DSK805 - Project

Henry Kirveslahti

Førår 2025

Evolutionary Anthropology

Background story

The topic of this analysis is primate molars and it is related to state-of-the-art research in evolutionary anthropology, specifically, *geometric morphometrics*.

Our data was collected by D.M.Boyer and it was first analyzed in [1]. This paper was an important breakthrough in the field that had been out-gunned by the gene expression data, which is abundant and easy to analyze. Our dataset builds on this story, but decade later, we have access to even fancier tools to analyze the data.

The Data

We have a dataset of 116 monkey molars. These are geometric objects, so we cannot compute their means. Luckily, we can extract numerical summaries of the data, and in our dataset we have four such summaries. These are as follows:

1. `File` This is just a file name that bears no value to our analysis (some sort of id)
2. `DNE` Dirichlet Normal Energy of the CT-scan of the tooth. This is a one-number summary that measures overall curvature of the tooth.
3. `positiveDNE` The positive component of the DNE above. Loosely speaking, this measures some sort of overall elliptic principal curvature (i.e. hills and valleys)
4. `surfacearea` The surface area of the mesh
5. `positivesurfacearea` The part of the surface area amounting to positive curvature (hills and valley type behavior)
6. `Diet`: The preferred diet of the monkey. We use these as the categories. Frugivore: fruits, Folivore: leaves, Omnivore: anything: Insectivore: Insects, Frugi-Folivore: Both fruits and insects.
7. `Genus`: The genus of the monkey - (some sort of species)

The data is available in the .csv file `data.csv`. The numbers were computed with signDNE algorithm using bandwidth parameter 0.1 [2] (which was developed last year here at SDU!)

You don't need to be expert on these numbers, as far as we are concerned, these are just numbers that tell something about the shape of the teeth - There are different kinds of curvature, for example, saw-like features are useful for processing leaves, more coarse geometry may be useful for cracking nuts. Also, this black-boxing has the advantage that we don't need to care about interpretation of the variables - we just explore and classify. (This is actually pretty close to what you might encounter if you did research in an interdisciplinary group.)

Some of the molars are of monkeys that are long extinct. This means that we don't know what kind of food they actually consumed. But we may contemplate this question, for it would be reasonable to believe that the shape of the molar is related to the dietary preference of its carrier. This is the subject of this project.

Your job

Your job is to explore the data and figure out some details related to the analysis.

0. You first job is to choose what you want to study. Some of the diets in the table are classified as NAs - this means that the monkey (the genus) in question is extinct, and there is no biological concensus (or good enough evidence) to put a label on them. Your ultimate job is to speculate what kind of food a particular extinct monkey might have consumed. So pick a genus that you would like to predict. If you are out of ideas, personally I would think Teilhardina is an interesting one, but you may choose any one of them.

1. Your main job is to speculate what kind of food did the monkey (genus) that you chose consume. Using a suitable classification model, predict the diets of these samples, (to obtain one prediction per sample in the genus that you chose), and from this prediction, contemplate on the question. Note: to do this, you should remove the species with NA diet from your analysis, and then do the prediction on the class that you are interested in. Also: Your model gives predictions to all the individuals, and you might get different diets for different individuals. There is nothing wrong with that, but if the predictions are all over the place, maybe the conclusion is that we are not sure. Also, you should read question 2 below before you tackle question 1.

2. The dataset describes some of the monkeys as being Frugivore-Folivores, meaning that they eat both fruits and leafs. In terms of the data, are the reasons to believe that this group is sufficiently different (in terms of the presented numbers) from Folivores? If not, you might consider merging these two groups together (and do this before you classify in Problem 1) Use a suitable analysis tool to answer this question. (You might for example compare if the means are different, and if you want to use LDA, see if it would be reasonable to see if they are normally distributed. You can probably make a case for either decision.)

This is a classification task: You might for example use LDA, QDA, logistic regression. (You might also try something else you have learned on some other course, if you want to.) The data is pretty clean and nice though, it should be amenable to simple methods. Simplicity is good.

Assignment rules:

1. You can work in groups (This is preferred). Free collaboration, free everything. One group submits one report, mention who are part of the group in the report. Max group size 5 - if you want to work in a larger group, you can just divide yourself into two groups. Across-group cooperation is welcome and encouraged, so you can still talk to your friends.

2. Report size should be max 6 pages - but aim for shorter. Include in the report, and communicate as clearly as possibly:

   (a) What did you do (i.e. what method and why [No need for hardcore sensitivity analysis])

   (b) What is the conclusion (what do you predict for the diet)

   (c) Relevant graphical summaries

No need to write any literature review or show code or specify mathetmatics. Just analyze the data and explain what you see.

This is supposed to be just a fun little project to get confidence in the tools. We have stressed about lot of technical things in the course - don't stress too much about this project.

The deadline for submission is June 2. (If you want feedback before exam, submit before May 30 5 am.). **Make sure to clearly indicate the names of all the group members.**

References

[1] Boyer et al. *Algorithms to automatically quantify the geometric similarity of anatomical surfaces*, Proceedings of the National Academy of Sciences, 108(45):18221–18226, 2011

[2] Hjerrild, Shan, Boyer and Daubechies *signDNE: A python package for ariaDNE and its sign-oriented extension* code available at https://github.com/frisbro303/SignDNE/blob/main/docs.md#Using-SignDNE-as-a-Python-library