UNIVERSITY OF CALIFORNIA SAN DIEGO

Statistical Methods for Longitudinal Data Analysis and Reproducible Feature Selection in Human Microbiome Studies

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Biostatistics

by

Lingjing Jiang

Committee in charge:

Professor Rob Knight, Co-Chair
Professor Wesley K. Thompson, Co-Chair
Professor Deborah M. Kado
Professor Loki Natarajan
Professor Xin Tu

2020

The dissertation of Lingjing Jiang is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____

Co-Chair

_____

Co-Chair

University of California San Diego

2020

DEDICATION

To my husband Paul: for your unconditional love, encouragement and tender care.

To my daughter Perisseia: for transforming my life.

To my sister Grace, and brother Isaac: for being my chefs, baby sitters and helpers.

To my Parents Laoba, and Laoma: for planting and watering.

To my spiritual Parents James, and Rebecca: for cherishing and nourishing.

To the Chungs, Loys, Sos, Tramels, Allisons, Palmas, Klimmeks and Barb: for being my extended families wherever I am.

EPIGRAPH

Statistics is a science, as well as an art. Like writing or painting, there are many ways of doing it. Be aware of your own style and what you are comfortable of doing.

— Andrew Gelman

But by the grace of God I am what I am; and His grace unto me did not turn out to be in vain, but, on the contrary, I labored more abundantly than all of them, yet not I but the grace of God which is with me.

— 1 Corinthians 15:10

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

As an African proverb says, “it takes a village to raise a child." This is not merely true to my parenting experience, but also incredibly applicable to my own PhD journey. I cannot imagine going through these exciting and challenging three years without the abundant support, academically and personally, from my advisors, colleagues, friends and families. I feel so privileged to grow in such a harmonious village of science and faith, and my heart is full of thanksgiving to all those who help and guide me to become a mature statistician, a curious scientist, a joyful mother, and a thankful believer.

I would like to first acknowledge my advisor Rob Knight, whose pioneering microbiome research has inspired so many statistical ideas in this dissertation, and also provided me rich experience in collaborating and leading numerous projects. Thanks for your insightful and forward-looking mentorship, and your trust and support in my being a female scientist and a PhD mother. I would also like to thank my co-advisor Wesley K. Thompson for affording me the freedom and flexibility to explore my personal research interest and dive into the uncomfortable zone.

If I could list three co-advisors in my dissertation, Loki Natarajan is undoubtedly one of them. Thanks for always making yourself available to address my all kinds of statistical questions and urgent requests. Thanks for your encouragement and standing behind me in the midst of difficulties. Thank you, Deborah M. Kado, for enlisting me on the MrOS studies to team up with Robert L. Thomas, Jian Shen and other international experts; thanks for your patience and appreciation while witnessing my growth as a microbiome researcher. Thank you, Xin Tu, for including me in your semi-parametric statistical project, which is a new learning experience to me.

My PhD journey could not be so fruitful without the strong support from the highly collaborative and intellectual members in the Knight Lab. Special thanks to Amnon Amir, and James T. Morton for guiding me into the new area of statistical method development for microbiome data analysis. Many thanks to multiple individuals that served as my mentors, namely Austin D. Swafford, Yoshiki Vázquez-Baeza, Antonio González, Shi Huang, Daniel McDonald, Mehrbod Estaki, Tomasz Kosciolek, Justine Debelius, Stefan Janssen, Qiyun Zhu, Sejin Song, Franck Lejzerowicz and Zhenjiang Xu. I would also like to thank my fellow peers who are always exciting of applying my new methods to their datasets, and approaching

me for interesting statistical questions, in particular Anupriya Tripathi, Cameron Martino, Clarisse A. Marotz, Celeste Allaband, Bryn C. Taylor, Alison Vrbanac and Robert H. Mills.

The Biostatistics PhD program is another pillar of my PhD training. It is such a joy to see the program develop and expand, while enjoying the intimacy with my fellow peers, and personal instructions from my knowledgeable professors. I want to express my gratitude to Armin Schwartzman, Karen Messer, Florin Vaida and Steven D. Edland, for perfecting me on statistical thinking and reasoning. I also want to thank my fellow PhD students, Brian Kwan, Anya Umlauf, Jinyuan Liu, Yuqi Qiu, Ruifeng Chen, Kristen Hansen and Wenyi Lin, for supporting one other and working together to produce the first group of Biostatistics PhD graduates.

Without the professional support to take care of the practical needs such as logistics, finances, and computing resource maintenance, I would not be able to complete my PhD studies. I want to thank my Biostatistics coordinators Melody Bazyar, Sarah Dauchez, and Stella Tripp, and Knight Lab professionals Jerry Kennedy, Gail Ackermann, Jeff DeReus, Michiko Souza, and Yna Villanueva.

Finally, I want to thank my friends and family for their tender love and care, in particular, my husband Paul, my daughter Perisseia, my siblings Grace and Isaac, my parents, my spiritual parents and my extended families in Hong Kong and the United States.

Chapter 2, in full, has been submitted for publication and is presented as it may appear in “Jiang, L.; Zhong, Y.; Elrod, C.; Natarajan, L.; Knight, R.; Thompson, W.K. *BayesTime: Bayesian Functional Principal Components for Sparse Longitudinal Data.* Journal of Computational and Graphical Statistics." The dissertation author was the primary investigator and author of this work.

Chapter 3, in full, has been submitted for publication and is presented as it may appear in “Jiang, L.; Elrod, C.; Swafford, A.D.; Knight, R.; Thompson, W.K. *Bayesian Multivariate Sparse Functional Principal Components Analysis with Applications to Longitudinal Microbiome Multi-Omics Data.* Annals of Applied Statistics." The dissertation author was the primary investigator and author of this work.

Chapter 4, in full, has been submitted for publication and is presented as it may appear in “Jiang, L.; Haiminen, N.; Carrieri, A.; Huang, S.; Vázquez-Baeza, Y.; Parida, L.; Kim, H.; Swafford, A.D.; Knight, R.; Natarajan, L. *Utilizing Stability Criteria in Choosing Feature Selection Methods Yields Reproducible Results in Microbiome Data.* Biometrics." The dissertation author was the primary investigator and author

of this work.

VITA

2010-2011	Undergraduate Visiting Fellow Williams College
2012	Bachelor of Arts in Translation The Chinese University of Hong Kong
2016	Master of Science in Statistics University of California San Diego
2016-2017	Statistician Shiley-Marcos Alzheimer's Disease Research Center
2017-2020	Graduate Research Fellow Knight Lab, University of California San Diego
2020	Doctor of Philosophy in Biostatistics University of California San Diego

PUBLICATIONS

Author names marked with † indicate shared first co-authorship.

Jiang, L., Zhong, Y., Elrod, C., Natarajan, L., Knight, R., Thompson, W.K. *BayesTime: Bayesian Functional Principal Components for Sparse Longitudinal Data.* Journal of Computational and Graphical Statistics. *Submitted*

Jiang, L., Haiminen, N., Carrieri, A., Huang, S., Vázquez-Baeza, Y., Parida, L., Kim, H., Swafford, A.D., Knight, R., Natarajan, L. *Utilizing Stability Criteria in Choosing Feature Selection Methods Yields Reproducible Results in Microbiome Data.* Biometrics. *Submitted*

Jiang, L., Elrod, C., Swafford, A.D., Knight, R., Thompson, W.K. *Bayesian Multivariate Sparse Functional Principal Components Analysis with Applications to Longitudinal Microbiome Multi-Omics Data.* Annals of Applied Statistics. *Submitted*

The following publications were not included as part of this dissertation, but were also significant works of my doctoral training.

Jiang, L.†, Amir, A.†, Morton, J.T., Heller, R., Arias-Castro, E. and Knight, R., 2017. *Discrete false-discovery rate improves identification of differentially abundant microbes.* MSystems, 2(6).

Vrbanac, A., Debelius, J.W., **Jiang, L.**, Morton, J.T., Dorrestein, P. and Knight, R., 2017. *An Elegan (t) Screen for Drug-Microbe Interactions.* Cell Host & Microbe, 21(5), pp.555-556.

Thompson, L.R., Sanders, J.G., McDonald, D., Amir, A., Ladau, J., Locey, K.J., Prill, R.J., Tripathi, A., Gibbons, S.M., Ackermann, G., Navas-Molina, J.A., **Jiang, L.**, et al., 2017. *A communal catalogue reveals Earth's multiscale microbial diversity.* Nature, 551(7681), pp.457-463.

Edland, S.D., Ard, M.C., Li, W. and **Jiang, L.**, 2017. *Design of pilot studies to inform the construction of composite outcome measures.* Alzheimer's and Dementia: Translational Research and Clinical Interventions, 3(2), pp.213-218.

Tripathi, A., Melnik, A.V., Xue, J., Poulsen, O., Meehan, M.J., Humphrey, G., **Jiang, L.**, Ackermann, G., McDonald, D., Zhou, D. and Knight, R., 2018. *Intermittent hypoxia and hypercapnia, a hallmark of obstructive sleep apnea, alters the gut microbiome and metabolome.* MSystems, 3(3).

Janssen, S., McDonald, D., Gonzalez, A., Navas-Molina, J.A., **Jiang, L.**, Xu, Z.Z., Winker, K., Kado, D.M., Orwoll, E., Manary, M. and Mirarab, S., 2018. *Phylogenetic placement of exact amplicon sequences improves associations with clinical information.* Msystems, 3(3).

McDonald, D., Hyde, E., Debelius, J.W., Morton, J.T., Gonzalez, A., Ackermann, G., Aksenov, A.A., Behsaz, B., Brennan, C., Chen, Y. and Goldasich, L.D., **Jiang, L.**, et al., 2018. *American Gut: an open platform for citizen science microbiome research.* Msystems, 3(3), pp.e00031-18.

Xu, Z.Z., Amir, A., Sanders, J., Zhu, Q., Morton, J.T., Bletz, M.C., Tripathi, A., Huang, S., McDonald, D., **Jiang, L.** and Knight, R., 2019. *Calour: an Interactive, Microbe-Centric Analysis Tool.* Msystems, 4(1).

Mills, R.H., Vázquez-Baeza, Y., Zhu, Q., **Jiang, L.**, Gaffney, J., Humphrey, G., Smarr, L., Knight, R. and Gonzalez, D.J., 2019. *Evaluating metagenomic prediction of the metaproteome in a 4.5-year study of a patient with Crohn's disease.* MSystems, 4(1).

Bolyen, E., Rideout, J.R., Dillon, M.R., Bokulich, N.A., Abnet, C.C., Al-Ghalith, G.A., Alexander, H., Alm, E.J., Arumugam, M., Asnicar, F., Bai, Y., **Jiang, L.**, et al., 2019. *Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2.* Nature biotechnology, 37(8), pp.852-857.

Gauglitz, J.M., Aceves, C.M., Aksenov, A.A., Aleti, G., Almaliti, J., Bouslimani, A., Brown, E.A., Campeau, A., Caraballo-Rodríguez, A.M., Chaar, R., da Silva, R.R., **Jiang, L.**, et al., 2020. *Untargeted mass spectrometry-based metabolomics approach unveils molecular changes in raw and processed foods and beverages.* Food chemistry, 302, p.125290.

Fields, F.J., Lejzerowicz, F., Schroeder, D., Ngoi, S.M., Tran, M., McDonald, D., **Jiang, L.**, Chang, J.T., Knight, R. and Mayfield, S., 2020. *Effects of the microalgae Chlamydomonas on gastrointestinal health.* Journal of Functional Foods, 65, p.103738.

Huang, S., Haiminen, N., Carrieri, A.P., Hu, R., **Jiang, L.**, Parida, L., Russell, B., Allaband, C., Zarrinpar, A., Vázquez-Baeza, Y. and Belda-Ferre, P., 2020. *Human Skin, Oral, and Gut Microbiomes Predict Chronological Age.* Msystems, 5(1).

Shardell, M., Parimi, N., Langsetmo, L., Tanaka, T., **Jiang, L.**, Orwoll, E., Shikany, J.M., Kado, D.M. and Cawthon, P.M., 2020. *Comparing Analytical Methods for the Gut Microbiome and Aging: Gut Microbial Communities and Body Weight in the Osteoporotic Fractures in Men (MrOS) Study.* The Journals of Gerontology: Series A.

Taylor, B.C., Lejzerowicz, F., Poirel, M., Shaffer, J.P., **Jiang, L.**, Aksenov, A., Litwin, N., Humphrey, G., Martino, C., Miller-Montgomery, S. and Dorrestein, P.C., 2020. *Consumption of Fermented Foods Is Associated with Systematic Differences in the Gut Microbiome and Metabolome.* Msystems, 5(2).

Casals-Pascual, C., González, A., Vázquez-Baeza, Y., Song, S.J., **Jiang, L.** and Knight, R., 2020. *Microbial Diversity in Clinical Microbiome Studies: Sample Size and Statistical Power Considerations.* Gastroenterology, 158(6), pp.1524-1528.

Raghuvanshi, R., Vasco, K., Vázquez-Baeza, **Jiang, L.**, L., Morton, J.T., Li, D., Gonzalez, A., Goldasich, L.D., Humphrey, G., Ackermann, G. and Swafford, A.D., 2020. *High-Resolution Longitudinal Dynamics of the Cystic Fibrosis Sputum Microbiome and Metabolome through Antibiotic Therapy.* Msystems, 5(3).

Estaki, M.†, **Jiang, L.**†, Bokulich, N.A., McDonald, D., González, A., Kosciolek, T., Martino, C., Zhu, Q., Birmingham, A., Vázquez-Baeza, Y. and Dillon, M.R., 2020. *QIIME 2 Enables Comprehensive End-to-End Analysis of Diverse Microbiome Data and Comparative Studies with Publicly Available Data.*, Current protocols in bioinformatics, 70(1), p.e100.

Martino, C., Shenhav, L., Marotz, C.A., Armstrong, G., McDonald, D., Vázquez-Baeza, Y., Morton, J.T., **Jiang, L.**, Dominguez-Bello, M.G., Swafford, A.D. and Halperin, E., 2020. *Context-aware Dimensionality Reduction Deconvolutes Gut Microbial Community Dynamics.* Nature Biotechnology, pp.1-4.

Huey, S.L., **Jiang, L.**, Fedarko, M.W., McDonald, D., Martino, C., Ali, F., Russell, D.G., Udipi, S.A., Thorat, A., Thakker, V. and Ghugre, P., 2020. *Nutrition and the Gut Microbiota in 10-to 18-Month-Old Children Living in Urban Slums of Mumbai, India.* mSphere, 5(5).

Thomas, R.L.†, **Jiang, L.**†, Xu, Z.Z., Shen, J., Janssen, S., Ackerman, G., Adams, J., Pauwels, S., Vanderscheuren, D. , Knight, R., Orwoll, E., Kado, D.M. *Vitamin D Metabolites and the Gut Microbiome in Older Men.*, Nature Communication, accepted

Vujkovic-Cvijin, I., Sklar, J., **Jiang, L.**, Natarajan, L., Knight, R., and Belkaid, Y., 2020. *Host Variables Confound Gut Microbiota Studies of Human disease.* Nature, pp.1-7.

Song, S.J., Wang, J., Martino, C., **Jiang, L.**, Thompson, W.K., Shenhav, L., McDonald, D., Marotz, C. Harris, P.R., Hernandez, C., Henderson, N., Ackley, E., Nardella, D., Gillihan, C., Montacuti, V., Schweizer, W., Jay, M., Combellick, J., Garcia-Mantrana, I., Raga, F.G., Collado, M.C., Rivera-Vinas, J.I., Campos-Rivera, M., Ruiz-Calderon, J.F.R., Knight, R., and Dominguez-Bellow, M.G. *Partial Restoration of the Microbiome Trajectory by Vaginal Seeding of C-section Born infants.* Nature Microbiology. *Submitted*

Allaband, C., Lingaraju, A., Russel, B.J., Tripathi, A., **Jiang, L.**, Poulsen, O. Haddad., G.G., Dorrestein, P.C., Knight, R., and Zarrinpar, A. *Sample Collection Time is Critical for Microbiome Result Reproducibility.* Nature Microbiology. *Submitted*

Xue, J., Allaband, C., Zhou, D., Poulsen, O., Martino, C., **Jiang, L.**, Tripathi, A., Elijah, E., Dorrestein, P., Knight, R., Zarrinpar, A. and Haddad, G.G. *Influence of Intermittent Hypoxia/Hypercapnia on Atherosclerosis, Gut Microbiome and Metabolome.* mSystems. *Submitted*

Orchanian, S.B., Gauglitz, J.M., Wandro, S., Weldon, K.C., Doty, M., Stillwell, K., Hansen, S., **Jiang, L.**, Vargas, F., Rhee, K.E., Lumeng, J.C., Dorrestein, P.C., Knight, R., Song, S.J., Kim, J.H., and Swafford, A.D., *Multiomic Analyses of Nascent Preterm Infant Microbiomes Differentiation Suggest Opportunities for Targeted Intervention.* Nature Medicine. *Submitted*

Liu, J., Zhang, X., Chen, T., Wu, T., Lin, T., **Jiang, L.**, Lang, S., Liu, L., Natarajan, L., Tu, J.X., Kiosciolek, T., Morton, J., Nguyen, T.T., Schnabl, B., Knight, R., Feng, C., and Tu, X.M. *A Semi-parametric Model for Between-Subject Attributes: Applications to Beta-diversity of Microbiome Data.* Biometrics. *Submitted*

Marotz, C., Belda-Ferre1, P., Ali, F., Das, P., Huang, S., Cantrel, K., **Jiang, L.**, Martino, C., Diner, R.E., Rahman, G., McDonald, D., Armstrong, G., Kodera, S., Donato, S., Ecklu-Mensah, G., Gottel, N., Garcia, M.S., Chiang, L., Benitez, R.S., Shaffer, J.P., Bryant, M., Sanders, K., Humphrey, G., Ackermann, G., Haiminen, N., Beck, K.L., Kim, H., Vázquez-Baeza, Y., Torriani, F.J., Knight, R., Gilbert, J., Sweeney, D.A., Allard, S.M., *Evaluating the Microbial Context of SARS-CoV-2 in Patients and the Hospital Built Environment.* Science Translational Medicine. *Submitted*

ABSTRACT OF THE DISSERTATION

Statistical Methods for Longitudinal Data Analysis and Reproducible Feature Selection in Human Microbiome Studies

by

Lingjing Jiang

Doctor of Philosophy in Biostatistics

University of California San Diego, 2020

Professor Rob Knight, Co-Chair
Professor Wesley K. Thompson, Co-Chair

The microbiome is inherently dynamic, driven by interactions among microbes, with the host, and with the environment. At any point in life, human microbiome can be dramatically altered, either transiently or long term, by diseases, medical interventions or even daily routines. Since the human microbiome is highly dynamic and personalized, longitudinal microbiome studies that sample human-associated microbial communities repeatedly over time provide valuable information for researchers to observe both inter- and intra-individual variability, or to measure changes in response to an intervention in real time. Despite this increasing need in longitudinal data analysis, statistical methods for analyzing sparse longitudinal microbiome data and longitudinal multi-omics data still lag behind. In this dissertation, we describe our

efforts in developing two novel statistical methods, Bayesian functional principal components analysis (SFPCA) for sparse longitudinal data analysis, and multivariate sparse functional principal components analysis (mSFPCA) for longitudinal microbiome multi-omics data analysis.

