Distance based linear regression model and its application to microbiome association studies

Masoumeh Sheikhi Kiasri

Abstract

In the past few decades, pairwise distance based statistical methods have been developed to identify spatial and/or temporal clusters of disease, study the association between the dissimilarity of ecological communities and distance in geographical locations. With emergence of high-throughput technologies, pairwise distance base methods are widely used in the analysis of genetics and genomics data, especially when the data structure fails the fundamental assumptions of classical multivariate analysis. However, much of existing knowledge has been around employing permutation techniques, which are known to be computationally expensive and sensitive to the choice of permutation. Majority of this thesis focuses on linear regression of pairwise distance matrices. We consider the pairwise correlation structure between the distances and investigate the large sample properties of the ordinary least square estimator of the model coefficients. Another major component of the thesis is the human microbiome data analysis. We analyze the integrative Human Microbiome Project (iHMP) data set of composition of microbial communities in the digestive tracts of humans by using multiple statistical methods, including our proposed method.

Department of Mathematics, UNC Charlotte, Charlotte, NC 28223