Clusters and Heatmaps

Jeff Oliver

14 August, 2017

Want to combine the visualization of quantitative data with clustering algorithms? Unsatisfied with the options provided by the base R packages? In this hands-on workshop, we'll use the `ggendro` package for R to make publication-quality graphics.

- Consider the `ggdendro` package? See: https://stackoverflow.com/questions/6673162/reproducing-lattice-dendrogram-gra
- ggdendro page

0. What are the data? (the otter data?)
1. Clustering approach (e.g. `hclust`)
2. Drawing just a cluster
3. Drawing just a heatmap (`geom_tile`); individuals on Y, measurement on X

Learning objectives

1. one
2. two
3. three

[DESCRIPTION OR MOTIVATION; 2-4 sentences that would be used for an announcement]

Getting started

Start by creating a new project in RStudio and creating two folders we'll use to organize our efforts. The two folders should be `data` and `output` and will store… data and output.

```
dir.create("data")
dir.create("output")
```

- Download data file from https://jcoliver.github.io/learn-r/data/otter-mandible-data.xlsx or http://tinyurl.com/otter-data (the latter just re-directs to the former). These data are a subset of those used in a study on skull morphology and diet specialization in otters doi: 10.1371/journal.pone.0143236.
- Open this file, otter-mandible-data.xlsx, in spreadsheet program like Microsoft Excel® or LibreOffice Calc.
- Save a copy of the file as a CSV (comma-separated values) file named 'otter-mandible-data.csv' in the data folder you created above:
  - In MS Excel®, select File > Save As… and in the dialog that appears, select CSV from the type dropdown menu.
  - In LibreOffice Calc, select File > Save As… and in the dialog that appears, select Text CSV (.csv) in the Format dropdown in the lower-right portion of the dialog.

```
otter <- read.csv(file = "data/otter-mandible-data.csv", header = TRUE)
```

Missing data can cause problems in downstream analyses, so we will just remove any rows that have missing data. Here we replace the original data object `otter` with one in which there are no missing values. Note, this *does not* alter the data in the original file we read into R; it only alters the data object `otter` currently in R's memory.

```r
otter <- na.omit(otter)
```

And because R *does not* automatically re-number the rows when we drop those with `NA` values, we can force re-numbering via:

```r
rownames(otter) <- NULL
```

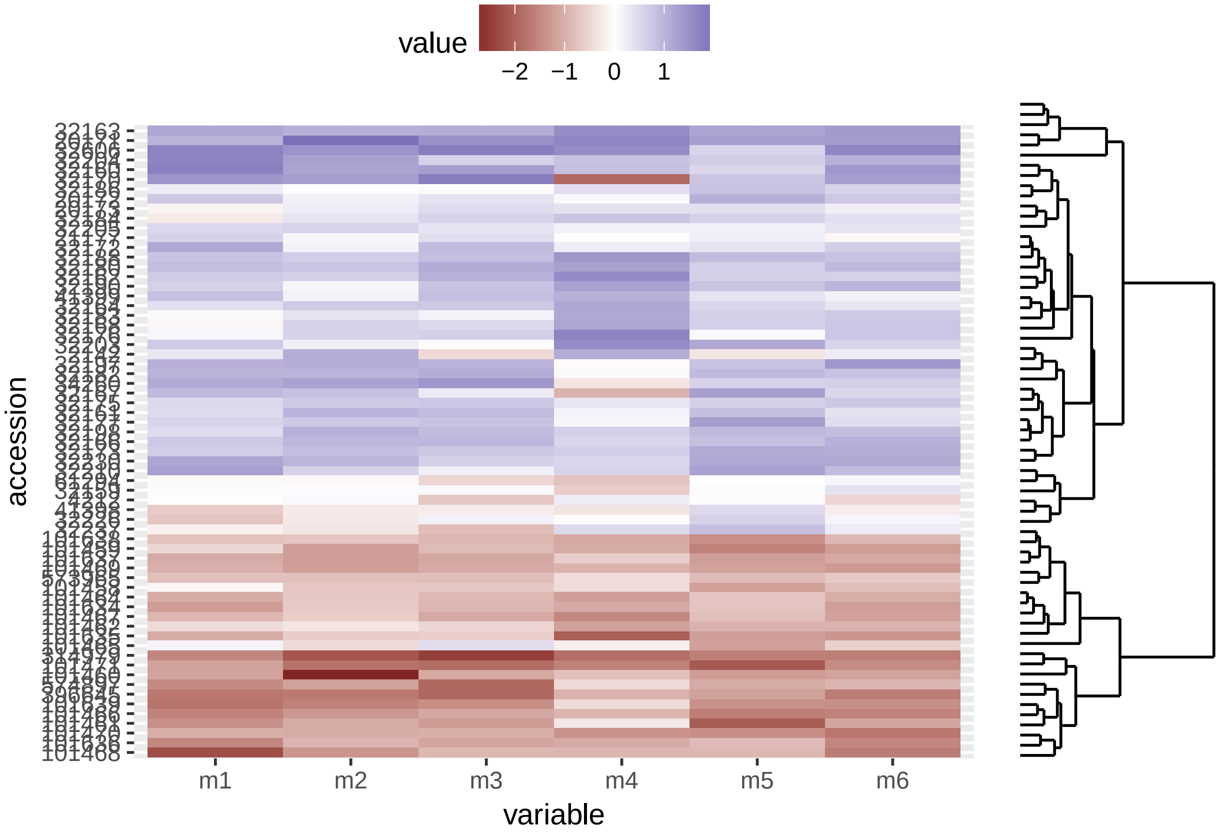
Clustering

Heatmap

Putting it all together

Final-ish

```r
otter <- read.csv(file = "data/otter-mandible-data.csv", header = TRUE)
two.species <- c("A. cinerea", "L. canadensis")
otter <- otter[otter$species %in% two.species, ]
otter <- na.omit(otter)
otter.scaled <- otter
otter.scaled[, c(4:9)] <- scale(otter.scaled[, 4:9])
otter.scaled$accession <- factor(otter.scaled$accession)

# Renumber rows
rownames(otter) <- NULL

# Run clustering
library("ggdendro")
otter.matrix <- as.matrix(otter.scaled[, -c(1:3)])
otter.dendro <- as.dendrogram(hclust(d = dist(x = otter.matrix)))
otter.dendro.data <- dendro_data(otter.dendro)
otter.order <- order.dendrogram(otter.dendro)

# Create dendro
library("ggplot2")
dendro.plot <- ggplot(data = segment(otter.dendro.data)) +
  geom_segment(aes(x = x, y = y, xend = xend, yend=yend)) +
  coord_flip() +
  theme_dendro()

# Heatmap
# Data wrangling
library("reshape2")
otter.long <- melt(otter.scaled, id = c("species", "museum", "accession"))
# Order the levels according to their position in the cluster
otter.long$accession <- factor(x = otter.long$accession, levels = otter.scaled$accession[otter.order], 
heatmap.plot <- ggplot(data = otter.long, aes(x = variable, y = accession)) +
  geom_tile(aes(fill = value)) +
  scale_fill_gradient2() +
  theme(legend.position = "top")

# All together
library("grid")
grid.newpage()
print(heatmap.plot, vp = viewport(x = 0.4, y = 0.5, width = 0.8, height = 1.0))
print(dendro.plot, vp = viewport(x = 0.90, y = 0.445, width = 0.2, height = 0.92))
```



Additional resources

- resource one
- resource two

Back to learn-r main page

Questions? e-mail me at jcoliver@email.arizona.edu.