Simulation study on parameter inference of multi-variate Ornstein–Uhlenbeck-type trait evolution models

Hao Chi Kiang <hao.chi.kiang@liu.se>

October 22, 2021

1 Background

Suppose n species (say, human, Neanderthal, chimpanzee, and so on) have evolved from a common ancestry according to a known phylogenetic tree, and we have measured a p-dimensional trait vector on each of the species, say, (skull size, body length) can be a trait vector. A commonly used model for this kind of data is to assume that the p-dimensional trait vector of each species has evolved independently once they had branched off from their common ancestor; and throughout the independent evolution the trait vector evolves according to the following stochastic differential equation:

$$dx_t = -H(x_t - \theta)dt + \Sigma dW_t$$

where t is time, W_t is the standard Brownian motion and (H, θ, Σ) are parameters. The p-vector θ represents an evolutionary optimum, square matrix H is usually assumed to be at least positively definite, and Σ is a Cholesky factor of a symmetric positively definite matrix. Note that $x_{t+\Delta t}|x_t$ is always Gaussian because W_t is a Gaussian process. Intuitively speaking, the model means that all the traits are doing a continous-time random walk but constantly being be "pulled" toward the optimum θ; once it has arrived near the optimum, the trait will wander around this optimum randomly. This stochastic process is called the Ornstein-Uhlenbeck (OU) process.

The likelihood of the model can be computed quickly[2], so this model can be estimated by the maximum-likelihood estimator. However, it has been observed that the result of fitting this model can be misleading when n is not much bigger than p[3]. For example, Cooper et al. [1] wrote that "the OU model is frequently incorrectly favoured over simpler models when using Likelihood ratio tests, and that many studies fitting this model use datasets that are small and prone to this problem;" This corresponds to the situation where H is in fact the zero matrix but it was "overfitted" because of insufficient data. It is therefore interesting to ask how large the data must be, and how the MLE and the geometry around the MLE behaves when there are not a lot of species.

2 Objective

1. Perform a simulation study (or potential analytical analysis) on how the maximum likelihood of the OU model behaves when the number of species, tree topology, number of traits, and distribution of the data changes.
2. Visualize, analyze, and summarize the findings.

3 Data

The student is expected to simulate data themselves.

4 Required student background

For this project, we seek students who

1. are capable of designing, performing, and diagnosing numerically intense computing;
2. have solid knowledge in linear algebra and frequentist estimation theory.

Formal training in stochastic differential equation is not strictly required.

5 Contact

You could contact Hao Chi Kiang <hao.chi.kiang@liu.se> if you have any questions.

References

[1] Cooper, N., Thomas, G. H., Venditti, C., Meade, A., & Freckleton, R. P. (2016). A cautionary note on the use of Ornstein Uhlenbeck models in macroevolutionary studies. Biological Journal of The Linnean Society, 118(1), 64–77.

[2] Mitov, V.,Bartoszek K., Asimomitis G., Stadler T. Fast likelihood calculation for multivariate Gaussian phylogenetic models with shifts. Theoretical Population Biology, 131, 66–78.

[3] Adams, D. C., & Collyer, M. L. (2018). Multivariate Phylogenetic Comparative Methods: Evaluations, Comparisons, and Recommendations. Systematic biology, 67(1), 14–31