Beyond longitudinal data analysis, we are also interested in utilizing statistical techniques for addressing the "reproducibility crisis" in microbiome research, especially in the indispensable task of feature selection. Instead of developing "the best" feature selection method, we focus on discovering a reproducible criterion called Stability for evaluating feature selection methods in order to yield reproducible results in microbiome analysis.

To set an appropriate motivation and context for our work, Chapter 1 reviews the importance of longitudinal studies in human microbiome research, and presents the crucial need of developing novel statistical methods to meet the new challenges in longitudinal microbiome data analysis, and of producing reproducible results in microbiome feature selection. Chapter 2 introduces Bayesian SFPCA, a flexible Bayesian approach to SFPCA that enables efficient model selection and graphical model diagnostics for valid longitudinal microbiome applications. Chapter 3 presents mSFPCA, an extension of Bayesian SFPCA from modeling a univariate temporal outcome to simultaneously characterizing multiple temporal measurements, and inferring their temporal associations based on mutual information estimation. Chapter 4 proposes to use reproducibility criterion such as Stability instead of popular model prediction metric such as mean squared error (MSE) to quantify the reproducibility of identified microbial features.

Chapter 1

Introduction

For centuries, a strong causal link has been drawn between bacteria and illness. But just over the past two decades, scientist have begun to shift that paradigm with mounting evidence that the dynamic community of trillions of microbes live in harmony with our human body, and they are crucial to our survival, influencing almost every aspect of our health from birth to death. We have only recently started to appreciate that the human body is home to at least 38 trillion microbial cells, outnumbering the 30 trillion human cells (Sender et al., 2016). Collectively, the microbial associates that reside in and on the human body constitute our microbiota, and the genes they encode are known as our microbiome. This complex community contains taxa from across the tree of life including bacteria, eukaryotes, viruses, and archaea, that interact with one another and with the host, greatly impacting the human health and physiology (Clemente et al., 2012). Only a small minority of these can be cultured, however, with the recent development of next generation sequencing techniques, scientists can now examine the "uncultured majority" of microbes by direct DNA sequencing, which greatly expanded the repertoire of known microbes both in our bodies and in the environment (Shendure and Ji, 2008; Whitman et al., 1998). The human microbiome contains about 150 times more genes than the human genome (Qin et al., 2010). In contrast to the mostly stationary human genome, the human microbiome is highly variable. It displays substantial intra- and inter- individual variation at different body sites (Huttenhower et al., 2012), and during different stages of human life (Dominguez-Bello et al., 2019).

The microbiota play a major role in human health from infancy to old age. At birth, infants are exposed to the maternal birth canal microbial population that influences the development of their gut microbiome. Infants born vaginally acquire their own mother's vaginal bacteria, while infants delivered through C-section harbor a characteristic of skin microbiota (Dominguez-Bello et al., 2010). Compared to vaginally delivered infants, the lack of the natural first inoculum in C-section babies might account for their increased susceptibility to certain pathogens, such as atopic diseases, allergies and asthma (Dominguez-Bello et al., 2011). Within a few hours of birth, breast milk is an early source of bacteria and nutrition introduced to the infant gut. In primarily breastfed infants, bacteria from mother's milk and areolar skin are most prominent in their infants' guts in the first month of life, accounting for nearly 40% of the gut bacteria (Pannaraj et al., 2017). In contrast, mother to infant microbe transmission was compromised in infants who were not primarily breastfed. This impact is not limited to the early childhood; breast milk bacteria seed the gut first influence and select for bacteria that follow, leaving a footprint that can be detected even in adulthood (Ding and Schloss, 2014). The infant microbial composition begins to converge toward an adult-like microbiota by the end of the first year (Palmer et al., 2007) and fully resembles the adult microbiota by 2.5 years of age (Koenig et al., 2011; Yatsunenko et al., 2012). Once the microbiota have reached maturity, they remain mostly stable until old age. However, the composition can be altered due to external perturbations such as antibiotic use, diet, and excessive hygiene. Repetitive use of antibiotics in humans is linked with an increase in antibiotic-resistant pathogens (Sommer et al., 2009). Although the particular bacterial taxa affected vary among individuals, some taxa do not recover even months after treatment, and in general, there is a long term decrease in bacterial diversity (Dethlefsen and Relman, 2011). Changing diet can alter the gut microbiota within days. For example, increased abundance of known butyrate producing bacteria in African Americans consuming a rural African diet caused butyrate production to increase 2.5 times and reduced synthesis of secondary bile acid (O'Keefe et al., 2015). The hygiene hypothesis postulates that excessive sanitation results in the lack of exposure to pathogenic and nonpathogenic microbial products in Western countries that might be a contributing factor to the inappropriate immune response to allergens (Clemente et al., 2012).

Given the rapid response of the microbiota to external perturbations throughout the human life, it will be of great value to model these dynamic changes and associate them with important biological and/or

clinical outcomes. Moreover, incorporating additional omics data into such analyses will be extremely helpful to understand the functional links between microbial communities and diseases. Last but not least, quantifying the reproducibility of identified bacterial taxa in microbial analyses will be crucial to ensure validity and future generalizability. In this dissertation, we presented three novel methods to address these critical needs.

1.1 Overview of Chapter Contents

In Chapter 2, we develop Bayesian Sparse Functional Principal Components Analysis (SFPCA), which is able to model the highly non-linear temporal trajectories, even when the longitudinal data are sparse with measurements for individuals occurring at possibly differing time points. Our Bayesian approach allows users to use the efficient leave-one-out cross-validation (LOO) with Pareto-smoothed importance sampling (PSIS) for model selection, and to utilise the estimated shape parameter from PSIS-LOO and also visual posterior predictive checks for graphical model diagnostics. This Bayesian implementation thus enables careful application of SFPCA to a variety of longitudinal microbiome data applications.

In Chapter 3, we extend Bayesian SFPCA from modeling a univariate temporal outcome to simultaneously characterizing multiple temporal measurements. Moreover, we utilized the correlations among the functional principal component (FPC) scores to estimate the inter-block and conditional inter-block relationship given other variables in order to infer the temporal associations between different measurements. This method could meet the growing need of simultaneously modeling multiple temporal outcomes and inferring their temporal associations in longitudinal microbiome multi-omics data.

In Chapter 4, we propose to use Stability index to quantify the reproducibility of identified microbial features. We showed that in both extensive simulations and real data applications, reproducibility criterion Stability is preferred over popular model prediction metric mean squared error (MSE). We thus suggest microbiome researchers use a reproducibility criterion such as Stability instead of a model prediction performance metric such as MSE for feature selection in microbiome data analysis.

Chapter 2

BayesTime: Bayesian Functional Principal Components for Sparse Longitudinal Data

2.1 Abstract

Modeling non-linear temporal trajectories is of fundamental interest in many application areas, such as in longitudinal microbiome analysis. Many existing methods focus on estimating mean trajectories, but it is also often of value to assess temporal patterns of individual subjects. Sparse principal components analysis (SFPCA) serves as a useful tool for assessing individual variation in non-linear trajectories; however its application to real data often requires careful model selection criteria and diagnostic tools. Here, we propose a Bayesian approach to SFPCA, which allows users to use the efficient leave-one-out cross-validation (LOO) with Pareto-smoothed importance sampling (PSIS) for model selection, and to utilize the estimated shape parameter from PSIS-LOO and also the posterior predictive checks for graphical model diagnostics. This Bayesian implementation thus enables careful application of SFPCA to a wide range of longitudinal data applications.

2.2 Introduction

Longitudinal data, i.e., multiple observations collected on the same subject over time, are ubiquitous in biomedical research. In addition to using longitudinal data to estimate mean trajectories, it is often of great interest to characterize individual subject variation. Both the mean trajectory and individual subject deviations from the mean trajectory may be highly non-linear and hard to characterize using typical modeling approaches for longitudinal data such as linear mixed-effects models. Additionally, longitudinal data are often collected at irregular timing and frequency across subjects (they are “sparse"), and methods for estimating trajectories need to be able to handle this common scenario.

For example, a question of fundamental interest in microbiome research is how the microbiome evolves in individual subjects as a response to subject-level perturbations, such as disease, diet and lifestyle (Kostic et al., 2015; Halfvarson et al., 2017; Smarr et al., 2017; Weingarden et al., 2015; David et al., 2014; Turnbaugh et al., 2009; Fierer et al., 2008). Accurate continuous monitoring of a subject’s microbiome may substantially improve prevention and treatment of some disorders. However, high-density temporal sampling is not currently feasible for microbiome studies; in practice, microbiome samples are collected infrequently and irregularly across time and subjects. Moreover, next generation sequencing techniques used to obtain estimates of microbial measurements are noisy, thus further hindering inference regarding the temporal evolution of a given subject’s microbial status. Finally, the microbiome exhibits highly nonlinear dynamics over time, which introduces an additional complication to traditional longitudinal analysis methods. While several analytical methods have been developed to model microbiome temporal dynamics addressing these challenges (Ridenhour et al., 2017; Gibson and Gerber, 2018; Silverman et al., 2018; Shenhav et al., 2019; Silverman et al., 2019), by and large the focus has been on mean trajectories, substantially ignoring potentially important information about variation in trajectories across subjects. Since microbiome progression is highly idiosyncratic, it would be of great interest to capture relevant individual deviation from the mean trajectories, perhaps resulting in personalized predictions and clustering of subjects based on progression patterns.

Sparse functional principal components analysis (SFPCA) serves as a useful tool to estimate smooth mean trajectories while at the same time estimating smooth principal modes of variation of

subject-level trajectories around the mean trajectory. SFPCA can be framed as an extension of linear random-effects models, where time effects are treated as random and non-linearity is achieved by choice of the functional basis (James et al., 2000; Kidziński and Hastie, 2018). The covariance structure of the trajectories is modeled as a low-rank matrix to produce efficient estimates of individual trajectories. Various fitting approaches, such as the EM algorithm, kernel smoothing and Newton-Raphson algorithm, have been proposed to estimate parameters of the SFPCA model (James et al., 2000; Yao et al., 2005; Peng and Paul, 2009). These approaches then use model selection techniques, such as cross-validation, Akaike information criterion (AIC) and leave-one-curve-out cross-validation, to select the dimension of basis and the number of principal components. However, due to their reliance on assumptions such as normally-distributed component scores and residuals, these models need to be carefully examined when applied to real data (Kidziński and Hastie, 2018).

We implemented the SFPCA model in a Bayesian framework to provide a flexible modeling approach that incorporates effective model selection and graphical diagnostic methods. Our BayesTime R package implementing the Bayesian SFPCA model allows users to use leave-one-out cross-validation (LOO) with Pareto-smoothed importance sampling (PSIS) for model selection (Vehtari et al., 2017), and to utilise the estimated shape parameter from PSIS-LOO and graphical posterior predictive checks for model diagnostics (Gelman et al., 1996; Gabry et al., 2019). This Bayesian implementation thus offers a flexible and comprehensive solution to real-date SFPCA applications, such as longitudinal microbiome data.

The Bayesian framework of SFPCA with PSIS-LOO is described in Section 2, and is implemented in the BayesTime package in R (Section 3). Section 4 presents Monte Carlo simulations evaluating the Baysian SFPCA model performance and further illustrates its use on a real longitudinal microbiome dataset, showing how individual microbiome trends can be visualized and explored with BayesTime. Future work is discussed in Section 5.

2.3 Methods

2.3.1 Sparse Functional Principal Components Analysis

The classical assumption of functional data analysis is that each trajectory is sampled over a dense grid of time points common to all individuals (Ramsay and Silverman, 2007). However, in practice, trajectories are often measured at an irregular and sparse set of time points that can differ widely across individuals. To address this scenario, James et al. (2000) proposed *sparse functional principal components analysis* (SFPCA) using a reduced rank mixed-effects framework. Let $Y_i(t)$ be the measurement at time t for the ith individual, $\mu(t)$ the overall mean function, f_j the jth principal component function and $f = [(f_1, f_2, \ldots, f_k)]^T$, where k is the number of principal components. Then the James et al. (2000) SFPCA model is given by

$$Y_i(t) = \mu(t) + \sum_{j=1}^{k} f_j(t)\alpha_{ij} + \varepsilon_i(t), \quad i = 1, ..., N \tag{2.1}$$

subject to the orthogonality constraint $\int f_j f_l = \delta_{jl}$, the Kronecker δ. The vector $\alpha_i = (\alpha_{i1}, \ldots, \alpha_{ik})^T$ is the component weights for the ith individual and $\varepsilon_i(t)$ is a normally-distributed residual, independent across subjects and across times within subject. The functions μ and f are approximated using cubic splines to allow for a smooth but flexible fit. Let $b(t)$ be a cubic spline basis with dimension $q > k$. The spline basis is orthonormalized so that $\int b_j b_l = \delta_{jl}$. Let Θ and θ_μ be, respectively, a $q \times k$ matrix and a q-dimensional vector of real-valued coefficients. For each individual i, denote their measurement times by $t = (t_{i1}, t_{i2}, \ldots, t_{in_i})^T$, and let $Y_i = (Y_i(t_{i1}), \ldots, Y_i(t_{in_i}))^T$ be the corresponding real-valued observations. Then $B_i = (b(t_{i1}), \ldots, b(t_{in_i}))^T$ is the $n_i \times q$ spline basis matrix for the ith individual. The reduced rank model can then be written as

$$\begin{aligned} Y_i &= B_i\theta_\mu + B_i\Theta\alpha_i + \varepsilon_i, \quad i = 1, ..., N, \\ \Theta^T\Theta &= I, \quad \alpha_i \sim N(0, D), \quad \varepsilon_i \sim N(0, \sigma_\varepsilon^2 I_{n_i}), \end{aligned} \tag{2.2}$$

where the covariance matrix D is restricted to be diagonal and I_{n_i} is the $n_i \times n_i$ identity matrix.

2.3.2 Bayesian SFPCA

We developed the SFPCA model in a Bayesian framework to allow for flexible prior specification and implementation of model selection and assessment methods. We implemented this Bayesian model using Hamilton Markov Chain Monte Carlo (MCMC) sampling algorithm in Stan (Carpenter et al., 2017). The real-valued observations $Y_i(t)$ are first standardized to have mean zero and standard deviation one. The prior distributions for parameters in Eq. (2.2) were chosen as follows:

$$\begin{aligned}
\theta_\mu &\sim N_q(0, I_q) \\
\alpha_i &\sim N_k(0, I_k) \\
\Theta_j &\sim N_q(0, I_q), j = 1, \ldots, k \\
\varepsilon_i &\sim N_{v_i}(0, \sigma_\varepsilon^2 I_{v_i}) \\
\sigma_\varepsilon &\sim Cauchy(0, 1),
\end{aligned}$$

where Θ_j is the jth column of the loading matrix Θ, and v_i is the total number of visits for the ith subject. The Bayesian implementation also enables use of leave-one-out cross-validation with Pareto-smoothed important sampling (PSIS-LOO) (Vehtari et al., 2017) to perform model selection on the number of principal components k and the number of basis functions q. Moreover, model fit can be assessed via diagnostics plots from PSIS-LOO as well as the graphical posterior predictive checks obtained from simulating posterior predictive data (Gelman et al., 1996; Gabry et al., 2019).

One difficulty in implementing the Bayesian SFPCA model is that the principal component loadings Θ are not uniquely specified. For a given $k \times k$ rotation matrix P, if $\Theta^* = \Theta P$ and Θ obeys the constraints in Eq.(3.2), then $\Theta^{*T}\Theta^* = P^T\Theta^T\Theta P = I$, and hence Θ is unidentifiable without additional restrictions. Instead of directly enforcing orthonormality when sampling from the conditional posteriors in the Bayesian model fitting, we sampled the parameters with no constraint on Θ and then performed a *post hoc* rotation for each iteration of the MCMC algorithm to meet the orthonormality constraint. Since the symmetric matrix $\Theta^T\Theta$ is identifiable and non-negative definite, we applied an eigenvalue decomposition $\Theta^T\Theta = VSV^T$, where V is the $q \times q$ matrix of orthonormal eigenvectors, and S is the diagonal matrix of

eigenvalues, with the q positive eigenvalues ordered from largest to smallest. Let $\Theta^* = V_k$ denote the $q \times k$ matrix consisting of the first k eigenvectors of V, which satisfies $\Theta^{*T}\Theta^* = I$. Finally, we rotated the FPC scores α_i to obtain $\alpha_i^* = \Theta^{*T}\Theta\alpha_i$, so that $\Theta^*\alpha_i^* = \Theta\alpha_i$.

2.3.3 Model Selection with PSIS-LOO

Leave-one-out cross-validation(LOO) with Pareto smoothed importance sampling (PSIS) is a stable model selection procedure which has been shown to be more robust in the presence of influential observations than other widely used criteria such as AIC (Akaike information criterion), DIC (deviance information criterion) and WAIC (widely applicable information criterion) (Vehtari et al., 2017). In Bayesian leave-one-out cross-validation, the estimate of the out-of-sample predictive fit (expected log pointwise predictive density) is defined as

$$elppd_{loo} = \sum_{i=1}^{n} \log p(y_i|y_{-i}), \tag{2.3}$$

where $p(y_i|y_{-i}) = \int p(y_i|\theta)p(\theta|y_{-i})d\theta$ is the leave-one-out predictive density given the data without the ith data point. Typically, LOO-CV requires re-fitting the model n times; however, a computational shortcut exists to enable only one model evaluation. As noted by Gelfand et al. (1992), if the n points are conditionally independent in the data model, we can then evaluate $p(y_i|y_{-i})$ with draw θ^s from the full posterior $p(\theta|y)$ using importance ratios, defined as

$$r_i^s = \frac{1}{p(y_i|\theta^s)} \propto \frac{p(\theta^s|y_{-i})}{p(\theta^s|y)}. \tag{2.4}$$

However, the posterior $p(\theta|y)$ is likely to have a smaller variance and thinner tails than the leave-one-out distribution $p(\theta|y_{-i})$, and thus a direct use of the formula above induces instability, because the importance ratios can have high or infinite variance.

