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: Inference for high-dimensional microbiome regression analysis

Objective and background: Compositional data refers to datasets where observations represent proportions, such as the distribution of various microbial species across different locations or among patients. With modern sequencing machines, the number of microbial species exceeds the number of observations (e.g., patients / locations), leading to high-dimensional settings. Using compositional data to predict an outcome requires the application of specialized methods, such as the log-contrast model [1], a constrained regression model, or the penalized counterpart [2].

The primary objective of this thesis is to implement and benchmark the FDR-controlled variable selection procedure proposed in [3] for sparse log-contrast models and robust extensions thereof [4, 5]. This procedure aims to identify microbial predictors while controlling for false discoveries, enhancing the reliability of the results.

To accomplish this, the student will first employ a realistic simulation to evaluate the efficacy of the methods compared to stability-based variable selection techniques [6]. Followed by applying the developed method to two real-world microbiome datasets:

- Ocean microbiome: Utilize data from the Tara Ocean project [7, 8] to determine groups of microbes predictive of environmental factors
- Large-scale health-related microbiome data: Analyze data from the American Gut Project (AGP) [9] to discover reliable microbes predictive of health status

Plan and deliverables: A successful completion of the M.Sc. thesis requires the following computational and scientific advances:

1. The procedure should be implemented and tested in either Python or R.
2. Implement a reproducible workflow for the simulation study and application part.

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

References

[1] J. Aitchison and J. Bacon-Shone, “Log contrast models for experiments with mixtures,” *Biometrika*, vol. 71, no. 2, pp. 323–330, 1984.

[2] W. Lin, P. Shi, R. Feng, and H. Li, “Variable selection in regression with compositional covariates,” *Biometrika*, vol. 101, no. 4, pp. 785–797, Dec. 2014, doi: 10.1093/biomet/asu031.

[3] P. Yuan, C. Jin, and G. Li, “FDR control for linear log-contrast models with high-dimensional compositional covariates,” *Comput. Stat. Data Anal.*, vol. 197, p. 107973, Sep. 2024, doi: 10.1016/j.csda.2024.107973.

[4] L. Simpson, P. Combettes, and C. Müller, “c-lasso - a Python package for constrained sparse and robust regression and classification,” *J. Open Source Softw.*, vol. 6, no. 57, p. 2844, Jan. 2021, doi: 10.21105/joss.02844.

[5] P. L. Combettes and C. L. Müller, “Regression Models for Compositional Data: General Log-Contrast Formulations, Proximal Optimization, and Microbiome Data Applications,” *Stat. Biosci.*, vol. 13, no. 2, pp. 217–242, Jul. 2021, doi: 10.1007/s12561-020-09283-2.

[6] N. Meinshausen and P. Bühlmann, “Stability Selection,” *J. R. Stat. Soc. Ser. B Stat. Methodol.*, vol. 72, no. 4, pp. 417–473, Sep. 2010, doi: 10.1111/j.1467-9868.2010.00740.x.

[7] S. Sunagawa *et al.*, “Structure and function of the global ocean microbiome,” *Science*, vol. 348, no. 6237, p. 1261359, 2015.

[8] G. Salazar *et al.*, “Gene Expression Changes and Community Turnover Differentially Shape the Global Ocean Metatranscriptome,” *Cell*, vol. 179, no. 5, pp. 1068-1083.e21, Nov. 2019, doi: 10.1016/j.cell.2019.10.014.

[9] D. McDonald *et al.*, “American Gut: an Open Platform for Citizen Science Microbiome Research,” *mSystems*, vol. 3, no. 3, pp. e00031-18, Jun. 2018, doi: 10.1128/mSystems.00031-18.