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: *Pre-training and Sparse Log-Contrast Regression in Compositional Microbiome Studies*

Objective: The human gut microbiome is composed of highly diverse microbial species whose relative abundances vary significantly across individuals. These species interact in complex ways, and their combined metabolic functions can have critical impacts on human health, such as in diseases like irritable bowel syndrome (IBS). Given the compositional nature of many microbiome datasets (i.e., only the relative abundances of species are observed), standard regression models are often inappropriate. Instead, methods such as sparse log-contrast regression [1], which account for the compositional structure of the data, are better suited for predicting health outcomes like disease status. A novel aspect of this thesis is to combine pre-training [2] with sparse log-contrast regression. By allowing the model to learn general patterns from a large dataset and then specialize on the smaller task-specific data, pre-training can significantly improve performance over models trained solely on the smaller dataset. In particular, this thesis will apply this approach to the MetaIBS study [3], which aggregates 16S rRNA amplicon data from ∽2,500 IBS and healthy individuals across 13 different studies. MetaIBS addresses inconsistencies in prior microbiome research by standardizing the processing of raw microbiome data and enabling coherent taxonomic assignments across diverse experimental protocols and sample types. It provides an unprecedented scale of data to explore how microbial composition influences IBS.

Plan and Deliverables: A successful completion of the M.Sc. thesis requires the student to familiarize themselves with compositional data analysis, particularly in the context of microbiome data. The first task involves conducting an in-depth review of existing literature [1, 2, 3] to understand the current methodologies and results. The next step is to create a reproducible GitHub repository that connects, tests, and improves existing code and pipelines. The goal of this thesis is to assess whether integrating pre-training improves the predictive accuracy of IBS outcomes when applied to the MetaIBS data. Additionally, the model will explore whether to make joint predictions across the entire cohort or fine-tune to specific subgroups.

References

1. Combettes, P. L., & Müller, C. L. (2021). Regression models for compositional data: General log-contrast formulations, proximal optimization, and microbiome data applications. *Statistics in Biosciences*, *13*(2), 217-242.
2. Craig, E., Pilanci, M., Menestrel, T. L., Narasimhan, B., Rivas, M., Dehghannasiri, R., ... & Tibshirani, R. (2024). Pretraining and the Lasso. *arXiv preprint arXiv:2401.12911*.
3. Carcy, S., Ostner, J., Tran, V., Menden, M., Müller, C. L. (2024). MetaIBS - large-scale amplicon-based meta analysis of irritable bowel syndrome. *bioRxiv 2024.01.22.575775*.