Vehtari et al. (2015) improve the LOO estimate using Pareto smoothed importance sampling (PSIS), which applies a smoothing procedure to the importance weights. As the distribution of the importance weights used in LOO may have a long right tail, the empirical Bayes estimate of Zhang and Stephens (2009) can be used to fit a generalized Pareto distribution to the tail (e.g. 20% largest importance

rations), and this is done separately for each held-out data point i. So for each i, the result is a vector of weights $\widetilde{w}_i^s = F^{-1}\left(\frac{z-\frac{1}{2}}{M}\right), z = 1,\ldots,M$, where M is the number of simulation draws used to fit the Pareto distribution (in this case, $M = 0.2S$), and F^{-1} is the inverse-CDF of the generalized Pareto distribution. Then each vector of weights is truncated at $S^{3/4}\bar{w}_i$, denoted as w_i^s. These results can then be combined to compute the PSIS estimate of the LOO expected log pointwise predictive density:

$$\widehat{elppd_{psis-loo}} = \sum_{i=1}^{n} \log \frac{\sum_{s=1}^{S} w_i^s p(y_i|\theta^s)}{\sum_{s=1}^{S} w_i^s}. \tag{2.5}$$

2.3.4 Model Diagnostics

PSIS-LOO is not only efficient, it can also provide useful diagnostics for model checking. The estimated shape parameter $\hat{k}$ of the fitted Pareto distribution can be used to assess the reliability of the estimate; this diagnostic approach can be used routinely with PSIS-LOO for any model with a factorizable likelihood. If $k < \frac{1}{2}$, the variance of the raw importance ratios is finite, the central limit theorem holds, and the estimate converges quickly. If $k \in [\frac{1}{2}, 1]$, the variance of the raw importance ratios is infinite but the mean exists, the generalized central limit theorem for stable distributions holds, and the convergence of the estimate is slower. If $k > 1$, the variance and the mean of the raw ratios distribution do not exist. Vehtari et al. (2017) suggested that if the estimated tail shape parameter $\hat{k}$ exceeds 0.5, the user should be warned, although in practice they have observed good performance of values of $\hat{k}$ up to 0.7. Hence, this threshold of 0.7 could be used in practice for model diagnostics. If the ith LOO predictive distribution has a large $\hat{k}$ value when holding out data point i to evaluate predictive density, it suggests that data point i is a highly influential observation that deserves further examination.

Moreover, since we are implementing a Bayesian SFPCA model, we can also compare the observed data to simulated data from the posterior predictive distribution (Gabry et al., 2019). The idea behind posterior predictive checks is simple: if a model is a good fit, then it should be able to generate data that resemble the observed data. The data used for posterior predictive checks are simulated from the posterior predictive distribution $p(\tilde{y}|y) = \int p(\tilde{y}|\theta)p(\theta|y)d\theta$, where y is the current observed data, $\tilde{y}$ is the new data to be predicted, and θ are model parameters. By comparing the observed and replicated data in the posterior predictive checks, we may find a need to extend or modify the model.

2.4 Implementation

The Bayesian SFPCA method has been implemented in R in the BayesTime package at *github.com/biocore/bayestime*. The user can choose a range of the number of principal components and the dimension of cubic spline basis (i.e. the number of internal knots + 4) by the PC_range and nknot_range argument in stan_fit function, with which the knots are placed by the quantile of the time range in the default setting. The following model comparisons are performed with the optimal_model function, which compare models based on their $\widehat{elppd}_{psis-loo}$ and standard errors. Moreover, model diagnostics on the chosen model can be visualized using plot_k_diagnostic and plot_posterior_diagnostic fuctions.

```
library(BayesTime)
# use PSIS-LOO for model selection
sfpca_stan_results <- stan_fit(sfpca_data = dat, Nsamples = 1000, Nchain = 3, Ncores
    =3, PC_range = c(1,2,3), nknot_range = c(1,2))
optimal_model_idx <- optimal(model_list = sfpca_stan_results)
optimal_model <- sfpca_stan_results[[optimal_model_idx]]
# model diagnostics with Pareto shape parameter k
plot_k_diagnostic(dat, optimal_model)
# model diagnostics with posterior predictive checking
plot_posterior_diagnostic(dat, optimal_model)
```

2.5 Examples

Using the BayesTime package, we evaluated the performance of the Bayesian SFPCA model in Monte Carlo simulation studies and applied it to a longitudinal microbiome dataset to demonstrate its utility in a practical example. Data and code for simulations and real data application are available at *https://github.com/knightlab-analyses/BayesTime-analyses*.

2.5.1 Simulation Studies

Due to potential sequencing errors and sample collection procedures, missing data and dropouts are the norm rather than the exception in longitudinal microbiome studies. Moreover, despite large-scale

cross-sectional microbiome studies such as the Human Microbiome Project (Turnbaugh et al., 2007; Methé et al., 2012) and American Gut Project (McDonald et al., 2018), studies characterizing human-associated microbial communities over time often have relatively small sample sizes (Dethlefsen and Relman, 2011; Flores et al., 2014; Caporaso et al., 2011). Hence, it is important to assess the performance of Bayesian SFPCA in simulations with various sample sizes and with different levels of sparsity. In our simulations, we varied the total number of subjects at 100, 50, 25, 10, and the proportion of missing values at 0%, 20%, 50% and 80% (i.e., the percentage of randomly deleted observations to create increasingly sparse functional datasets) over observations at 10 time points. To better mimic the reality, we simulated longitudinal trajectories based on an SFPCA model using parameters initially estimated from the real microbiome data in the following way:

1. applying SFPCA to a real longitudinal microbiome dataset (Dominguez-Bello et al., 2016);
2. selecting the optimal number of PCs k and dimension of basis q using PSIS-LOO;
3. extracting the estimated values for population mean curve (θ_μ), FPC loadings (Θ), diagonal covariance matrix of FPC scores (D), and error variance (σ^2).

Then we simulate the data by varying the number of subjects, the number of time points and proportion of missing data as follows:

1. choosing the total number of subjects (N) and of time points (N_T) in order to place possible time points between $[0, 1]$;
2. specifying the average number (μ_T) of time points across all subjects in order to vary the proportion of missing data (approximated as $1 - \mu_T/N_T$) by simulating the observed number of time points for each individual with $n_i \sim Poisson(\mu_T)$ and then randomly placing the observed time points in the possible time locations (chosen in the previous step);
3. generating the cubic spline basis matrix $b(t)$ for each subject (orthonormality obtained through Gram-Schmidt orthonormalization);
4. simulating for each subject FPC scores $\alpha_i \sim N(0, D)$ and noise $\varepsilon_i \sim N(0, \sigma^2 I)$;

5. obtaining the temporal trajectory for each individual with $Y_i(t) = B_i\theta_\mu + B_i\Theta\alpha_i + \varepsilon_i$, where $B_i = (b(t_{i1}), \ldots, b(t_{in_i}))^T$

6. repeating steps 1 – 5 100 times for each simulation scenario with different number of subjects and proportion of missing data, thus generating 1600 simulated datasets in total.

Before describing the simulation results, we want to use the scenario of 100 subjects with 80% missing data to demonstrate how to perform model selection with PSIS-LOO and how to use its estimated shape parameter $\hat{k}$ to assess the reliability of the model. Models are compared based on their values of $\widehat{elppd_{psis-loo}}$: the larger the value, the better the model is. Among nine models tested (with the number of PCs and the number of internal knots ranging from 1 to 3), the model with two PC's and one internal knot had the highest $\widehat{elppd_{psis-loo}}$. The second best model (with three PC's and one internal knot) is lower in $\widehat{elppd_{psis-loo}}$ by 1.86, and the standard error of the difference between the two models is 2.21, indicating that the second model provides a similarly good fit. But since the first model is more parsimonious and all of its estimated shape parameters $\hat{k}$ are smaller than 0.7 (Figure 2.1A), we chose this as our best model.

We also generated graphical displays comparing observed data to simulated data from the posterior predictive distribution. In Figure 2.1B, the dark line is the distribution of the observed outcomes y and each of the lighter lines is the kernel density estimate of one of the replicates of y from the posterior predictive distribution. This figure shows that there is very little discrepancy between real and simulated data from the model, confirming the model validity for this application.

To evaluate the performance of Bayesian SFPCA, we investigated how well it recovered the mean trajectory and two PC functions. With 100 subjects, even as the proportion of missing data increased from 0% to 80%, the estimated overall mean curves and PC curves accurately recovered the ground truth in both scenarios (Figure 2.2). For the scenarios with 50 or 25 subjects with 80% missing data, the estimated mean curves were still close to the ground truth, except for a slight deviation at the two ends due to the large proportion of missing data there (Figure 2.3A, B). The PC curves were estimated well for both cases on two PCs, despite slight underestimation toward the end on both PCs (Figure 2.3C, D). As for the challenging scenarios of 10 samples with 50% or 80% missing data, the estimated mean curves in both scenarios and the PC curves for the scenario with 50% missingness were still robust (Figure 2.4A, B, C). However, for the

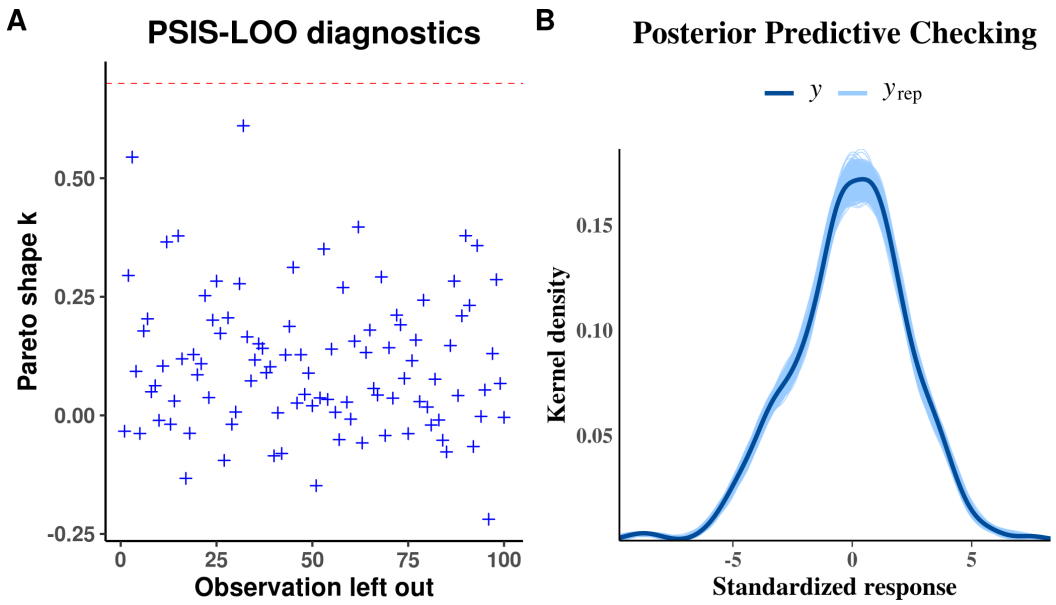

Figure 2.1: Graphical model checking with PSIS-LOO diagnostic plot and posterior predictive checks for Bayesian SFPCA simulated scenario of 100 subjects with 80% missing data.

case with 80% missing data, the estimated PC 1 curve did not capture the decreasing trend as accurately as before and displayed an artificial curvature towards the end; the estimated PC 2 curve also exhibited some deviations from the ground truth (Figure 2.4D). A closer look at the simulated trajectories (Figure 2.4B) indicated that few trajectories exhibited the decreasing trend at the beginning in both PCs due to the loss of data, hence the deviated estimation was caused by the limitation of the underlying data. Note that the visual comparisons above were demonstrated using one representative case from each simulation scenario. Results over all 100 simulated datasets for each scenario were summarized in table 2.1 and 2.2, showing that in the estimations of both mean (θ_μ) and FPC spline coefficients (Θ), mean squared errors increase as sample size decreases at each given missing proportion, although the variabilities are still within the 95% credible intervals. Moreover, errors for the mean and FPC estimations remain similar despite increasing missing proportion at fixed sample size. In summary, the performance of Bayesian SFPCA is robust to limited sample size and a high proportion of missing data.

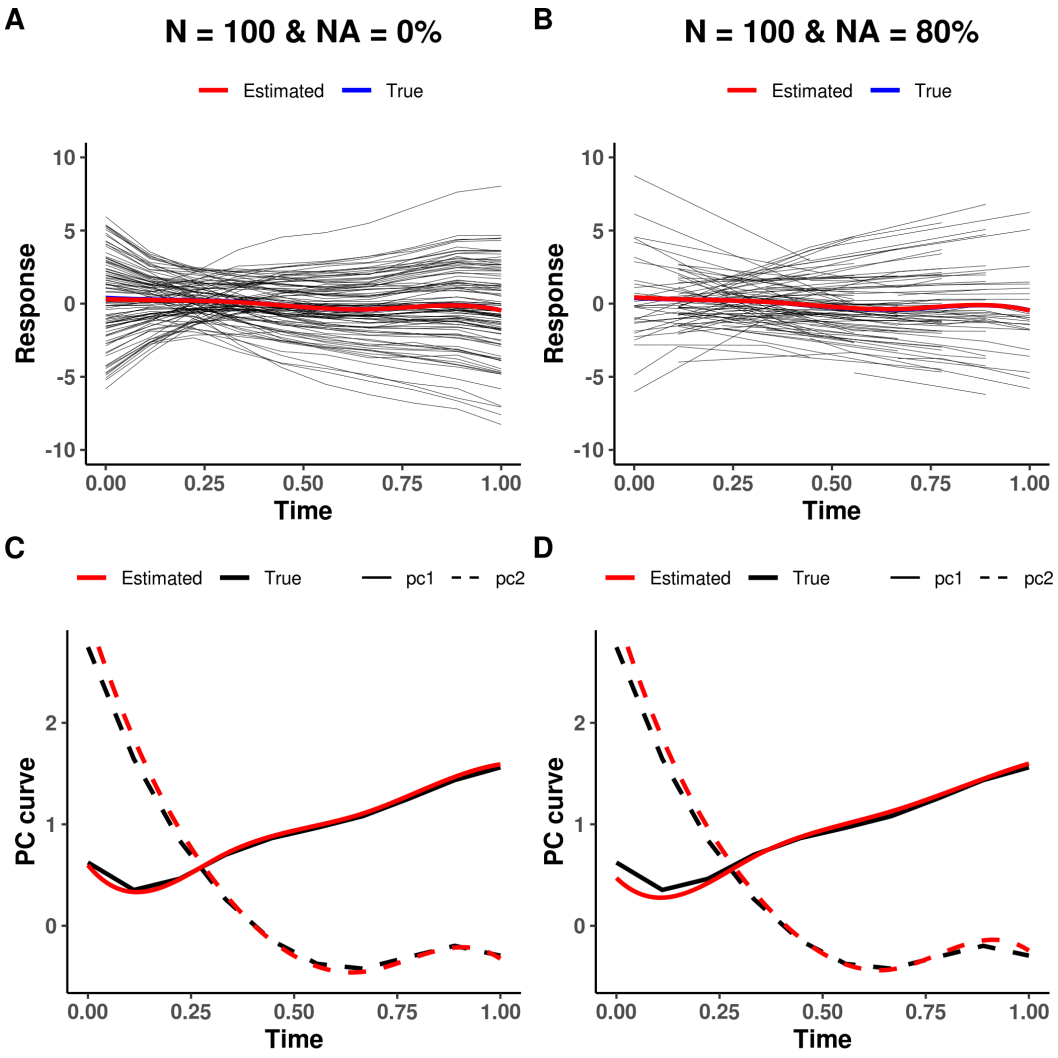

Figure 2.2: Results of Bayesian SFPCA on simulated data with 100 total samples of 0% vs. 80% missing data.

Table 2.1: Average mean squared errors with 95% CIs for estimating mean spline coefficients.

	N = 100	N = 50	N = 25	N = 20
NA = 0%	0.008 (0, 0.029)	0.015 (0, 0.066)	0.031 (0, 0.173)	0.044 (0.002, 0.188)
NA = 20%	0.008 (0, 0.034)	0.01 (0, 0.052)	0.021 (0.001, 0.088)	0.04 (0.001, 0.138)
NA = 50%	0.008 (0, 0.034)	0.011 (0, 0.053)	0.021 (0.001, 0.082)	0.041 (0.001, 0.154)
NA = 80%	0.007 (0, 0.025)	0.013 (0.001, 0.08)	0.026 (0.001, 0.105)	0.069 (0.009, 0.223)

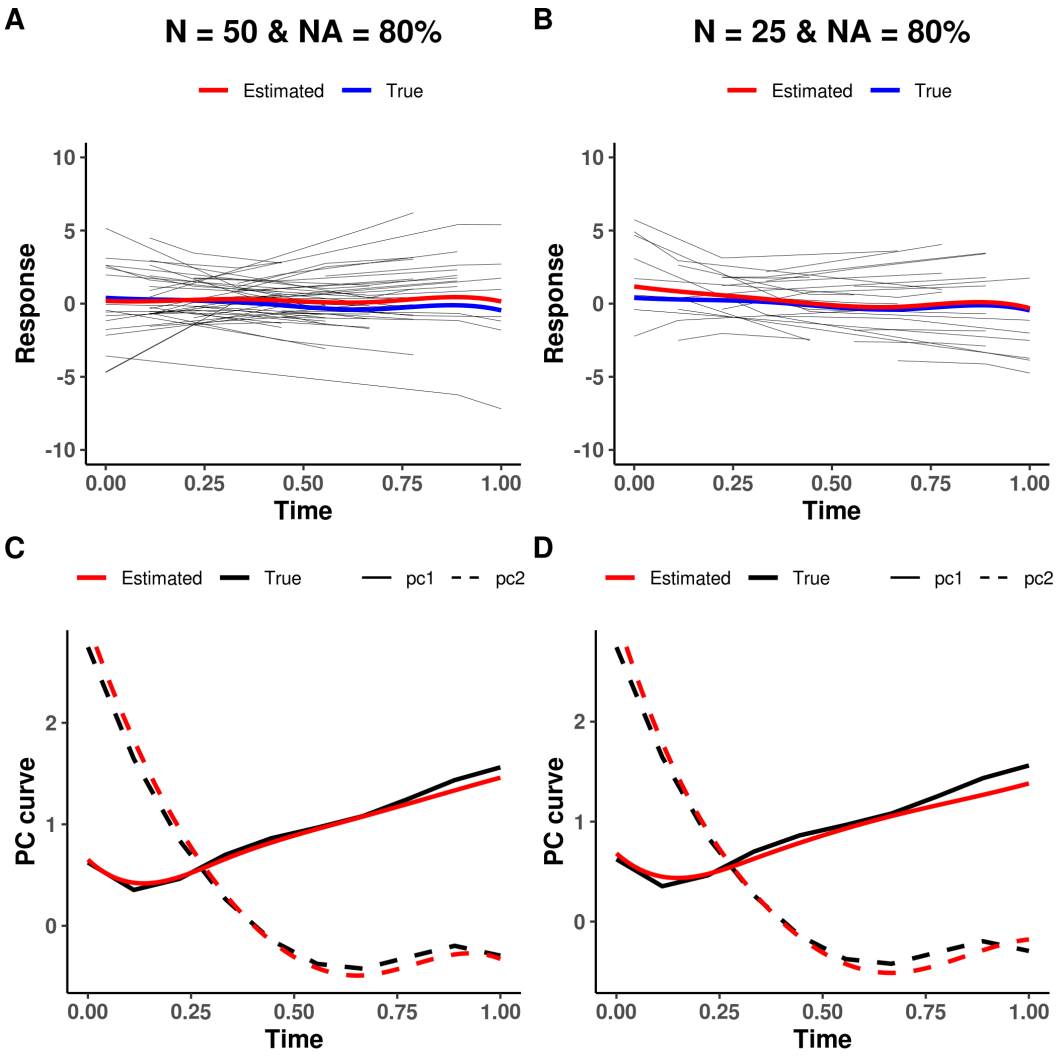

Figure 2.3: Results of Bayesian SFPCA on simulated data with 50 vs. 25 total samples of 80% missing data.

