Statistica Sinica **33** (2023), 1-26
doi:https://doi.org/10.5705/ss.202021.0147

ROBUST SHAPE MATRIX ESTIMATION FOR HIGH-DIMENSIONAL COMPOSITIONAL DATA WITH APPLICATION TO MICROBIAL INTER-TAXA ANALYSIS

Danning Li[1], Arun Srinivasan[2], Lingzhou Xue[2] and Xiang Zhan[3]

[1]*Northeast Normal University,* [2]*Pennsylvania State University and* [3]*Peking University*

Abstract: Estimating the dependence structure in the data is a key task when analyzing compositional data. Real-world compositional data sets are often complex owing to high-dimensionality, heavy tails, and the possible existence of outliers. We consider a general class of elliptical distributions to model the heavy-tailed distribution of latent log-basis variables, which is characterized by a latent shape matrix. The latent shape matrix is a scalar multiple of the latent covariance matrix, when it exists, and it can preserve the directional properties of the dependence in a distribution when the covariance matrix does not exist. We propose using a robust composition-adjusted thresholding procedure based on Tyler's M-estimator to estimate the latent shape matrices of high-dimensional compositional data from different groups. We prove appealing theoretical properties under the high-dimensional setting. Simulation studies and a real application to microbial inter-taxa analysis demonstrate the numerical properties of the proposed method.

Key words and phrases: Compositional data, elliptical distribution, human microbiome research, shape matrix, thresholding, Tyler's M-estimation.

1. Introduction

Compositional data arise naturally in many research topics in biology, ecology, finance, geology, and other areas. For example, compositional data are used to assess the relative proportions of chemicals within stones across different geographical locations in geology (Thomas and Aitchison (2005)), and to analyze relative market share while dynamically accounting for the total market size in economics (Arata and Onozaki (2017)). This study is motivated by inter-taxa analyses of microbiome compositional data in the rapidly growing field of human microbiome research (Cho and Blaser (2012)). It is known that an accurate estimation of the dependence structure between bacteria leads to a better understanding of the underlying mechanisms of disease development

Corresponding author: Lingzhou Xue. E-mail: lzxue@psu.edu. Xiang Zhan. E-mail: zhanx@bjmu.edu.cn.