Modeling incomplete compositional datasets

Andriette Bekker[1], **Jason Pillay**[1], **Antonio Punzo**[2], **and Cristina Tortora**[3]

[1] *University of Pretoria, South Africa*
[2]*University of Catania, Italy*
[3]*San Jose State University, California, USA*

Abstract

Incomplete compositional data analysis is hindered by the lack of likelihood-based methods that directly handle missing proportions on the simplex. This paper introduces an estimation technique that operates directly on the original data, thereby accommodating missing values while preserving the interpretability of the variables within their natural domain. Simulation studies demonstrate robust performance in parameter estimation and imputation across missingness mechanisms. To showcase its practicality, the proposed algorithm is applied to two real datasets exhibiting distinct missingness patterns. Analysis of the American Time Use dataset (with values missing at random) identifies six distinct categories of daily activities, yielding interpretable profiles of daily activity patterns and the demographics associated with them. Application to PM2.5 species data from the Air Quality System, supports a four-component mixture model, providing a descriptive profile of measured PM2.5 speciation across the United States.

Keywords

Compositional data, Missing data imputation, Mixture models
Mixture models

.