Table 2.2: Average mean squared errors with 95% CIs for estimating FPC spline coefficients.

	N = 100	N = 50	N = 25	N = 20
NA = 0%	0.002 (0.001, 0.01)	0.011 (0.001, 0.116)	0.02 (0.001, 0.143)	0.049 (0.003, 0.21)
NA = 20%	0.002 (0.001, 0.005)	0.005 (0.001, 0.03)	0.018 (0.001, 0.157)	0.05 (0.003, 0.208)
NA = 50%	0.002 (0.001, 0.005)	0.005 (0.001, 0.026)	0.02 (0.001, 0.162)	0.054 (0.003, 0.203)
NA = 80%	0.004 (0.001, 0.016)	0.01 (0.001, 0.079)	0.027 (0.002, 0.208)	0.081 (0.017, 0.246)

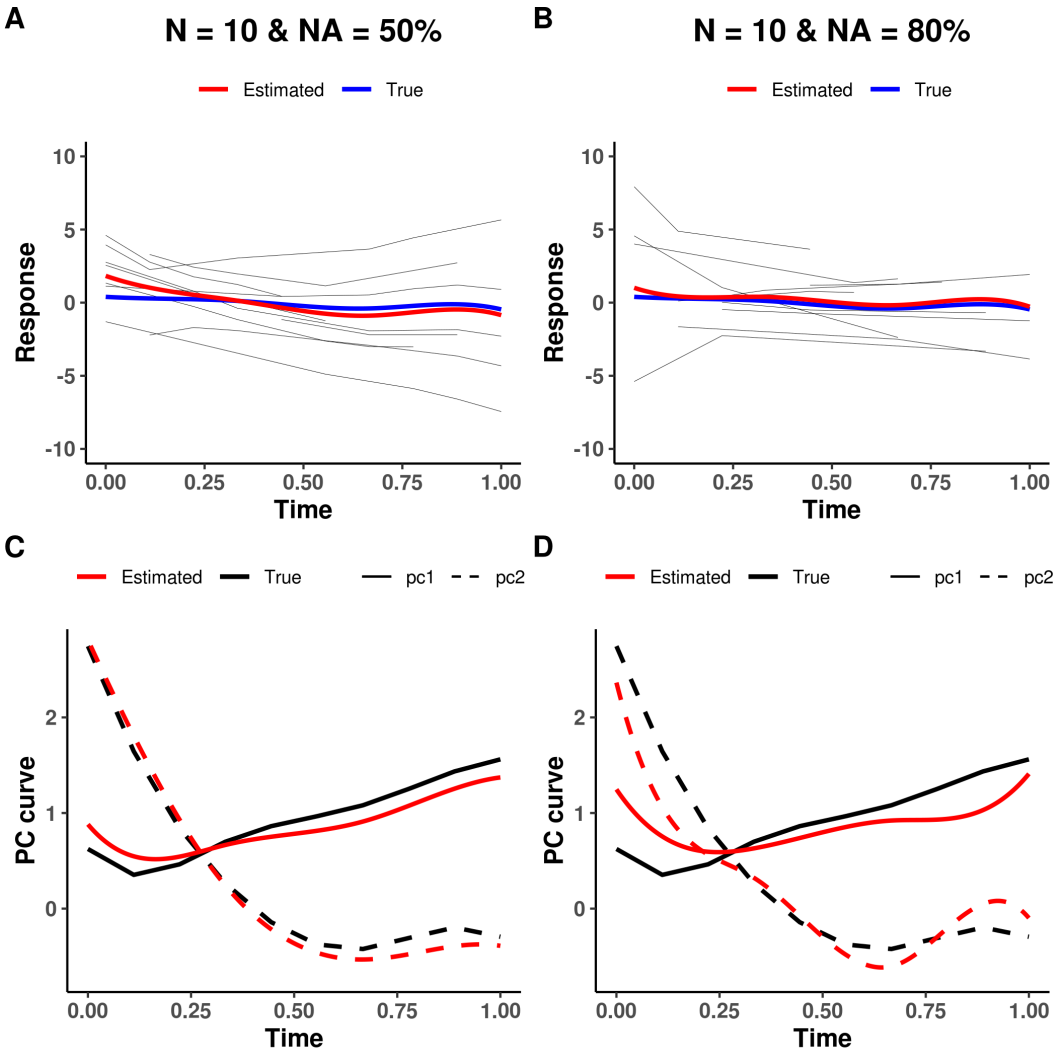

Figure 2.4: Results of Bayesian SFPCA on simulated data with 10 total samples of 50% vs. 80% missing data.

2.5.2 Impact of Skin Care Products on Microbiome Dynamics

In this example, researchers want to know how the skin microbiome would be altered when the hygiene routine is modified, and whether this alteration is similar across different body sites (Bouslimani et al., 2019). Twelve healthy subjects participated in this 9-week study, and samples were collected from each individual on four skin body sites (face, armpits, front forearms and toes). For the baseline (week 0), subjects performed their normal routine of using their personal skin care products. During the first three weeks (w1-w3), all volunteers used only the same head-to-toe shampoo and no other product was applied. In the following 3 weeks (w4-w6), four selected commercial products were applied daily by all volunteers on the specific body site: sunscreen for the face, deodorant antiperspirant for the armpits, moisturizer for the front forearm, and soothing foot powder for the toes, and continued use of the same shampoo. For the last three weeks (w7-w9), all volunteers went back to their normal routine using their personal products. Due to its specific study design, the perturbations of the skin microbiome are expected to occur around the intervention time points. The outcome of interest in this example is the longitudinal pattern of Shannon microbial diversity of the microbiome, defined as $Shannon = -\sum_{i=1}^{S} p_i ln(p_i)$, where S is the total number of species, and p_i is the relative proportion of species i relative to the entire population.

The best SFPCA model was selected by PSIS-LOO to have four PCs and three internal knots for the cubic spline basis. The estimated difference of expected leave-one-out prediction errors between the models with three and four PCs was smaller than the standard error, hence they could both be considered as adequate models. We chose the model with the highest value of $\widehat{elppd_{psis-loo}}$, which has four principal components and three internal knots. The model diagnostics using graphical posterior predictive checks showed that the simulated data from the posterior predictive distribution was able to cover the distribution of observed outcomes well (Figure 2.5A). Moreover, the estimated shape parameters from PSIS-LOO were all under the threshold of 0.7, except for one subject with a marginal value at 0.71 and another with an extreme value at 1.15 (Figure 2.5B). To examine these two potential outliers, we compared their observed trajectories with the predicted curves. Figure 2.5C, D showed that the observed trajectories (black) were closely followed by the predicted curves (red) and fell within the 95% credible intervals (blue). All these suggested that our selected SFPCA model was able to fit this dataset well.

As seen in Figure 2.6A, the population mean curve reveals an overall trend of an initial decrease in microbial diversity during the first 3 weeks due to the cessation of using personal skin care products, an increase in the middle 3 weeks because of the introduction of four additional products, and a decrease toward the end due to the resumption of normal routines. Figure 2.6B shows the first four estimated PCs, with the first two PCs explaining over 90% of the variance. The first principal component captures variation in changes in microbial diversity around week 2.5 and week 5. The second component captures additional variation in changes around week 8. In Figure 2.6C-D, by adding a PC with ± 1 standard deviation of PC scores to the population mean curve, we illustrate how the first and second PCs impact the trajectories. The first PC represents an overall vertical shift of the mean microbial diversity, and explains about 80% of the variance. An individual with a high score on this component has on average higher microbial diversity than one with a lower score, and *vice versa*. The second PC curve explains 12% of the variance, and captures variation during the middle three weeks. Since a trajectory of each individual is represented as a weighted sum of these principal modes of variation, we can use each individual's PC scores to gain insight about microbial perturbations in different body sites (Figure 2.6E-F). The scores of the first PC unveil the order of microbial diversity from highest to lowest in the four body sites, where arm has the highest diversity over time, while armpit the lowest. The signs of mean scores (positive or negative) indicate that arm and face share one similar temporal pattern, corresponding to the orange curve in Figure 2.6C, while foot and armpit share another temporal pattern, corresponding to the blue curve in Figure 2.6C. A similar temporal clustering of face and arm, versus foot and armpit was observed in scores of the second PC as well.

2.6 Discussion

We have introduced a Bayesian approach to SFPCA, providing users an efficient Bayesian model selection technique like PSIS-LOO and reliable model diagnostics methods such as examining the estimated shape parameters from PSIS-LOO and utilizing the graphical posterior predictive checks. Moreover, our Bayesian modeling approach is flexible in incorporating alternative prior distributions, for example, a t-distribution to capture heavy tails in the distribution of principal component scores α_i, which are easily implemented in Stan. The examples in Section 4 demonstrate the potential of this Bayesian approach to

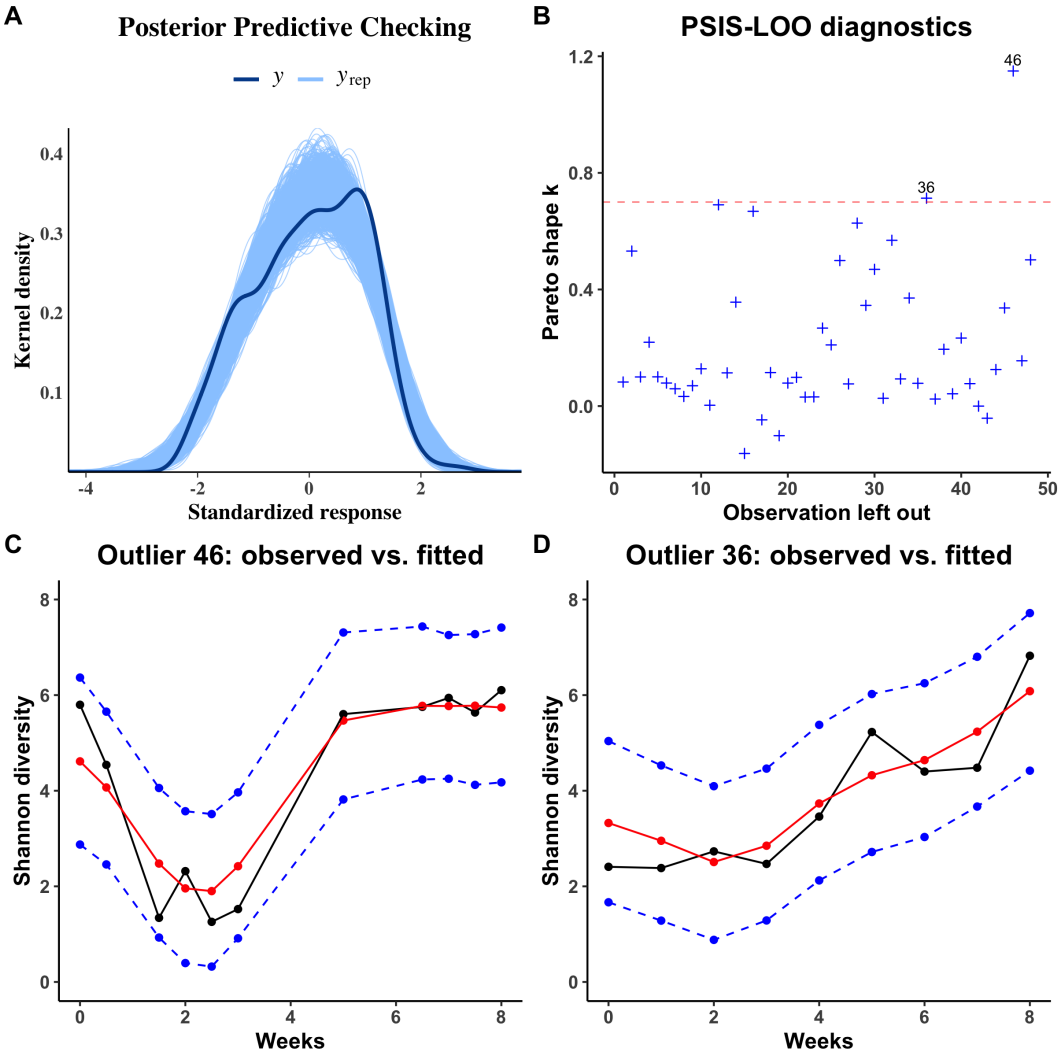

Figure 2.5: Graphical model diagnostics with posterior predictive checks and PSIS-LOO diagnostic plot for Bayesian SFPCA application to skin microbiome dataset.

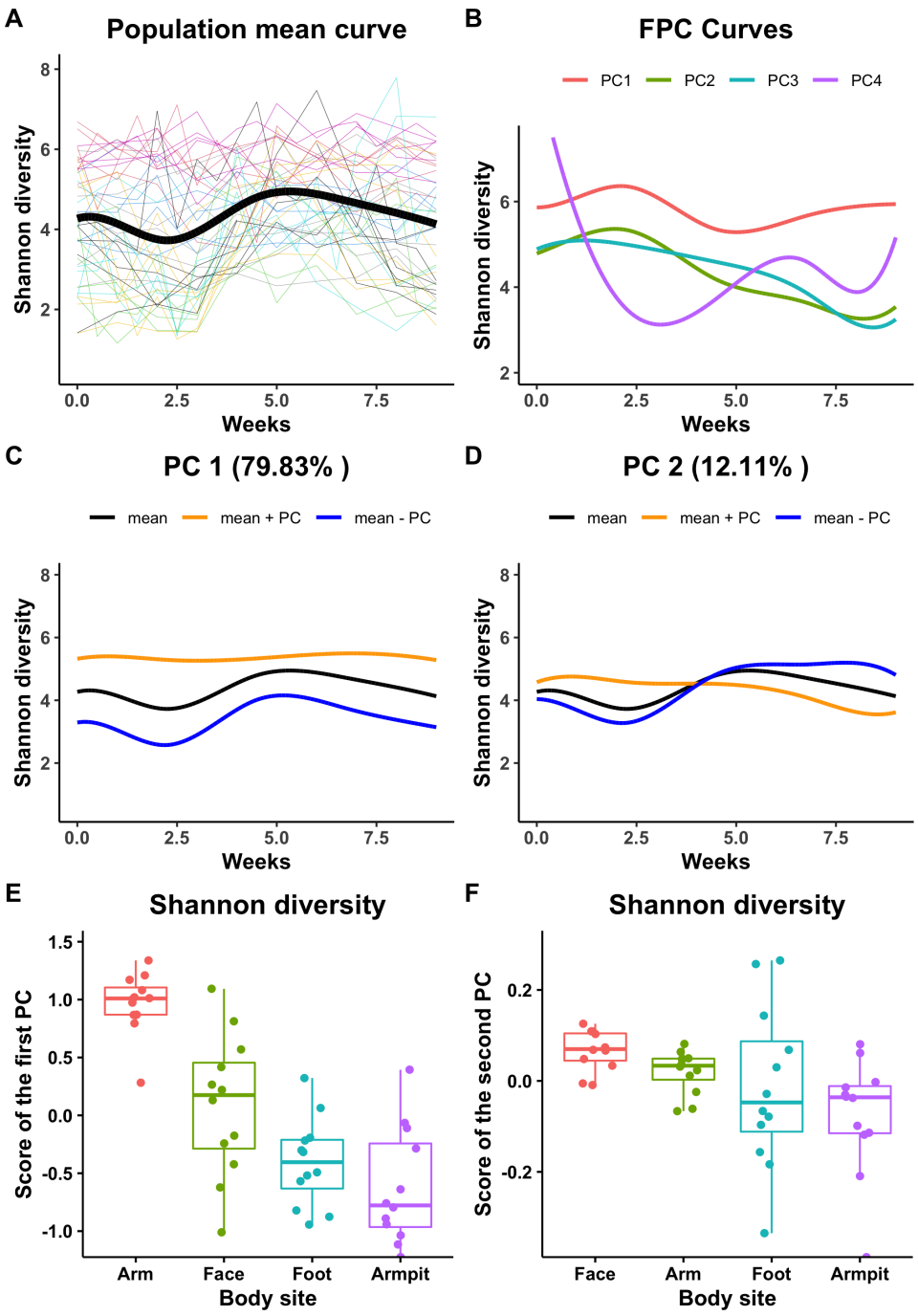

Figure 2.6: Results of Bayesian SFPCA on skincare impact microbiome dataset.

select the optimal model and uncover meaningful biological insights after careful model implementation and diagnostics. The first limitation of our current Bayesian SFPCA method is that it can only model one temporal measurement for different subjects over time, while microbiome data are typically comprised of thousands of microbes. This drawback would restrict the microbiome applications of this method to mainly analyses of alpha diversity, changes or differences in beta diversity, or measurement of a specific microbe. But given the flexibility in Bayesian modeling, it is feasible to extend the current model to multiple outcome measures simultaneously. The second limitation of our method is that the Bayesian implementation of SFPCA is less computationally efficient than frequentist approaches. However, with the goal of building more valid and reliable models in real data applications, our flexible modeling options, model selection and diagnostics with PSIS-LOO grants advantages over other available SFPCA approaches. Moreover, since the SFPCA model is implemented in Stan, a programming language with a very active user base, the BayesTime R package will be able to be updated with more efficient MCMC sampling algorithms and also incorporate other groundbreaking model selection and diagnostic techniques whenever they become available. Hence, we believe that the Bayesian approach to SFPCA will enable broader applications to a wider range of longitudinal data analysis going forward.

2.7 Acknowledgements

Chapter 2, in full, has been submitted for publication and is presented as it may appear in "Jiang, L.; Zhong, Y.; Elrod, C.; Natarajan, L.; Knight, R.; Thompson, W.K. *BayesTime: Bayesian Functional Principal Components for Sparse Longitudinal Data.* Journal of Computational and Graphical Statistics." The dissertation author was the primary investigator and author of this work.

We thank Yoshiki Vázquez-Baeza, Tomasz Kościółek, and Antonio González for suggestions and insights on microbiome data analysis, and Jeff De Reus for advice on high performance computing.

LN was partially supported by funding from the National Institute of Health grants: NIDDK 1R01DK110541-01A1 and NIA 1P01AG052352-01A1. RK was supported by NIH under grant 1DP1AT010885, NIDDK under grant 1P30DK120515, and CCFA under grant 675191. WT was supported by NIH/NIMH under grants RF1 MH120025 and R01 MH122688.

