}{K_1}, \ldots, \frac{\alpha_1}{K_1}\right) \\
\pi_2 &\sim \text{Dirichlet}\left(\frac{\alpha_2}{K_2}, \ldots, \frac{\alpha_2}{K_2}\right)
\end{aligned}
$$

These weights form the basis of the local clustering within each dataset. A third mixure distribution is used to link the two local clusters. This has a Dirichlet prior over the global mixture weights, ρ, which is defined over the outer product of the local weights:

$$
\rho \sim \text{Dirichlet}\left(\gamma \ \text{vec}(\pi_1 \otimes \pi_2)\right).
$$

The outer product is defined:

$$
\begin{aligned}
\pi_1 \otimes \pi_2 &= \pi_1 \pi_2^T \\
&= \begin{pmatrix}
\pi_{11}\pi_{21} & \pi_{11}\pi_{22} & \cdots & \pi_{11}\pi_{2K_2} \\
\pi_{12}\pi_{21} & \pi_{12}\pi_{22} & \cdots & \pi_{12}\pi_{2K_2} \\
\vdots & \vdots & \ddots & \vdots \\
\pi_{1K_1}\pi_{21} & \pi_{1K_1}\pi_{22} & \cdots & \pi_{1K_1}\pi_{2K_2}
\end{pmatrix}
\end{aligned}
$$

Then the vector function takes the column vectors of the matrix and places them in one vector:

$$
\text{vec}(\pi_1 \otimes \pi_2) = \begin{pmatrix}
\pi_{11}\pi_{21} \\
\vdots \\
\pi_{1K_1}\pi_{21} \\
\pi_{11}\pi_{22} \\
\vdots \\
\pi_{1K_1}\pi_{22} \\
\pi_{11}\pi_{23} \\
\vdots \\
\pi_{1K_1}\pi_{2K_2}
\end{pmatrix}
$$

This is the basis for the non-symmetric concentration parameter of the Dirichlet distribution over the global mixture weights, ρ.

A global membership variable, c, is drawn from a Catgeorical distribution with concentration parameter ρ:

$$\begin{aligned} c_i &= (c_{1i}, c_{2i}) \\ c &\sim \text{Categorical}(\rho) \\ X_{li}|c_{li} = k &\sim f_l(X_{li}|\theta_{lk}), l \in \{1, 2\} \end{aligned}$$

In this way the local clusters, c_{li}, in the lth context are projections of the global clustering, c_i, onto this space.

The model achieves the property that the local assignment should affect the global assignment posterior probability. This can be seen as local assignments affect the posterior probability of the context-specific weights, π_l, which in turn define the probabilities of the global combinatorial clusters through the hierarchical model (a change in the membership of the kth component in the lth dataset affects an entire slice of column vectors in the tensors defining the concentration parameter for the global component weights, ρ).

The model also represents different degrees of dependence by allowing any combination of cluster assignments across contexts. When there is a single common cluster structure across the two contexts, the occupied clusters will be concentrated along the diagonal of the probability matrix of ρ.

When $L > 2$ then the outer product of the local componenet weights generalises to a tensor product. In this case the hierarchical model is given by:

$$\begin{aligned} \pi_l|\alpha_0 &\sim \text{Dirichlet}\left(\frac{\alpha_0}{K_l}, \ldots, \frac{\alpha_0}{K_l}\right) \\ \rho|\gamma, \{\pi_1 \ \ldots, \pi_L\} &\sim \text{Dirichlet}\left(\gamma \left\{\bigotimes_{l=1}^{L} \pi_l\right\}\right) \\ c_i|\rho &\sim \text{Categorical}(\rho), c_i = (c_{1i}, \ldots, c_{Li}) \\ \theta_{lk} &\sim h_l \\ X_{li}|c_{li} = k &\sim f_l(X_{li}|\theta_{lk}) \end{aligned}$$

The model for L contexts has a scaling issue as each additional dataset brings an additional layer of calculations in the probability tensor. The second model is designed to reduce this cost by avoiding calculations for uninhabited components.

Second formulation: the quick model

This model attempts to reduce the number of combinations required. This requires decoupling the number of local clusters and the number of global clusters. A mixture over S global clusters is defined:

$$\begin{aligned} \rho|\gamma_0 &\sim \text{Dirichlet}\left(\frac{\gamma_0}{S}, \ldots, \frac{\gamma_0}{S}\right) \\ c_i|\rho &\sim \text{Categorical}(\rho) \end{aligned}$$

where ρ is the global mixture weights and c_i is the components assignment indivator variable as before. A variable z_{ls} is then defined that associates the sth global cluster with context specific clusters:

$$\begin{aligned} \pi_l|\alpha_0 &\sim \text{Dirichlet}\left(\frac{\alpha_0}{K_l}, \ldots, \frac{\alpha_0}{K_l}\right), \\ z_{ls}|\pi_l &\sim \text{Categorical}(\pi_l). \end{aligned}$$

Here $z_{ls} \in \{1, \ldots, K_l\}$ assignes the sth global cluster to the a specific local cluster in the lth dataset. One may consider $(z_{sl})_{s=1}^{S}$ as the coordinates of the global clusters in the lth dataset. This link means that the c_i maps a point to a global cluster as well as the local clusters.

In the previous model the mapping of global clusters to local clusters was implicit, because each combination of context clusters mapped to a unique global cluster. In this model, the mapping is probabilistic and forms a part of the model. One may now state S, the number of global components, and thus limit the number of computations required in contrast to the preceding model.

This hierarchial model is defined by:

$$\begin{aligned}
\rho|\gamma_0 &\sim \text{Dirichlet}\left(\frac{\gamma_0}{S}, \ldots, \frac{\gamma_0}{S}\right), \\
c_i|\rho &\sim \text{Categorical}(\rho), \\
\pi_l|\alpha_0 &\sim \text{Dirichlet}\left(\frac{\alpha_0}{K_l}, \ldots, \frac{\alpha_0}{K_l}\right), \\
z_{ls}|\pi_l &\sim \text{Categorical}(\pi_l) \\
\theta_{kl}|h_l &\sim h_l \\
X_{li}|c_i, (z_{sl})_{s=1}^{S}, (\theta_{lk})_{k=1}^{K} &\sim f_l(X_{li}|\theta_{lk_{c_i}})
\end{aligned}$$

Inference

The quick, decoupled model is implemented in the R package `clusternomics` and uses Gibbs sampling as the inference algorithm.

References

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

Green, Peter J, and Sylvia Richardson. 2001. "Modelling Heterogeneity with and Without the Dirichlet Process." *Scandinavian Journal of Statistics* 28 (2): 355–75.

Savage, Richard S, Zoubin Ghahramani, Jim E Griffin, Paul Kirk, and David L Wild. 2013. "Identifying Cancer Subtypes in Glioblastoma by Combining Genomic, Transcriptomic and Epigenomic Data." *arXiv Preprint arXiv:1304.3577*.