School of Engineering and Applied Science (SEAS), Ahmedabad University

B.Tech (ICT Semester IV): Probability and Random Processes (MAT 202)
Special Assignment Abstract Submission #2
Submission Deadline: January 31, 2020 (11:59 PM)

- **Group No.:** H_B29
- **Project Area:** Biology
- **Project Title:** Seriation in Paleontological Data Using Markov Chain Monte Carlo Methods
- **Group members:**
  1. Yashraj Kakkad (AU1841036)
  2. Prayag Savsani (AU1841035)

Abstract

Problem Background

We are given a collection of fossil sites with information about the taxa (species) that occur in each site. Our task is to perform "Seriation" - the task of chronology. We intend to figure out appropriate estimates for the temporal (time-based) ordering of these sites. Seriation is considered a fundamental problem in Palaeontology - the branch of science concerned with fossil animals and plants.

Large databases of fossil sites and their corresponding occurrences of taxa are available today. Conventional methods of seriation (biostratigraphy) are not easy to apply in large datasets, because of the limited age-related information. Therefore, there is an increasing need to opt for newer methods which do not rely solely on them.

Brief of the base article

The base article[1] describes a simple probabilistic model which estimates site ordering given taxon occurences and other kinds of readily available stratigraphic information.

A fossil site would have several taxa living at the same time. However, the mere presence or absence of taxa cannot help in accurately determining their chronology. Some taxa will be sparsely founded in a site, while in another it might be densely predominant. Similarly, there might be errors in available information which needs to be considered from a mathematical viewpoint.[1]

As discussed, the recent advent of large, readily available fossil databases and the increased influence of probabilistic methods can help solve this problem. These computational solutions have become feasible in the past decade as an effective solution.[1]

Theoretically, we could find the set of parameters which maximize the likelihood of the data. But, the probability of one site pre-dating another might be close to 0.5, in which case, the seriation is uncertain. Therefore, we resort to an analytical approach - Markov Chain Monte Carlo Simulations to find multiple sets of parameters such that their posterior distribution can be calculated.

Planned contributions

Our first goal in mind is to extract relevant information from publicly available fossil databases using pandas (a Python data analysis library). The information can be operated on using NumPy. We are going to use the European late Cenozoic large land mammals database (http://www.helsinki.fi/science/now). We will understand the concept of Markov Chain Monte Carlo simulations to model the uncertainty. We will implement the model in Python using the relevant packages.

If time permits, we will generate an artificial dataset with predefined parameters and error probabilities and testify our model. We also wish to express our inferences using appropriate graphical tools.

References

[1] K. Puolamäki, M. Fortelius, and H. Mannila, "Seriation in paleontological data using markov chain monte carlo methods," *PLOS Computational Biology*, vol. 2, pp. 1–9, 02 2006.