Chapter 3

Bayesian Multivariate Sparse Functional Principal Components Analysis with Applications to Longitudinal Microbiome Multi-Omics Data

3.1 Abstract

Microbiome researchers often need to model the temporal dynamics of multiple complex, nonlinear outcome trajectories simultaneously. This motivates our development of *multivariate Sparse Functional Principal Components Analysis* (mSFPCA), which extends existing FPCA models to simultaneously characterize multiple temporal trajectories and their inter-relationships. As with existing FPCA methods, the mSFPCA algorithm characterizes each trajectory as a smooth mean plus a weighted combination of the major (smooth) modes of variation about the mean, where the weights are given by the component scores for each subject. Unlike existing FPCA methods, the mSFPCA algorithm allows for estimation of multiple trajectories, such that the component scores, which are constrained to be independent within a particular outcome for identifiability, may be arbitrarily correlated with component scores for other

outcomes. A Cholesky decomposition is used to estimate the component score covariance matrix efficiently and guarantee positive semi-definiteness given these constraints, and mutual information to assess the strength of marginal and conditional temporal associations across outcomes. Importantly, we implement mSFPCA as a Bayesian algorithm using R and stan, which enables the usage of PSIS-LOO for model selection and graphical posterior predictive checks to assess the validity of mSFPCA models. While we focus on application of mSFPCA to microbiome data in this paper, the mSFPCA model is of general utility and can be used in a wide range of real data applications.

3.2 Introduction

Numerous diseases, including inflammatory bowel disease (IBD), heritable immune-mediated diseases such as asthma, neurological conditions including autism, and genetically driven diseases such as cancer, have been linked to dysregulation of human microbiota (Holleran et al., 2018; Lloyd-Price et al., 2019; Frati et al., 2019; Sharon et al., 2019; Ballen et al., 2016). However, the complex influence of microbiota on human health is not yet functionally understood. Ultimately, to understand the link between the human microbiome and disease it is necessary to determine which microbe genes are being expressed as well as the timing of their expression (Sberro et al., 2019). Thus, in addition to obtaining microbiome data using 16S ribosomal RNA gene sequencing or whole genome shotgun sequencing (Kuczynski et al., 2010; Ranjan et al., 2016; Gill et al., 2006), an increasing number of studies are also collecting transcriptomics data in order to understand microbial gene expression, proteomics data to study expressed proteins, and metabolomics data to define the functional status of host-microbial relationships (iHMP Consortium, 2014; Lloyd-Price et al., 2019; Bouslimani et al., 2019). This complex combination of data types, called *microbiome multi-omics*, is essential for understanding the links between microbial communities and disease and may enable translation of microbiome research into effective treatments.

An increasing number of microbiome multi-omics studies are longitudinal, aimed at simultaneously characterizing microbiome and host temporal changes in order to provide a more comprehensive picture of the dynamic changes during healthy and diseased states (iHMP Consortium, 2014; Lloyd-Price et al., 2019; Vatanen et al., 2018; Stewart et al., 2018). Despite these breakthroughs in microbiome study

designs and data collections, few statistical methods are available to analyze these complex longitudinal omics data. Recently, several new methods based on network analysis were developed for multi-omics integration of microbiome data in cross-sectional studies (Jiang et al., 2019; Morton et al., 2019), however, analytical methods for longitudinal microbiome multi-omics data are still in their infancy. The challenges include irregular timing and frequency across subjects, unmatched time points between different data types, non-linear temporal patterns, missing data, and high individual variability (Bodein et al., 2019).

The field of functional data analysis sheds some light on modeling such challenging and complex type of longitudinal data. The statistical framework of functional data analysis (FDA) is a term introduced by Ramsay and Silverman (Ramsay and Silverman, 1997), where the basic unit of information is the entire function, such as a curve or image. Functional principal component analysis (FPCA) has been widely used and serves as a fundamental tool for developing advanced methods for functional data analysis. The fundamental aims of this method include capturing the principal directions of variation and dimension reduction. FPCA summarizes the subject-specific features as the coordinates (called principal component scores) of subject curves in the basis spanned by the principal components (Di et al., 2009). Recent works include approaches of the smoothed FPCA based on a roughness penalty (Rice and Silverman, 1991), the FPC methods for sparsely sampled functional data (James et al., 2000; Yao et al., 2005; Peng and Paul, 2009; Di et al., 2014; Kidziński and Hastie, 2018), and asymptotic properties of the classical FPCA (Hall and Hosseini-Nasab, 2006; Li et al., 2010). Despite this burgeoning interest in FPCA research, most work has been focused on univariate functional data. Chiou et al. (2014) proposed a multivariate FPCA method to simultaneously model multiple temporal measurements and infer the component dependencies through the pairwise cross-covariance functions. However, this method is limited to the classical functional data, where the curves are observed longitudinally over densely sampled time points.

To meet the need for modeling irregularly and sparsely sampled, non-linear multivariate microbiome multi-omics trajectories, we developed multivariate sparse functional principal components analysis (mSFPCA). The major novelty of our approach is that it focuses on a set of functions which are not necessarily independent. Smoothing is accomplished through a few PC functions via the one-dimensional reduced rank mixed-effect model proposed by James et al. (2000), and then modeling the association of curves by jointly modeling the PC scores, whose covariance matrix is efficiently estimated by Cholesky

decomposition. Our proposed method allows for simultaneously characterizing multiple temporal measurements, such as microbiome, metabolome, inflammatory markers, and self-report measures, and to infer the temporal associations among these measures both marginally and conditionally, based on estimation of marginal and partial mutual information. Our model is employed in a Bayesian formulation and we use Hamilton Markov chain Monte Carlo (MCMC) methods in stan to sample from the posterior distribution of the model parameters. Our Bayesian implementation enables the usage of PSIS-LOO for model selection and graphical posterior predictive checks to assess the validity of mSFPCA models. While we focus on application of mSFPCA to microbiome data in this paper, the mSFPCA model is of general utility and can be used in a wide range of real data applications.

The remainder of the paper is organized as follows. Section 2 reviews the sparse functional principal component analysis (SFPCA), and introduces multivariate SFPCA, our statistical framework for longitudinal microbiome multi-omics data. Section 3 describes extensive simulation studies to evaluate performance of mSFPCA in realistic settings. Section 4 describes the application of our methodology to a challenging longitudinal microbiome multi-omics data on type 2 diabetes. Section 5 presents our conclusion. To ensure reproducibility of our results accompanying software, simulations and analysis results are posted at https://github.com/knightlab-analyses/mfpca-analyses.

3.3 Methodology

3.3.1 Sparse functional principal components analysis

The classical assumption of functional data analysis is that each trajectory is sampled over a dense grid of time points common to all individuals (Ramsay and Silverman, 2007). However, in practice, trajectories are often measured at an irregular and sparse set of time points that can differ widely across individuals. To address this issue, James et al. (2000) proposed *sparse functional principal components analysis* (SFPCA) using a reduced rank mixed-effects framework. Let $Y_i(t)$ be the measurement at time t for the ith individual, $\mu(t)$ the overall mean function, f_j the jth principal component function and $f = [(f_1, f_2, \ldots, f_k)]^T$, where k is the number of principal components. Then the James et al. (2000)

SFPCA model is given by

$$Y_i(t) = \mu(t) + \sum_{j=1}^{k} f_j(t)\alpha_{ij} + \varepsilon_i(t), \quad i = 1, ..., N \tag{3.1}$$

subject to the orthogonality constraint $\int f_j f_l = \delta_{jl}$, the Kronecker δ. The vector $\alpha_i = (\alpha_{i1}, \ldots, \alpha_{ik})^T$ is the component weights for the ith individual and $\varepsilon_i(t)$ is a normally-distributed residual, independent across subjects and across times within subject. The functions μ and f are approximated using cubic splines to allow for a smooth but flexible fit. Let $b(t)$ be a cubic spline basis with dimension $q > k$. The spline basis is orthonormalized so that $\int b_j b_l = \delta_{jl}$. Let Θ and θ_μ be, respectively, a $q \times k$ matrix and a q-dimensional vector of real-valued coefficients. For each individual i, denote their measurement times by $t = (t_{i1}, t_{i2}, \ldots, t_{in_i})^T$, and let $Y_i = (Y_i(t_{i1}), \ldots, Y_i(t_{in_i}))^T$ be the corresponding real-valued observations. Then $B_i = (b(t_{i1}), \ldots, b(t_{in_i}))^T$ is the $n_i \times q$ spline basis matrix for the ith individual. The reduced rank model can then be written as

$$\begin{aligned} Y_i &= B_i\theta_\mu + B_i\Theta\alpha_i + \varepsilon_i, \quad i = 1, ..., N, \\ \Theta^T\Theta &= I, \quad \alpha_i \sim N(0, D), \quad \varepsilon_i \sim N(0, \sigma_\varepsilon^2 I_{n_i}), \end{aligned} \tag{3.2}$$

where the covariance matrix D is restricted to be diagonal and I_{n_i} is the $n_i \times n_i$ identity matrix. Various fitting approaches, such as the EM algorithm, kernel smoothing and Newton-Raphson algorithm, have been proposed to estimate parameters of the SFPCA model (James et al., 2000; Yao et al., 2005; Peng and Paul, 2009). These approaches then use model selection techniques, such as cross-validation, Akaike information criterion (AIC) and leave-one-curve-out cross-validation, to select the dimension of basis and the number of principal components. However, due to their reliance on assumptions such as normally-distributed component scores and residuals, these models need to be carefully examined when applied to real data (Kidziński and Hastie, 2018).

3.3.2 Bayesian SFPCA

Jiang et al. (2020) proposed a SFPCA model in a Bayesian framework to allow for flexible prior specification and implementation of model selection and assessment methods. This Bayesian implementation used Hamilton MCMC sampling algorithm in Stan (Carpenter et al., 2017). The real-valued observations $Y_i(t)$ are first standardized to have mean zero and standard deviation one. The prior distributions for parameters in Eq. (3.2) were chosen as follows:

$$\begin{aligned}
\theta_\mu &\sim N_q(0, I_q) \\
\alpha_i &\sim N_k(0, I_k) \\
\Theta_j &\sim N_q(0, I_q), j = 1, \ldots, k \\
\varepsilon_i &\sim N_{v_i}(0, \sigma_\varepsilon^2 I_{v_i}) \\
\sigma_\varepsilon &\sim Cauchy(0, 1),
\end{aligned}$$

where Θ_j is the jth column of the loading matrix Θ, and v_i is the total number of visits for the ith subject. This Bayesian implementation enables use of leave-one-out cross-validation with Pareto-smoothed important sampling (PSIS-LOO) (Vehtari et al., 2017) to perform model selection on the number of principal components k and the number of basis functions q. Moreover, model fit can be assessed via diagnostics plots from PSIS-LOO as well as the graphical posterior predictive checks obtained from simulating posterior predictive data (Gelman et al., 1996; Gabry et al., 2019). This Bayesian implementation thus offers a flexible and comprehensive solution to real-date SFPCA applications.

3.3.3 Multivariate SFPCA

To model the P- dimensional multivariate response, we extend Bayesian SFPCA to simultaneously model multiple temporal measurements, and infer both their marginal and conditional temporal associations. For the pth temporal measurement, let K_p be the number of PCs, Q_p be the corresponding number of basis functions, V_{ip} be the total number of visits for ith subject in the pth temporal measurement, B_{ip} be the

transpose of the cubic spline basis, and Θ_p be the corresponding FPC loadings. Then the total number of principal components across p measurements are $K = \sum_{p=1}^{P} K_p$, the total number of basis functions are $Q = \sum_{p=1}^{P} Q_p$, and the total number of visits for subject i is $V_i = \sum_{p=1}^{P} V_{ip}$. The SFPCA model can be extended to be multivariate Sparse Functional PCA (mSFPCA) as

$$Y_i = B_i\theta_\mu + B_i\Theta\alpha_i + \varepsilon_i, i = 1,...,N, \tag{3.3}$$

where Y_i is a P- dimensional observed response, residuals $\varepsilon_i \sim N_{V_i}(0, \sigma_\varepsilon^2 I_{V_i})$, spline basis B_i is a $V_i \times Q$ matrix with

$$B_i = \begin{bmatrix} B_{i1} & 0^T & \cdots & 0^T \\ 0 & B_{i2} & \cdots & 0^T \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & B_{iP} \end{bmatrix},$$

Θ is the $Q \times K$ matrix of FPC loadings, subject to the orthonormality constraint $\Theta^T\Theta = \mathrm{I}$, defined as

$$\Theta = \begin{bmatrix} \Theta_1 & 0^T & \cdots & 0^T \\ 0 & \Theta_2 & \cdots & 0^T \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \Theta_P \end{bmatrix},$$

α_i is the K- dimensional FPC scores with $\alpha_i \sim N(0, \Sigma_\alpha)$ and Σ_α is restricted to the form

$$\begin{bmatrix} D_1 & C_{21}^T & \cdots & C_{P1}^T \\ C_{21} & D_2 & \cdots & C_{P2}^T \\ \vdots & \vdots & \ddots & \vdots \\ C_{P1} & C_{P2} & \cdots & D_P \end{bmatrix},$$

where D_p is the within-measurement diagonal covariance matrix for the pth measurement, and C_{lm} is the between-measurement covariance matrix for the lth and mth measurements.

Σ_α can also be written as $\Sigma_\alpha = S_\alpha R_\alpha S_\alpha$, where S_α is the diagonal matrix of standard deviations for PC scores, and R_α is the correlation matrix with the restricted form as

$$\begin{bmatrix} I_1 & R_{21}^T & \cdots & R_{P1}^T \\ R_{21} & I_2 & \cdots & R_{P2}^T \\ \vdots & \vdots & \ddots & \vdots \\ R_{P1} & R_{P2} & \cdots & I_P \end{bmatrix},$$

where I_p is the identity matrix corresponding to the pth measurement.

Similar to Bayesian SFPCA, we used Hamilton Markov Chain Monte Carlo (MCMC) sampling algorithm in Stan to estimate parameters, PSIS-LOO for model selection, and diagnostics plots from PSIS-LOO and graphical posterior predictive checks for model diagnostics. The prior distributions for θ_μ, Θ and ε_i are set as follows

$$\begin{aligned} \theta_\mu &\sim N_Q(0, I_Q) \\ \Theta_{kp} &\sim N_{Q_p}(0, I_{Q_p}), \\ \varepsilon_i &\sim N_{V_i}(0, \sigma_\varepsilon^2 I_{V_i}) \\ \sigma_\varepsilon &\sim Cauchy(0, 1), \end{aligned}$$

where Θ_{kp} is the kth column of the FPC loadings in the pth block.

Orthonormality constraint

One difficulty in implementing the Bayesian mSFPCA model is that the principal component loadings Θ are not uniquely specified. For a given $K \times K$ rotation matrix P, if $\Theta^* = \Theta P$ and Θ obeys the constraints in Eq.(3.3), then $\Theta^{*T}\Theta^* = P^T\Theta^T\Theta P = I$, and hence Θ is unidentifiable without additional restrictions. Instead of directly enforcing orthonormality when sampling from the conditional posteriors in the Bayesian model fitting, we sampled the parameters with no constraint on Θ and then performed

a *post hoc* rotation for each iteration of the MCMC algorithm to meet the orthonormality constraint. Since the symmetric matrix $\Theta\Sigma_\alpha\Theta^T$ is identifiable and non-negative definite, we applied an eigenvalue decomposition $\Theta\Sigma_\alpha\Theta^T = VSV^T$, where V is the $Q \times Q$ matrix of orthonormal eigenvectors, and S is the diagonal matrix of eigenvalues, with the Q positive eigenvalues ordered from largest to smallest. Let $\Theta^* = V_k$ denote the $Q \times K$ matrix consisting of the first K eigenvectors of V, which satisfies $\Theta^{*T}\Theta^* = I$. Finally, we rotated Σ_α and FPC scores α_i, to obtain $\Sigma^*_\alpha = \Theta^{T*}\Theta\Sigma_\alpha\Theta^T\Theta^*$, and $\alpha^*_i = \Theta^{*T}\Theta\alpha_i$, so that $\Theta^*\Sigma^*_\alpha\Theta^{T*} = \Theta\Sigma_\alpha\Theta^T$, and $\Theta^*\alpha^*_i = \Theta\alpha_i$.

Modeling covariance

Since the covariance matrix of FPC scores Σ_α has the constraint of positive semi-definiteness and it is restricted to the form of diagonal within-measurement covariance and any arbitrary form of between-measurement covariance structure, it is a challenge to model this covariance matrix effectively. Barnard et al. (2000) proposed a separation strategy for modeling $\Sigma = SRS$ by assuming independent priors for the standard deviations S and the correlation matrix R. To account for the dependent structure about correlations among different subsets of variables, Liechty et al. (2004) proposed the common correlation model for R, which assumes a common normal prior for all correlations with the additional restriction that the correlation matrix is positive definite. However, the awkward manner in which r_{ij}, the ijth element in the correlation matrix R, is embedded in the full conditional posterior density, leading to use a Metropolis-Hastings algorithm to update one coefficient r_{ij} at a time (Liechty et al., 2004). This consecutive updating procedure for correlation estimation is inefficient, and could lead to heavy computational cost when the correlation matrix is large or when the correlation has to be estimated separately from other parameters in mSFPCA model when implemented in Stan (Carpenter et al., 2017). For example, in a simulated data with 3 temporal measurements from 100 subjects over 10 time points, it would take 40 hours for a mSFPCA model using existing covariance estimation method to estimate all the parameters when implemented in Stan. However, the computational time can be reduced over 130 times (to only 18 minutes) by using our proposed method due to the avoidance of additional Metropolis-Hastings algorithm.

To pursue an efficient numerical solution to the covariance estimation, we took advantage of the Cholesky decomposition (Nash, 1990) and imposed the diagonal constraint on the within-measurement

covariance matrices. Since the covariance matrix of FPC scores Σ_α has full rank (highly unlikely that PC scores are correlated 100% across outcomes), it has a unique Cholesky decomposition in the form of

$$\Sigma_\alpha = LL^T,$$

where L is a real lower triangular matrix with positive diagonal entries (Gentle, 2012) . Given a lower triangular matrix L divided into P blocks, we have

$$L = \begin{bmatrix} L_{1,1} & 0 & \cdots & 0 \\ L_{2,1} & L_{2,2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ L_{P,1} & L_{P,2} & \cdots & L_{P,P} \end{bmatrix},$$

$$\text{then } LL^T = \begin{bmatrix} L_{1,1}L_{1,1}^T & * & \cdots & * \\ L_{2,1}L_{1,1}^T & L_{2,1}L_{2,1}^T + L_{2,2}L_{2,2}^T & \cdots & * \\ \vdots & \vdots & \ddots & \vdots \\ L_{P,1}L_{1,1}^T & L_{P,1}L_{2,1}^T + L_{P,2}L_{2,2}^T & \cdots & L_{P,1}L_{P,1}^T + \ldots + L_{P,P}L_{P,P}^T \end{bmatrix},$$

where $*$ denotes the transpose of the corresponding sub-diagonal block. To ensure that LL^T is positive definite with diagonal within-block covariance matrices, the lower triangular Cholesky factor L needs to meet the following two conditions:

1. Within-block covariance matrices $\sum_{m=1}^{M} L_{M,m}L_{M,m}^T, M = 1, \ldots, P$, are diagonal.

2. The diagonal entries of $L_{M,M}, M = 1, \ldots, P$ are positive.

We will focus on defining the diagonal blocks $L_{M,M}$ to achieve these, and leave the off-diagonal blocks $L_{M,m}, m = 1, \ldots, M-1$ to be arbitrary, unconstrained (i.e. the unconstrained parameter elements from the Hamiltonian MCMC sampling).

