Statistical Methods, Final Assignment

Università della Svizzera italiana

Due Date: 06/25/2024

This assignment is to be completed using a Jupyter Notebook. Your code, figures, and analyses should be documented clearly within the notebook.

Simulation and Learning

You will use stochastic methods to simulate evolutionary processes and train systems that learn from these simulations. You have two alternatives to choose from:

- Augmenting evolutionary trees and then training a Generalized Additive Model (GAM)
- Simulating full evolutionary trees and then training a Neural Network (NN)

Additionally, you will select one of the following species diversification models for your simulations:

- BiSSE (Binary State Speciation and Extinction)
- MuSSE (Multiple State Speciation and Extinction)
- QuaSSE (Quantitative State Speciation and Extinction)
- GeoSSE (Geographic State Speciation and Extinction)
- BiSSEness (BiSSE-Node Enhanced State Shift)
- ClaSSE (Cladogenetic State Speciation and Extinction)
- Diversity-Dependent Model formulated as:

$$\lambda_t = \max\{(\lambda_0 - \beta_N N_t), 0\}$$

Choose one of the following approaches:

Option 1: Augmenting Trees and Training a GAM

In this approach, you will extend the evolutionary trees generated in the previous assignment with additional simulated data to train a Generalized Additive Model (GAM). The GAM will be used to predict the likelihood function as a function of the model parameters. Follow these steps:

1. Select one of the species diversification models listed above. Describe the chosen model in detail, including the dynamics and the probabilities involved in the process.
2. Simulate additional evolutionary trees by varying the parameters according to the parameter grid specific to the chosen model.
3. Choose one of those simulated trees. Use the simulated trees to augment the dataset, as in the previous assignment.
4. Estimate the joint probability as done in the previous assignment, for different sets of parameters.
5. Train a GAM using the augmented dataset to predict the likelihood function based on the parameters.
6. Implement Differential Evolution (DE) or Stochastic Gradient Descent (SGD) to optimize the GAM and retrieve the appropriate parameters of the tree.
7. Evaluate the performance of the GAM by comparing its predictions to the known rates used in the simulations.

Option 2: Simulating Full Trees and Training a NN

In this approach, you will simulate entire evolutionary trees and use them to train a Neural Network (NN). Follow these steps:

1. Select one of the species diversification models listed above. Describe the chosen model in detail, including the dynamics and the probabilities involved in the process.
2. Simulate a diverse set of full evolutionary trees with varying parameters according to the parameter grid specific to the chosen model.
3. Encode the trees in a suitable format for neural network input (e.g., summary statistics, lineage through time plots).
4. Split the dataset into training, validation, and test sets.
5. Design and train a Neural Network to predict the parameters from the encoded trees.
6. Evaluate the NN's performance on the test set by comparing predicted parameters to the true parameters used in the simulations.
7. Comment your results.

Comparative Analysis for Advanced Students

For advanced students who want to explore both approaches, perform a comparative analysis:

1. Compare the predictive accuracy of the GAM and NN models.
2. Discuss the computational efficiency and scalability of each method.
3. Analyze the strengths and weaknesses of each approach.

Appendix: Detailed Description of Species Diversification Models

BiSSE (Binary State Speciation and Extinction)

Dynamics: The BiSSE model integrates the constant-rate birth-death process with a two-state Markov model, allowing species to exist in one of two states. Speciation (λ_1, λ_2), extinction (μ_1, μ_2), and transition rates (q_{12}, q_{21}) between states are constant over time.

Parameters:

$$\begin{aligned}
\text{Speciation rates: } & \lambda_1, \lambda_2 \\
\text{Extinction rates: } & \mu_1, \mu_2 \\
\text{Transition rates: } & q_{12}, q_{21}
\end{aligned}$$

MuSSE (Multiple State Speciation and Extinction)

Dynamics: The MuSSE model generalizes BiSSE to multiple states, allowing for the analysis of traits with more than two discrete states. Each state has its own speciation and extinction rates, and transition probabilities between states.

Parameters:

$$\begin{aligned}
\text{Speciation rates: } & \lambda_i \quad \text{where} \quad i = 1, ..., K \\
\text{Extinction rates: } & \mu_i \quad \text{where} \quad i = 1, ..., K \\
\text{Transition rates: } & q_{ij} \quad \text{where} \quad i, j = 1, ..., K; \; i \neq j
\end{aligned}$$

QuaSSE (Quantitative State Speciation and Extinction)

Dynamics: The QuaSSE model allows speciation and extinction rates to vary as continuous functions of a quantitative trait. This model uses trait values that change gradually over time, as opposed to discrete states.

Parameters:

$$\begin{aligned}
\text{Speciation rate function: } & \lambda(x) \\
\text{Extinction rate function: } & \mu(x)
\end{aligned}$$

Trait evolution model: typically modeled as a Brownian motion with a drift parameter.

GeoSSE (Geographic State Speciation and Extinction)

Dynamics: The GeoSSE model combines geographic range dynamics with speciation and extinction processes. Species can occupy region A, region B, or both (AB), and speciation can occur within a region or between regions.

Parameters:

Speciation rates within regions:	s_A, s_B
Speciation rate between regions:	s_{AB}
Extinction rates:	x_A, x_B
Dispersal rates:	d_A, d_B

BiSSEness (BiSSE-Node Enhanced State Shift)

Dynamics: BiSSEness enhances the BiSSE model by incorporating state shifts during speciation events. It differentiates between changes occurring along lineages (anagenetic) and those occurring during speciation events (cladogenetic).

Parameters:

Speciation rates:	λ_0, λ_1
Extinction rates:	μ_0, μ_1
Transition rates:	q_{12}, q_{21}
Cladogenetic shift probabilities:	p_{0c}, p_{1c}
Anagenetic shift probabilities:	p_{0a}, p_{1a}

ClaSSE (Cladogenetic State Speciation and Extinction)

Dynamics: ClaSSE distinguishes between anagenetic evolution, which occurs within a lineage, and cladogenetic evolution, which occurs at speciation events. It models the rates of these processes based on character states.

Parameters:

Cladogenesis rates:	λ_{klm}	where k, l, m represent different character states
Extinction rates:	μ_k	for character state k
Transition rates:	q_{lm}	from state l to m