Phylofactorization - theory and challenges

Alex D. Washburne[1]
[1]Duke University; *alex.d.washburne@gmail.com*

Abstract

Data from biological communities are compositions whose parts are connected by an important sequential binary partition - the "phylogeny", or evolutionary history of the parts. Compositional data with a natural sequential binary partition suggest the isometric log-ratio transform as a means of analyzing community ecological data. Balances in an ilr transform of the phylogeny will correspond to nodes and contrast abundances of sister clades, but traits, such as the wings of birds, arise along edges and so a natural contrast may not be between sister clades but between organisms with and without a trait. A greedy algorithm - 'phylofactorization' - was developed to construct an ilr transform whose balances correspond to edges along which traits arose, thereby contrasting birds to non-birds as opposed to contrasting birds to crocodiles.

In this paper, the general theory of phylofactorization is presented as a graph partitioning algorithm. A special case - regression phylofactorization - chooses ilr coordinates based on sequential maximization of objective functions from regression. The connections between regression phylofactorization and other methods is discussed, including matrix factorization, hierarchical regression, factor analysis and latent variable models. Open challenges in the statistical analysis of phylofactorization are presented, including criteria for choosing the number of factors and approximating null-distributions of commonly used test-statistics and objective functions. As a graph-partitioning algorithm, cross-validation of phylofactorization across datasets requires graph-topological considerations, such as how to deal with novel nodes and edges and whether or not to control for partition order. These challenges carry major implications for the biological sciences and are a promising area of future work.

Key words: compositional data, community ecology, isometric log-ratio, phylogeny, phylofactorization, graph-partitioning, greedy algorithm, regression, cross-validation