Let $D_M, M = 1, ..., P$ be the Mth within-block covariance matrix, then

$$\begin{aligned} D_M &= \sum_{m=1}^{M} L_{M,m} L_{M,m}^T \\ &= L_{M,M} L_{M,M}^T + \sum_{m=1}^{M-1} L_{M,m} L_{M,m}^T, \\ L_{M,M} L_{M,M}^T &= D_M - \sum_{m=1}^{M-1} L_{M,m} L_{M,m}^T = A. \end{aligned} \tag{3.4}$$

Since all the off-diagonal elements of D_M are known to be zero and the off-diagonal blocks $L_{M,m}, m = 1, ..., M-1$ are defined earlier with unconstrained estimates, we have thus defined all the off-diagonals of this matrix A, leaving only the diagonals. Because $L_{M,M}$ needs to have positive diagonal entries, $L_{M,M} L_{M,M}^T$ must be positive definite, thus $L_{M,M}$ is the Cholesky factor of A. To derive $L_{M,M}$, we can proceed with the Cholesky–Banachiewicz and Cholesky–Crout algorithm on A, where entries for the lower triangular factor L are

$$\begin{aligned} L_{j,j} &= \sqrt{Aj, j - \sum_{k=1}^{j-1} L_{j,k} L_{j,k}^T} \\ L_{i,j} &= \frac{1}{L_{j,j}} (A_{i,j} - \sum_{k=1}^{j-1} L_{i,k} L_{j,k}^T) \text{ for } i > j \end{aligned} \tag{3.5}$$

For the diagonal entries $L_{j,j}$, instead of using Eq.3.5, we substitute it with an exponential term $exp(0.5 * O + 2)$ to ensure it is positive, where O is the corresponding unconstrained parameter estimates. Here 0.5 was chosen to mimic the square root in the original formula, and 2 was added to bound initial values of diagonal entries away from zero, given the default initial values are drawn uniformly from the interval $(-2, 2)$ in Stan. Finally, we update the off-diagonal entries $L_{i,j}$ using the existing formula Eq.3.5.

In short, in our Bayesian implementation, we set the off-diagonal entries in within-measurement covariance matrices to be zero, estimate the rest of parameters without constraint using Cholesky algorithm and uninformative prior $uniform(-\infty, +\infty)$, and then substitute the diagonal entries of our exponential term. In this way, we are able to estimate covariance matrix efficiently and guarantee it to be positive semi-definite with our desired constrained form. Once we obtained the covariance matrix, we can then decompose it into correlation matrix R_α and standard deviations in order to estimate temporal associations.

Estimating inter-block association

Apart from simultaneously modeling multivariate longitudinal measurements, we want to estimate the association among measurements of interest via the correlations among the FPC scores, where the correlation matrix R_α obtained earlier will play a crucial role. We propose a new measure of inter-block association by calculating the mutual information of FPC scores from different measurements. Intuitively, mutual information measures the shared information between measurements: how much information is communicated, on average, in one measurement about the other.

We define the inter-block association between measurements p_1 and p_2 as the mutual information of FPC scores α_{ip_1} and α_{ip_2}, $1 \leq p_1, p_2 \leq P$, with

$$MI(\alpha_{ip_1}, \alpha_{ip_2}) = H(\alpha_{ip_1}) + H(\alpha_{ip_2}) - H(\alpha_{ip_1}, \alpha_{ip_2}),$$

where $H(X)$ is the entropy of X and $H(X) = -E[log(f_X(X))]$ with $f_X(X)$ being the probability density function of X(Cover, 1999).

If K-dimensional random variable X follows multivariate normal distribution with covariance matrix Σ, then according to Ahmed and Gokhale (1989)

$$H(X) = \frac{k}{2} + \frac{k}{2} log(2\pi) + \frac{1}{2}|\Sigma|.$$

Since the K-dimensional FPC scores $\alpha_i \sim N(0, \Sigma_\alpha)$, and any subvector of α_i is of the same structure with the correlation matrix being a submatrix of R_α, then according to Arellano-Valle et al. (2013), the mutual information of α_{ip_1} and α_{ip_2} could be simplified as

$$MI(\alpha_{ip_1}, \alpha_{ip_2}) = -\frac{1}{2} log|R_{\alpha\{p_1,p_2\}}|, \tag{3.6}$$

where

$$R_{\alpha\{p_1,p_2\}} = \begin{vmatrix} I_{p1} & R_{p_1p_2} \\ R^T_{p_1p_2} & I_{p2} \end{vmatrix}.$$

Moreover, we can estimate the conditional inter-block association between any two measurements of interest given the other measurements in the model by calculating the partial mutual information of FPC scores. The conditional inter-block association between measurements p_1 and p_2 is defined as the partial mutual information of α_{ip_1} and α_{ip_2}, $1 \leq p_1, p_2 \leq P$, with

$$\begin{aligned} MI(\alpha_{ip_1}, \alpha_{ip_2} | \alpha_{i\{1,\ldots,P\setminus p_1,p_2\}}) &= H(\alpha_{ip_1}, \alpha_{i\{1,\ldots,P\setminus p_1,p_2\}}) + H(\alpha_{ip_2}, \alpha_{i\{1,\ldots,P\setminus p_1,p_2\}}) \\ &\quad - H(\alpha_{i\{1,\ldots,P\setminus p_1,p_2\}}) - H(\alpha_{ip_1}, \alpha_{ip_2}, \alpha_{i\{1,\ldots,P\setminus p_1,p_2\}}) \\ &= H(\alpha_{i\{1,\ldots,P\setminus p_2\}}) + H(\alpha_{i\{1,\ldots,P\setminus p_1\}}) - H(\alpha_{i\{1,\ldots,P\setminus p_1,p_2\}}) - H(\alpha_i) \quad (3.7) \\ &= \frac{1}{2} log|R_{\alpha\{1,\ldots,P\setminus p_2\}}| + \frac{1}{2} log|R_{\alpha\{1,\ldots,P\setminus p_1\}}| \\ &\quad - \frac{1}{2} log|R_{\alpha\{1,\ldots,P\setminus p_1,p_2\}}| - \frac{1}{2} log|R_\alpha|, \end{aligned}$$

where $R_{\alpha\{1,\ldots,P\setminus p_2\}}$, $R_{\alpha\{1,\ldots,P\setminus p_1\}}$, and $R_{\alpha\{1,\ldots,P\setminus p_1,p_2\}}$ are defined in the similar way as $R_{\alpha\{p_1,p_2\}}$ in Eq.(3.6).

Inter-block association obtained from this way ranges from 0 to infinity. By analogy with the way Person's contingency coefficient was obtained, we can apply a simple transformation proposed by Joe (1989) to obtain a normalized version of the mutual information as

$$MI^*(\alpha_{ip_1}, \alpha_{ip_2}) := \sqrt{1 - exp[-2MI(\alpha_{ip_1}, \alpha_{ip_2})]}. \quad (3.8)$$

In this way, the inter-block and conditional associations now take its value in [0,1]. The interpretation is that the closer $MI^*(\alpha_{ip_1}, \alpha_{ip_2})$ or $MI^*(\alpha_{ip_1}, \alpha_{ip_2} | \alpha_{i\{1,\ldots,P\setminus p_1,p_2\}})$ is to 1, the higher the temporal association between measurements is.

3.4 Simulation studies

To evaluate the performance of mSFPCA in modeling multiple temporal measurements, especially in its covariance estimation and temporal association inference, we simulated sparse longitudinal trajectories with three temporal measurements under four different covariance structures. To better mimic the reality, our data was simulated based on an mSFPCA model using parameters initially estimated from a real longitudinal microbiome multi-omics data (Kostic et al., 2015) in the following way:

1. Applying mSFPCA to model three temporal measurements in the real multi-omics dataset.

2. Selecting the optimal number of PCs and dimension of basis using PSIS-LOO: the chosen model has the number of PCs as 2, 2, 1, and the number of basis as 6, 5, 5 for each measurement respectively.

3. Extracting the estimated values for population mean curve (θ_μ), FPC loadings (Θ), and residual variance σ_ε.

Then under four distinct covariance structures on FPC scores (Σ_α), we simulate the trajectories for 100 subjects with an average of 20% missing data over observations at 10 time points. Observations were randomly deleted to create increasingly sparse functional datasets. In the 1st covariance structure, all PCs are independent; in the 2nd covariance structure, only 1 strong correlation of 0.75 exists across all PCs; in the 3rd covariance structure, 1 strong and 1 medium correlation exists, at values of 0.75 and 0.5 respectively; in the 4th covariance structure, 1 strong, 1 medium and 1 weak correlations exists at strength of 0.75, 0.5 and 0.25. In short, there are increasing dependence structures among PCs as the covariance structure moves from the first to the last. Based on these pre-specified covariance structures and initially estimated parameters, we simulate the sparse longitudinal trajectories as follows:

1. Choosing the total number of subjects to be 100, and the number of time points to be 10 in order to place possible time points between $[0, 1]$.

2. Simulating the observed number of time points for each individual with $n_i \sim Poisson(8)$, where 8 represents the average number of time points across all subjects, and then randomly placing the observed time points in the possible time locations (chosen in the previous step).

3. Generating the cubic spline basis matrix B_i for each subject (orthonormality obtained through Gram-Schmidt orthonormalization).

4. Simulating for each subject FPC scores $\alpha_i \sim N(0, \Sigma_\alpha)$ and noise $\varepsilon_i \sim N(0, \sigma_\varepsilon^2 I)$.

5. Obtaining the temporal trajectory for each individual with $Y_i = B_i\theta_\mu + B_i\Theta\alpha_i + \varepsilon_i$.

6. Repeating step 1– 5 1000 times for each covariance structure, thus generating 4000 simulated datasets in total.

To evaluate the mSFPCA model performance in simulated data, we want to examine three main results: 1. how well mSFPCA can capture the temporal patterns embodied in the overall mean curve and FPC curves for each temporal measurement; 2. the accuracy of covariance estimation; 3. the inference on temporal associations based on mutual information estimation.

Figure 3.1 shows that the estimated overall mean curves and PC curves accurately recovered the ground truth for all three outcome variables under covariance structure I. This accurate capturing of major temporal patterns was seen in other three covariance structures as well (Figure 3.6, 3.7, 3.8). Figure 3.2 summarizes the performance of covariance estimation across all 4 scenarios in terms of the coverage probabilities of 95% credible intervals on estimated covariance parameters. The coverage probabilities are lowest in the 1st covariance structure (independent, Figure 3.2A), improved when more dependence structures are introduced (Figure 3.2B-D), and reach highest with the 4th covariance structure (having most correlations across PCs, Figure 3.2D). Despite these subtle differences in the coverage probabilities for each covariance parameter, the average coverage probability across all estimated parameters, represented by the dashed line, is around 95% within each covariance structure. This indicates that our mSFPCA model is able to estimate the covariance matrix properly, and its performance is affected by the structure of covariance matrix itself: the more sparse the covariance is, the more challenging the estimation. But even with the most sparse scenario (Figure 3.2 A), our mSFPCA model is still able to achieve about 95% average coverage probability.

Regarding the inference on temporal associations, Table 3.1 shows the mutual information estimates in each simulation scenario, which estimates the temporal association between each pair of temporal

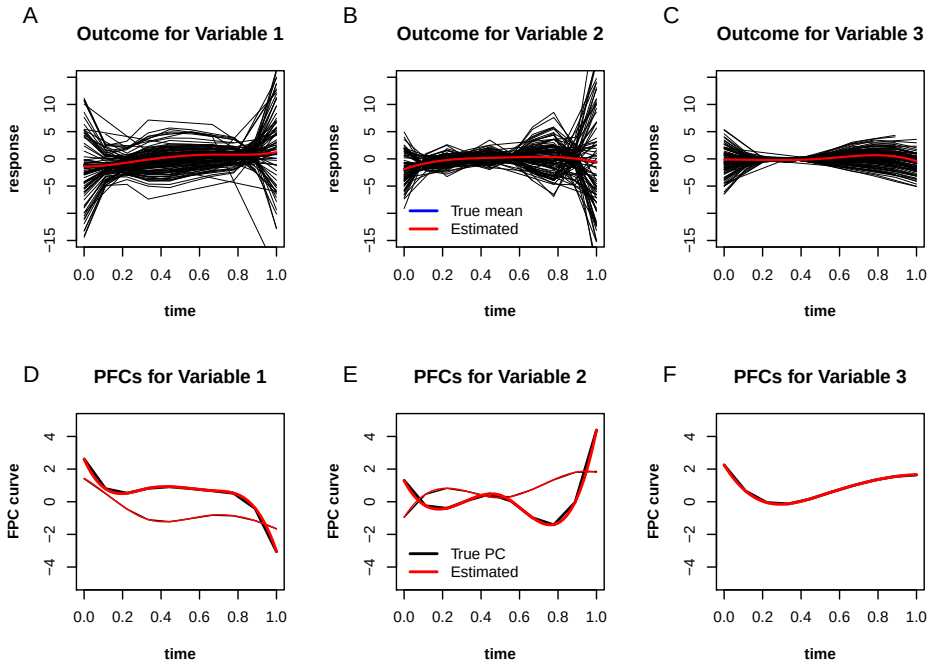

Figure 3.1: Estimated mean and FPC curves from mSFPCA on simulated data with covariance structure I.

measurements. MI_{ij} denotes the normalized mutual information between ith and jth temporal measurements. Except for the slightly lower coverage probability with 0.92 for MI_{12} in the 3rd scenario, or 0.94 for MI_{23} in the 4th scenario, all the coverage probabilities are close to 95%. Table 3.2 shows the conditional mutual information estimates in each simulation scenario, which estimates the temporal association between each pair of temporal measurements given the other measurement in the model. CMI_{ij} denotes the normalized conditional mutual information between ith and jth temporal measurements. All coverage probabilities are close to 95% on the estimation of conditional mutual information.

In short, our simulation results have demonstrated the good performance of mSFPCA in modeling sparse longitudinal data with multiple temporal measurements and providing valid inference on temporal associations.

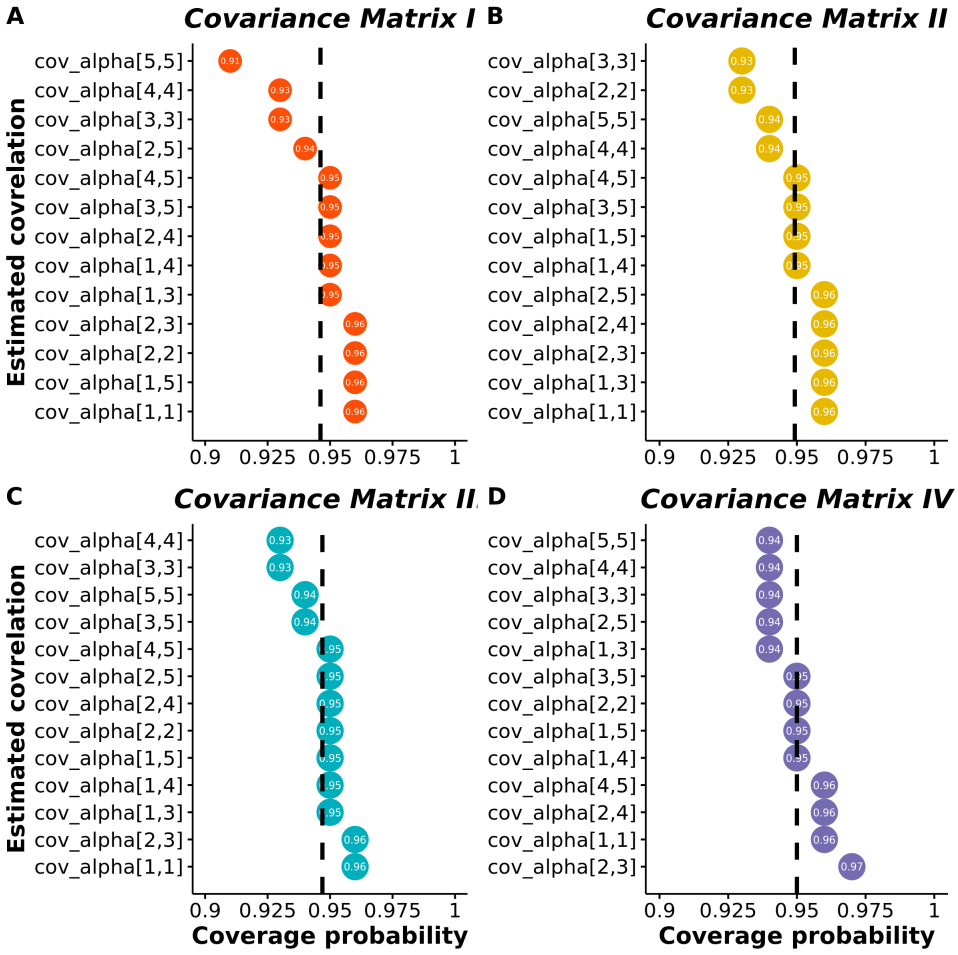

Figure 3.2: Coverage probability of 95% credible interval on estimated covariance parameters in four simulation scenarios.

Table 3.1: Mutual information estimates for each simulation scenario

				95% credible interval		
Simulation scenario	Parameter	Truth	Median	Cov.prob.	2.5%	97.5%
Covariance I	MI_{12}	0	0.26	0*	0.11	0.42
	MI_{13}	0	0.17	0*	0.04	0.34
	MI_{23}	0	0.18	0*	0.04	0.35
Covariance II	MI_{12}	0	0.26	0*	0.11	0.42
	MI_{13}	0	0.17	0*	0.04	0.34
	MI_{23}	0.75	0.75	0.96	0.65	0.83
Covariance III	MI_{12}	0.5	0.54	0.92	0.39	0.66
	MI_{13}	0	0.17	0*	0.04	0.34
	MI_{23}	0.75	0.75	0.96	0.65	0.83
Covariance IV	MI_{12}	0.5	0.54	0.94	0.38	0.66
	MI_{13}	0.25	0.29	0.94	0.12	0.46
	MI_{23}	0.75	0.75	0.93	0.66	0.83

Table 3.2: Conditional mutual information estimates for each simulation scenario

				95% credible interval		
Simulation scenario	Parameter	Truth	Median	Cov.prob.	2.5%	97.5%
Covariance I	CMI_{12}	0	0.26	0*	0.11	0.42
	CMI_{13}	0	0.17	0*	0.04	0.34
	CMI_{23}	0	0.18	0*	0.04	0.35
Covariance II	CMI_{12}	0	0.26	0*	0.12	0.42
	CMI_{13}	0	0.17	0*	0.04	0.34
	CMI_{23}	0.75	0.75	0.95	0.65	0.83
Covariance III	CMI_{12}	0.76	0.77	0.94	0.68	0.84
	CMI_{13}	0.66	0.66	0.95	0.53	0.76
	CMI_{23}	0.87	0.87	0.96	0.81	0.91
Covariance IV	CMI_{12}	0.81	0.82	0.95	0.74	0.88
	CMI_{13}	0.76	0.76	0.95	0.66	0.83
	CMI_{23}	0.89	0.90	0.95	0.85	0.93

3.5 Real data application

For the real data application, we want to model multiple temporal measurements simultaneously in a large and challenging dataset, with a special interest in utilizing conditional mutual information to infer temporal association. This dataset comes from the type 2 diabetes (T2D) longitudinal studies in the Integrative Human Microbiome Project (iHMP Consortium, 2014). In this example, an over 3 years' study has been conducted in approximately 100 individuals at high risk for T2D, in order to better understand the biological changes that occur during the onset and progression of T2D. Multiple sample types were collected from the study participants every 2-3 months during their healthy periods, with more frequent sampling during periods of respiratory illness and other environmental stressors. These data include multi-omics assays such as stool microbiome data using 16S rRNA sequencing, host protein expression profiles in fecal samples using LC-MS/MS, and cytokine profiles that quantify the levels of 50 diverse inflammatory proteins and insulin peptides in host serum, as well as standard clinical tests results like hemoglobin A1c (HbA1c), insulin and glucose. Moreover, behavior changes of patients, such as emotional and psychological stress, were documented using the Perceived Stress Scale instrument. Our outcomes of interest are the longitudinal pattern of Shannon diversities in bacteria, proteins and cytokines, and of clinical test results on HbA1c. Shannon diversity is defined as $Shannon = -\sum_{i=1}^{S} p_i ln(p_i)$, where S is the total number of species, and p_i is the relative proportion of species i relative to the entire population. We are particularly interested in utilizing mutual information to investigate which omics data have strongest association with HbA1c, and whether additional omics data improve the temporal association based on conditional mutual information.

