A new method of discriminant analysis, phylogeny-informed and applicable in large dimension.

Anaïs Duhamel*[1]

[1]Laboratoire de Géologie de Lyon - Terre, Planètes, Environnement [Lyon] – École Normale Supérieure - Lyon, Université Claude Bernard Lyon 1, Centre National de la Recherche Scientifique – France

Résumé

Understanding the link between ecology and morphology is a key issue, notably in paleontology since understanding this relationship for extant species allows the inference of extinct species ecologies from their osteological features. A particularly adapted statistical tool for such predictions is discriminant analysis. Indeed, this method first fits a model on a training set containing individuals for which both categorical traits (e.g. ecological classes) and continuous traits (e.g. morphology) are known, to be then able to predict the class of an individual for which only continuous traits are known. However, morphology is not the only signal which can help to infer past ecologies. Indeed, when the issue is classification and not studying the structure/function relationship stricto sensu, the phylogenetic position of an individual can be highly informative since closely related species often share the same ecology. Still, the only attempt of phylogenetic discriminant analysis so far removes the phylogenetic signal from the dataset, and thus potentially discards an informative signal for ecological classes discrimination. Moreover, with the rise of 2D and 3D geometric morphometrics, datasets with more traits than species are now commonplace, but classical discriminant analysis methods significantly lose statistical power when the number of morphological traits (p) approaches the number of species (n), and are not even computable when p is higher than n. For now, there exists no discriminant analysis method that both includes the phylogenetic signal (instead of removing it) and is applicable to high dimensional datasets, i.e. when p is higher than n. We thus propose here a new method of discriminant analysis which is both phylogeny-informed and "penalized" in order to be computable even in high dimensions. The performances of this newly implemented method was assessed on simulated and empirical datasets, compared with pre-existing methods of linear discriminant analysis (LDA). It appears that this new method performs at least as good as other LDA, performs better in datasets showing a non-null phylogenetic signal, and performs even better when p gets closer to n. In the light of these encouraging results, it seems recommended to use the penalized phylogeny-informed method instead of other LDA, and this regardless of the dimensionality of the dataset and of the strength of its phylogenetic signal.

*Intervenant