Statistica Sinica **36** (2026), 1-22
doi:https://doi.org/10.5705/ss.202024.0104

A SEMIPARAMETRIC QUANTILE SINGLE-INDEX MODEL FOR ZERO-INFLATED OUTCOMES

Zirui Wang and Tianying Wang*

Tsinghua University and Colorado State University

Abstract: We consider the complex data modeling problem motivated by the zero-inflated and overdispersed data from microbiome studies. Analyzing how microbiome abundance is associated with human biological features, such as BMI, is of great importance for host health. Methods based on parametric distributional assumptions, such as zero-inflated Poisson and zero-inflated Negative Binomial regression, have been widely used in modeling such data, yet the parametric assumptions are restricted and hard to verify in real-world applications. We relax the parametric assumptions and propose a semiparametric single-index quantile regression model. It is flexible to include a wide range of possible association functions and adaptable to the various zero proportions across subjects, which relaxes the strong parametric distributional assumptions of most existing zero-inflated data modeling approaches. We establish the asymptotic properties for the index coefficients estimator and quantile regression curve estimation. Through extensive simulation studies, we demonstrate the superior performance of the proposed method regarding model fitting.

Key words and phrases: Microbiome count data, profile principle, quantile regression, single-index model, zero-inflation.

1. Introduction

The human microbiota consists of the microorganisms that reside in or on the human body and contribute essential functions to human beings (Cani, 2018). Human microbiome research studies the dynamic interactions among microbiomes, host, and environment (Xia and Sun, 2017). It is of great importance to build more accurate predictive models of taxa and identify the relationship between taxa and clinical parameters (Lloyd-Price, Abu-Ali and Huttenhower, 2016). The main challenges in modeling microbiome data are zero inflation and overdispersion (Kaul et al., 2017). It is common that the proportion of zeros in gut microbiota counts can reach 70%–80% (Yatsunenko et al., 2012). Meanwhile, the non-zero counts of the microbiota counts could be as large as thousands and cause overdispersion (McMurdie and Holmes, 2014). The inflated zeros in microbiome data are commonly caused by two reasons: microbes are present in the environment but not detected due to low sequencing

*Corresponding author. E-mail: Tianying.Wang@colostate.edu