LMU

LUDWIG-
MAXIMILIANS-
UNIVERSITÄT
MÜNCHEN

Prof. Dr. Christian L. Müller
Ludwig-Maximilians-Universität München
Institut für Statistik
Ludwigstr. 33
80539 München
christian.mueller@stat.uni-muenchen.de

M.Sc. Thesis Proposal: *Bayesian modeling of high-throughput sequencing data*

Objective: The goal of this M.Sc. Thesis is to extend a Bayesian model for inferring cell composition changes in single-cell RNA sequencing data to be used with higher-dimensional data. The scCODA model (Büttner et al., 2020; https://www.biorxiv.org/content/10.1101/2020.12.14.422688v2) allows researchers to infer sparse credible changes in cell count data in low- to moderate-dimensional settings. The key objectives of the thesis are to adjust the hierarchical model structure of scCODA to account for special properties of higher-dimensional sequencing data, for example from microbiome analysis, and to optimize the Bayesian inference process.

Plan and deliverables: A successful completion of the M.Sc. thesis requires the following computational and scientific advances. The scCODA model was specifically designed for use in very low-dimensional settings, where some properties of high-throughput sequencing data like zero-inflation and overdispersion are less pronounced.

Therefore, the model should first be tested on large-scale observational microbiome data, e.g. the American Gut Project (https://msystems.asm.org/content/3/3/e00031-18), and adjusted for these types of data. Code is already available in Python (tensorflow) to solve the associated estimation problem via Hamiltonian Monte Carlo sampling, but will require enhancements and testing.

Secondly, we will look at how to perform inference on the model via Variational Inference methods or a Neural Network, and compare the performance of different methods.

We will finally use the model to formulate hypotheses about links between covariates and microbial abundance, and validate those via literature review. A successful outcome could be to link diseases or other factors to changes in abundance of a subset of species in the microbiome.

A write-up in thesis form and commented code on GitHub are mandatory deliverables at the end of the thesis.