The estimated mean curves in Figure 3.3 show different temporal trends in each outcome, where Shannon bacterial diversity decreases slowly over time, protein diversity increases steadily over time, cytokine diversity increases over the first 300 days, decreases between day 300 and 900, and then increases, and HbA1c decreases during the first 2 years, and increases slightly afterwards. As indicated by the observed trajectories for each individual (black curves), there are great subject-level variations in each outcome. This additional temporal information is captured by the FPC curves in Figure 3.4. Figure 3.4A shows the first two PCs in Shannon bacterial diversity, of which PC 1 explains 79% and captures variation

around day 750, while PC 2 explains 21% of the variation and emphasizes variation around day 300 and 1100. Figure 3.4B shows the first four PCs in Shannon protein diversity, of which the first two PCs explain over 90% of the variance. The first PC captures variation around day 750, and the second PC emphasizes variation around day 450 and 1200. Figure 3.4C shows the first four PCs in Shannon cytokine diversity, of which PC 1 explains 83% and exhibits an almost flat curve over time, while PC 2 explains 13% of the variation and emphasizes variation around day 300 and 800. Figure 3.4D shows the first four PCs in HbA1c: PC1 exhibits a slight increasing curve over time, accounting for 70% variation, and PC 2 captures variation around day 300 and 800, explaining for an additional 21% variation. In short, although principal patterns in each measurement vary, changing time points are pretty consistent, suggesting coherent responses to changes in patients' mental or physical conditions.

Among omics' temporal associations with standard clinical test result HbA1c, Table 3.3 suggests that Shannon protein diversity has the highest association with HbA1c, at an estimated mutual information of 0.91 with 95% credible interval (0.786, 0.971). Cytokine diversity is the second highest, with MI at 0.849 (0.668, 0.954), and bacteria diversity has the lowest association at 0.71 (0.441, 0.854). However, when information about other omics measurements are provided, all the pairwise temporal associations increase to over 0.95, as indicated by the conditional MI results. Regarding temporal associations among omics measurement, Shannon protein and bacteria diversities have highest temporal association, with mutual information at 0.982 (0.875, 0.999); Shannon protein and cytokine diversities also have high association at 0.966 (0.886, 0.998); the association between Shannon bacteria and cytokine is medium at 0.798 (0.454, 0.977). Similar to earlier results, when information about other measurements are available, all conditional information increase to 0.99. In short, host protein expression profiles data has highest temporal association with patients' diabetes status (i.e. HbA1c), but this information can still be further improved with additional omics data.

We need model diagnostics to conclude on the validity of our mSFPCA application. The optimal model selected by PSIS-LOO has 2 PCs for Shannon bacterial diversity, and 4 PCs for the other measurements, and the number of internal knot is chosen to be one for all outcomes. PSIS-LOO diagnostics in Figure 3.5A show that the selected mSFPCA model fit the majority of the data well, except for 4 outliers with Pareto shape k values higher than the warning threshold 0.7. Graphical posterior predictive checks in

Table 3.3: Mutual information estimates for type 2 diabetes multi-omics dataset application

temporal associations with HbA1c	$MI(95\%CI)$	$CMI(95\%CI)$
HbA1c—protein	0.910 (0.786, 0.971)	0.994 (0.968, 0.999)
HbA1c—cytokine	0.849 (0.668, 0.954)	0.986 (0.951, 0.999)
HbA1c—bacteria	0.710 (0.441, 0.854)	0.957 (0.820, 0.999)
temporal associations among omics	$MI(95\%CI)$	$CMI(95\%CI)$
protein—bacteria	0.982 (0.875, 0.999)	0.999 (0.996, 0.999)
protein—cytokine	0.966 (0.886, 0.998)	0.999 (0.997, 0.999)
bacteria—cytokine	0.798 (0.454, 0.977)	0.995 (0.958, 0.999)

Figure 3.5B suggests good model fit as the simulated data from the posterior predictive distribution was able to cover the distribution of observed outcomes well. Figure 3.5C-F highlight the observed trajectories of the 4 outliers detected by PSIS-LOO diagnostic plot. The red subject has highest curve in Shannon cytokine diversity and low value in HbA1c. A closer look at his/her metadata shows that this subject went through stages of healthy, infection and back to healthy. The green subject shows high oscillation pattern in Shannon protein diversity, as he/she oscillated between stages of healthy, inflammation, and infection. The blue subject exhibits high oscillation pattern in Shannon protein diversity, because he/she went through a complicated interweaving stages of healthy, inflammation, infection, post-travel and allergy. The purple subject, who has the highest Pareto shape k value in Figure 3.5A, experienced drastic change in HbA1c, as he/she went through stages of infection, stress and back to healthy. In short, our mSFPCA model generally fits this dataset well, and our diagnostic tools were able to highlight biologically meaningful outliers for further examination.

3.6 Discussion

We have introduced multivariate sparse functional PCA, an extension to the sparse functional principal components analysis, in modeling multiple temporal measurements simultaneously, and inferring the temporal associations among interested measurements based on estimation of mutual information. Our greatest methodological novelty lies in covariance matrix estimation, where we utilized Cholesky

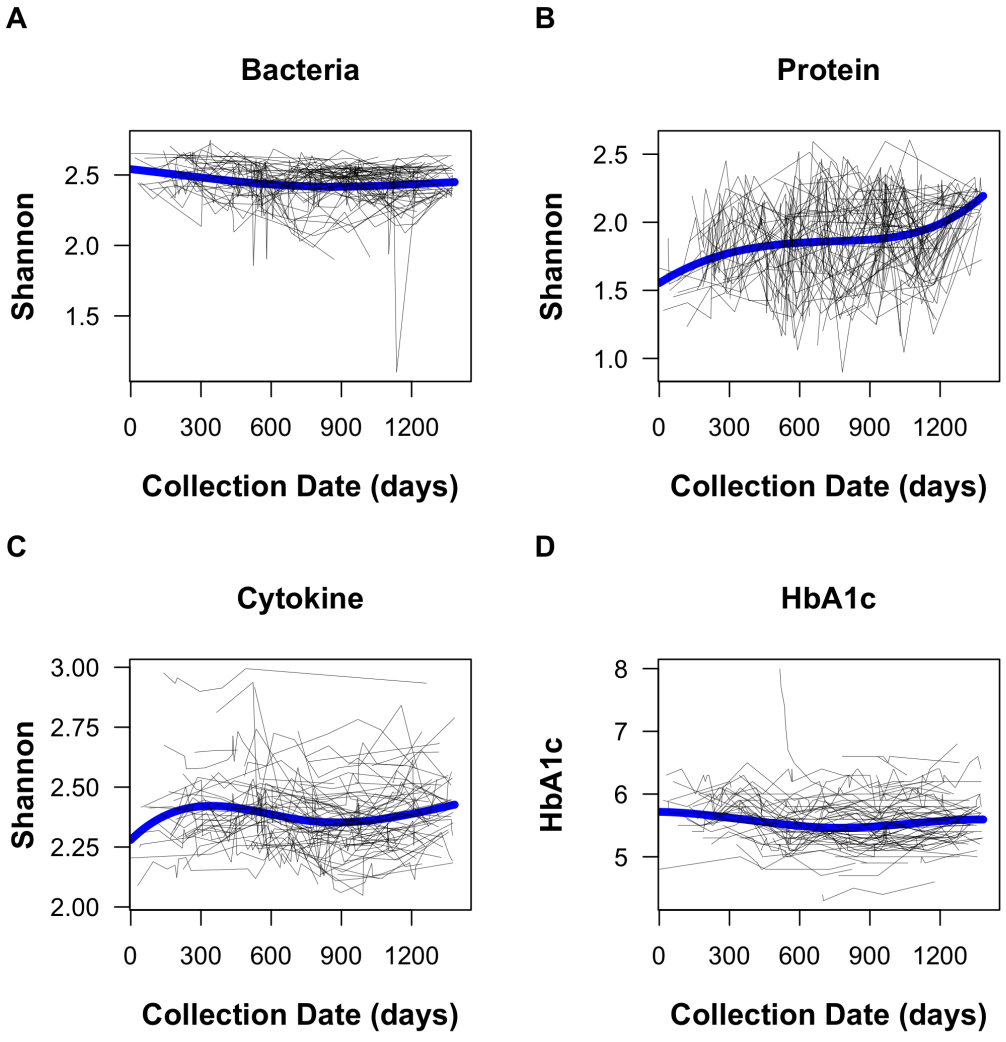

Figure 3.3: Estimated mean curves from mSFPCA application on Type 2 diabetes multi-omics dataset.

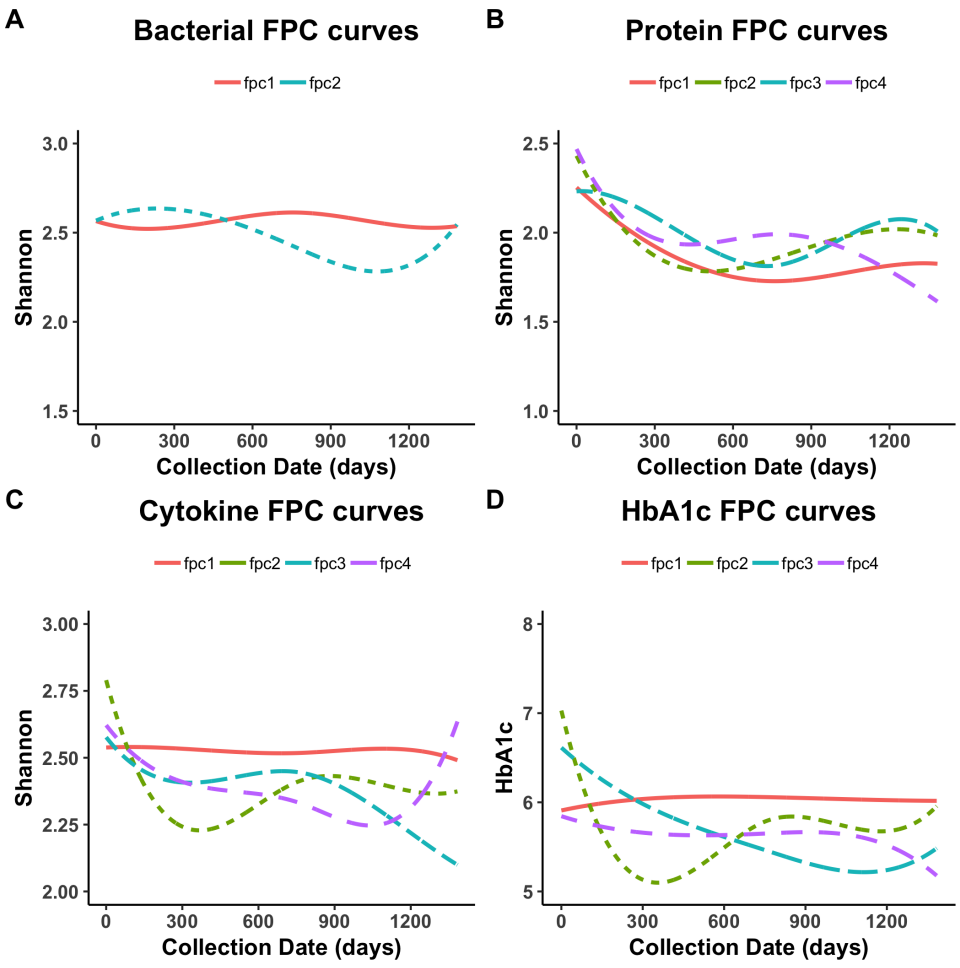

Figure 3.4: Estimated FPC curves from mSFPCA application on Type 2 diabetes multi-omics dataset.

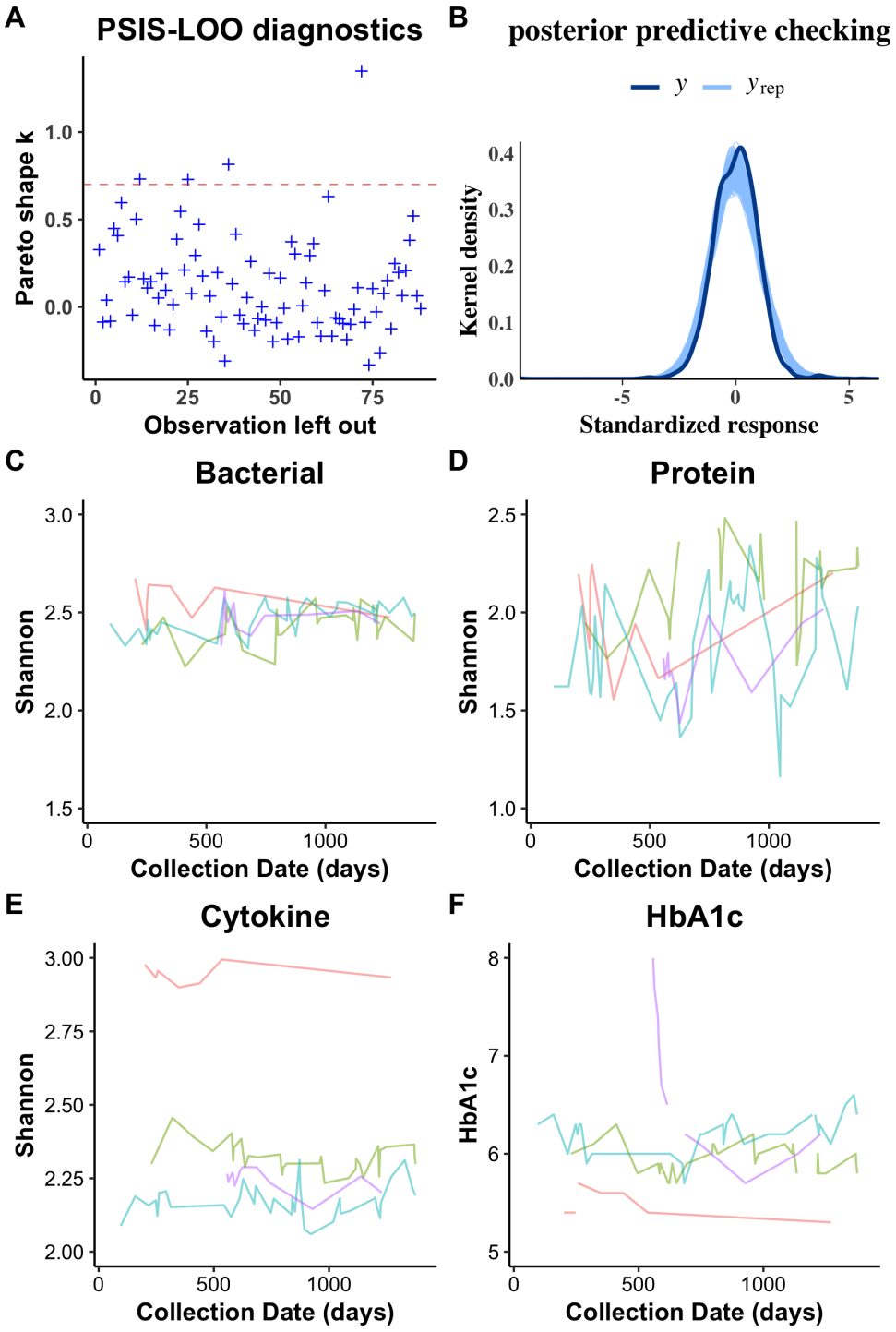

Figure 3.5: Graphical model diagnostics and examination of outliers for mSFPCA application on Type 2 diabetes multi-omics dataset.

decomposition to estimate covariance matrix efficiently and guarantee it to be positive semi-definite under the constrained form of covariance structure, i.e. diagonal within-measurement covariance and any arbitrary form of between-measurement covariance structure. Moreover, we utilized the concept of mutual information to define the marginal and conditional temporal associations, which provides a meaningful and interpretable measure to quantify temporal associations. Last but not least, our Bayesian implementation in Stan enables the usage of PSIS-LOO for efficient model selection, and visual model diagnostic methods, such as examining the estimated shape parameters from PSIS-LOO and utilizing the graphical posterior predictive checks, to evaluate the validity of mSFPCA models and highlight potential outliers.

In both our real-data based simulations and application to longitudinal microbiome multi-omics datasets, we have demonstrated that mSFPCA is able to accurately uncover the underlying principal modes of variation over time, including both the average population pattern and subject-level variations, and estimate the temporal associations properly. These enabled us to detect biologically meaningful signals in a large and challenging longitudinal cohort with irregular sampling, missing data, and four temporal measurements. Moreover, the model diagnostics plots from real data application show that our mSFPCA method can provide reliable model fitting to real microbiome multi-omics dataset. All these results highlight the great value of our method in modeling longitudinal data with multiple temporal measurements. Despite our applications to microbiome data in this paper, our mSFPCA method is in fact a general framework, and can be applied to a wide range of data, beyond the scope of longitudinal microbiome data.

One limitation of our method is that we assume the principal component scores and residuals to be normally distributed as in the original SFPCA model. This normality assumption would restrict our method from applying to highly skewed data. However, improper application of our method to such data could be detected by our model diagnostics tools, especially the graphical posterior predictive checks. Moreover, users could also modify our mSFPCA model by incorporating alternative prior distributions, for example, a t-distribution to capture heavy tails in the distribution of principal component scores, which can be easily implemented in Stan. Last but not least, since the mSFPCA model is implemented in Stan, a programming language with a very active user base, this method will be able to be updated with more efficient MCMC sampling algorithms and also incorporate other groundbreaking model selection and diagnostic techniques

whenever they become available. Hence, we believe that our mFPCA method will become a useful and up-to-date tool for researchers in various fields to analyze longitudinal data with multiple measurements in order to detect meaningful signals.

3.7 Acknowledgements

Chapter 3, in full, has been submitted for publication and is presented as it may appear in "Jiang, L.; Elrod, C.; Swafford, A.D.; Knight, R.; Thompson, W.K. *Bayesian Multivariate Sparse Functional Principal Components Analysis with Applications to Longitudinal Microbiome Multi-Omics Data.* Annals of Applied Statistics." The dissertation author was the primary investigator and author of this work.

RK was supported by NIH under grant 1DP1AT010885, NIDDK under grant 1P30DK120515, and CCFA under grant 675191. WT was supported by NIH/NIMH under grants RF1 MH120025 and R01 MH122688.

3.8 Supplementary Material

Three supplementary figures for simulation results are included as supplementary materials.

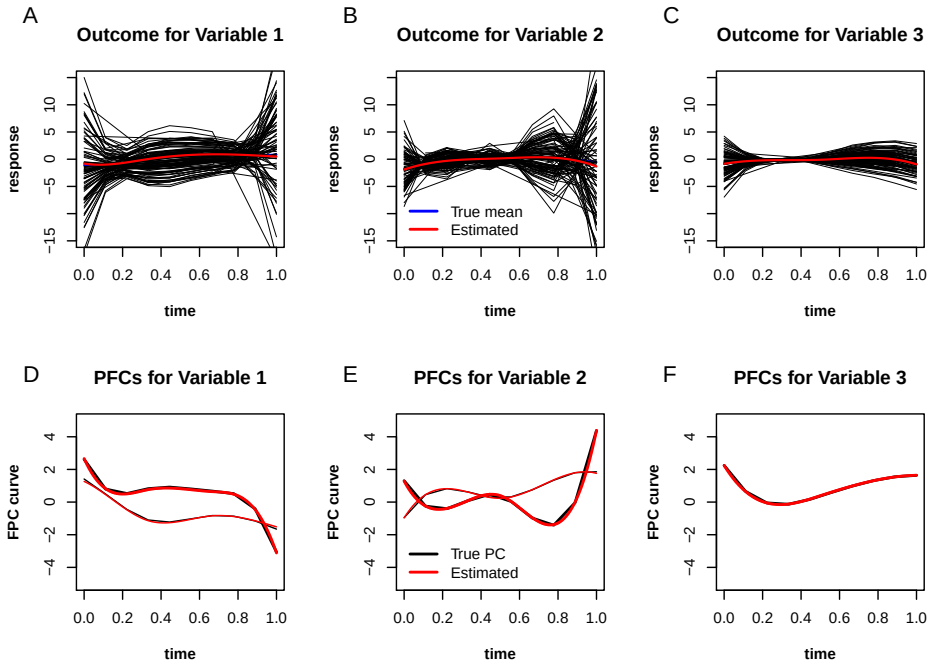

Figure 3.6: Estimated mean and FPC curves from mSMFPCA on simulated data with covariance structure II.

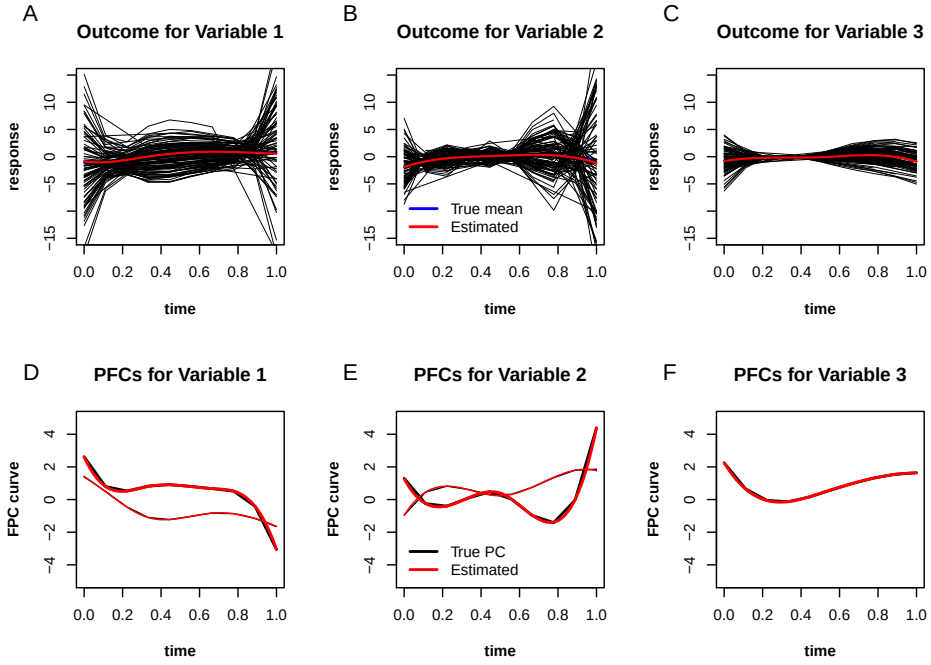

Figure 3.7: Estimated mean and FPC curves from mSMFPCA on simulated data with covariance structure III.

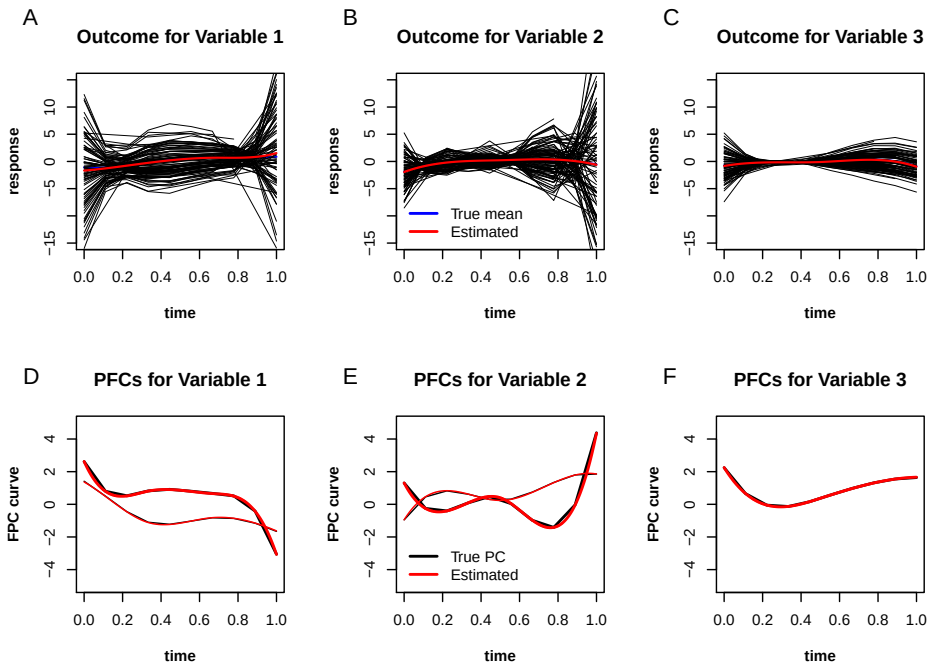

Figure 3.8: Estimated mean and FPC curves from mSMFPCA on simulated data with covariance structure IV.

Chapter 4

Utilizing Stability Criteria in Choosing Feature Selection Methods Yields Reproducible Results in Microbiome Data

4.1 Abstract

Feature selection is indispensable in microbiome data analysis, but it can be particularly challenging as microbiome data sets are high-dimensional, underdetermined, sparse and compositional. Great efforts have recently been made on developing new methods for feature selection that handle the above data characteristics, but almost all methods were evaluated based on performance of model predictions. However, little attention has been paid to address a fundamental question: how appropriate are those evaluation criteria? Most feature selection methods often control the model fit, but the ability to identify meaningful subsets of features cannot be evaluated simply based on the prediction accuracy. If tiny changes to the training data would lead to large changes in the chosen feature subset, then many of the biological features that an algorithm has found are likely to be a data artifact rather than real biological signal. This crucial need of identifying relevant and reproducible features motivated the reproducibility evaluation criterion such as Stability, which quantifies how robust a method is to perturbations in the data. In our paper, we

compare the performance of popular model prediction metric MSE and proposed reproducibility criterion Stability in evaluating four widely used feature selection methods in both simulations and experimental microbiome applications. We conclude that Stability is a preferred feature selection criterion over MSE because it better quantifies the reproducibility of the feature selection method.

4.2 Introduction

Feature selection is indispensable for predicting clinical or biological outcomes from microbiome data as researchers are often interested in identifying the most relevant microbial features associated with a given outcome. This task can be particularly challenging in microbiome analyses, as the datasets are typically high-dimensional, underdetermined (the number of features far exceeds the number of samples), sparse (a large number of zeros are present), and compositional (the relative abundance of taxa in a sample sum to one). Current methodological research has been focusing on developing and identifying the best methods for feature selection that handle the above characteristics of microbiome data, however, methods are typically evaluated based on overall performance of model prediction, such as Mean Squared Error (MSE), R-squared or Area Under the Curve (AUC). While prediction accuracy is important, another possibly more biologically relevant criterion for choosing an optimal feature selection method is reproducibility, i.e. how reproducible are all discovered features in unseen (independent) samples? If a feature selection method is identifying true signals in a microbiome dataset, then we would expect those discovered features to be found in other similar datasets using the same method, indicating high reproducibility of the method. If a feature selection method yields a good model fit yet poor reproducibility, then its discovered features will mislead related biological interpretation. The notion of reproducibility for evaluating feature selection method seems intuitive and sensible, yet in reality we neither have access to multiple similar datasets to estimate reproducibility, nor have a well-defined mathematical formula to define reproducibility. The many available resampling techniques (Efron and Tibshirani, 1994) enable us to utilize well-studied methods, for example bootstrapping, to create replicates of real microbiome datasets for estimating reproducibility. Moreover, given the burgeoning research in reproducibility estimation in the field of computer science (Kalousis et al., 2005, 2007; Nogueira, 2018), we can borrow their concept of

Stability to approximate the reproducibility of feature selection methods in microbiome data analysis.

In this paper, we investigate the performance of a popular model prediction metric MSE and the proposed feature selection criterion Stability in evaluating four widely used feature selection methods in microbiome analysis (lasso, elastic net, random forests and compositional lasso) (Tibshirani, 1996; Zou and Hastie, 2005; Breiman, 2001; Lin et al., 2014). We evaluate both extensive simulations and experimental microbiome applications, with a focus of feature selection analysis in the context of continuous outcomes. We find that Stability is a superior feature selection criterion to MSE as it is more reliable in discovering true and biologically meaningful signals. We thus suggest microbiome researchers use a reproducibility criterion such as Stability instead of a model prediction performance metric such as MSE for feature selection in microbiome data analysis.

4.3 Methods

4.3.1 Estimation of stability

The Stability of a feature selection method was defined as the robustness of the feature preferences it produces to differences in training sets drawn from the same generating distribution (Kalousis et al., 2005). If the subsets of chosen features are nearly static with respect to data changes, then this feature selection method is a *stable* procedure. Conversely, if small changes to the data result in significantly different feature subsets, then this method is considered *unstable*, and we should not trust the output as reflective of the true underlying structure influencing the outcome being predicted. In biomedical fields, this is a proxy for reproducible research, in the latter case indicating that the biological features the method has found are likely to be a data artifact, not a real clinical signal worth pursuing with further resources (Lee et al., 2013). Goh and Wong (2016) recommend augmenting statistical feature selection methods with concurrent analysis on stability and reproducibility to improve the quality of selected features prior to experimental validation (Sze and Schloss, 2016; Duvallet et al., 2017).

While the intuition behind the concept of stability is simple, there is to date no single agreed-upon measure for precisely quantifying stability. Up to now, there have been at least 16 different measures proposed to quantify the stability of feature selection algorithms in the field of computer science (Nogueira

et al., 2017). Given the variety of stability measures published, it is sensible to ask: which stability measure is most valid in the context of microbiome research? A multiplicity of methods for stability assessment may lead to publication bias in that researchers may be drawn toward the metric that extracts their hypothesized features or that reports their feature selection algorithm as more stable (Boulesteix and Slawski, 2009). Under the perspective that a useful measure should obey certain properties that are desirable in the domain of application, and provide capabilities that other measures do not, Nogueira and Brown aggregated and generalized the requirements of the literature into a set of five properties (Nogueira et al., 2017). The first property requires the stability estimator to be fully defined for any collection of feature subsets, thus allowing a feature selection algorithm to return a varying number of features. The second property requires the stability estimator to be a strictly decreasing function of the average variance of the selection of each feature. The third property requires the stability estimator to be bounded by constants not dependent on the overall number of features or the number of features selected. The fourth property states that a stability estimator should achieve its maximum if and only if all chosen feature sets are identical. The fifth property requires that under the null model of feature selection, where we independently draw feature subsets at random, the expected value of a stability estimator should be constant. These five properties are desirable in any reasonable feature selection scenario, and are critical for useful comparison and interpretation of stability values. Among all the existing measures, only Nogueira's stability measure (defined below) satisfies all five properties, thus we adopted this measure in the current work.

We assume a data set of n samples $\{x_i, y_i\}_{i=1}^{n}$ where each x_i is a p-dimensional feature vector and y_i is the associated biological outcome. The task of feature selection is to identify a feature subset, of size $k < p$, that conveys the maximum information about the outcome y. An ideal approach to measure stability is to first take M data sets drawn randomly from the same underlying population, to apply feature selection to each data set, and then to measure the variability in the M feature sets obtained. The collection of the M feature sets can be represented as a binary matrix Z of size $M \times p$, where a row represents a feature set (for a particular data set) and a column represents the selection of a given feature over the M data sets as

follows

$$Z = \begin{pmatrix} Z_{1,1} & \cdots & Z_{1,p} \\ \vdots & \ddots & \vdots \\ Z_{M,1} & \cdots & Z_{M,p} \end{pmatrix}$$

Let $Z_{.f}$ denote the f^{th} column of the binary matrix Z, indicating the selection of the f^{th} feature among the M data sets. Then $Z_{.f} \sim Bernoulli(p_f)$, where $\hat{p}_f = \frac{1}{M}\sum_{i=1}^{M} Z_{i,f}$ as the observed selection probability of the f^{th} feature. Nogueira defined the stability estimator as

$$\hat{\Phi}(Z) = 1 - \frac{\frac{1}{p}\sum_{f=1}^{p}\sigma_f^2}{E[\frac{1}{p}\sum_{f=1}^{p}\sigma_f^2|H_0]} = 1 - \frac{\frac{1}{p}\sum_{f=1}^{p}\sigma_f^2}{\frac{\bar{k}}{p}(1-\frac{\bar{k}}{p})} \tag{4.1}$$

where $\sigma_f^2 = \frac{M}{M-1}\hat{p}_f(1-\hat{p}_f)$ is the unbiased sample variance of the selection of the f^{th} feature, H_0 denotes the null model of feature selection (i.e. feature subsets are drawn independently at random), and $\bar{k} = \frac{1}{M}\sum_{i=1}^{M}\sum_{f=1}^{p} Z_{i,f}$ is the average number of selected features over the M data sets.

In practice, we usually only have one data sample (not M), so a typical approach to measure stability is to first take M bootstrap samples of the provided data set, and apply the procedure described in the previous paragraph. Other data sampling techniques can be used as well, but due to the well understood properties and familiarity of bootstrap to the community, we adopt the bootstrap approach.

4.3.2 Four selected feature selection methods

Lasso, elastic net, compositional lasso and random forests were chosen as benchmarked feature selection methods in this paper due to their wide application in microbiome community (Knights et al., 2011). Lasso is a penalized least squares method imposing an L_1-penalty on the regression coefficients (Tibshirani, 1996). Owing to the nature of the L_1-penalty, lasso does both continuous shrinkage and automatic variable selection simultaneously. One limitation of lasso is that if there is a group of variables among which the pairwise correlations are very high, then lasso tends to select one variable from the group and ignore the others. Elastic net is a generalization of lasso, imposing a convex combination of the L_1 and L_2 penalties, thus allowing elastic net to select groups of correlated variables when predictors are highly correlated (Zou and Hastie, 2005). Compositional lasso is an extension of lasso to compositional data

analysis (Lin et al., 2014), and it is one of the most highly cited compositional feature selection methods in microbiome analysis (Kurtz et al., 2015; Li, 2015; Shi et al., 2016; Silverman et al., 2017). Compositional lasso, or the sparse linear log-contrast model, considers variable selection via L_1 regularization. The log-contrast regression model expresses the continuous outcome of interest as a linear combination of the log-transformed compositions subject to a zero-sum constraint on the regression vector, which leads to the intuitive interpretation of the response as a linear combination of log-ratios of the original composition. Suppose an $n \times p$ matrix X consists of n samples of the composition of a mixture with p components, and suppose Y is a response variable depending on X. The nature of composition makes each row of X lie in a $(p-1)$-dimensional positive simplex $S^{p-1} = \{(x_1, \ldots, x_p) : x_j > 0, j = 1, \ldots, p \text{ and } \sum_{j=1}^{p} x_j = 1\}$. This compositional lasso model is then expressed as

$$y = Z\beta + \varepsilon, \sum_{j=1}^{p} \beta_j = 0 \tag{4.2}$$

where $Z = (z_1, \ldots, z_p) = (logx_{ij})$ is the $n \times p$ design matrix, and $\beta = (\beta_1, \ldots, \beta_p)^T$ is the p-vector of regression coefficients. Applying the L_1 regularization approach to this model is then

$$\hat{\beta} = argmin(\frac{1}{2n}||y - z\beta||_2^2 + \lambda||\beta||_1), \text{ subject to } \sum_{j=1}^{p} \beta_j = 0 \tag{4.3}$$

where $\beta = (\beta_1, \ldots, \beta_p)^T, \lambda > 0$ is a regularization parameter, and $||.||_2$ and $||.||_1$ denote the L_2 and L_1 norms, respectively.

Random forests is regarded as one of the most effective machine learning techniques for feature selection in microbiome analysis (Belk et al., 2018; Liu et al., 2017; Namkung, 2020; Santo et al., 2019; Statnikov et al., 2013). Random forests is a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest (Breiman, 2001). Since random forests do not select features but only assign importance scores to features, we choose features from random forests using Altmann's permutation test (Altmann et al., 2010), where the response variable is randomly permuted S times to construct new random forests and new importance scores computed. The S importance scores are then used to compute the p-value for the feature,

which is derived by computing the fraction of the S importance scores that are greater than the original importance score.

4.3.3 Simulation settings

We compared the performance of the popular model prediction metric MSE and the proposed criterion Stability in evaluating four widely used feature selection methods for different data scenarios. We simulated features with Independent, Toeplitz and Block correlation structures for datasets with the number of samples and features in all possible combinations of $(50, 100, 500, 1000)$, resulting in the ratio of p (number of features) over n (number of samples) ranging from 0.05 to 20. Our simulated compositional microbiome data are an extension of the simulation settings from Lin et al. (2014) as follows:

1. Generate an $n \times p$ data matrix $W = (w_{ij})$ from a multivariate normal distribution $N_p(\theta, \Sigma)$. To reflect the fact the components of a composition in metagenomic data often differ by orders of magnitude, let $\theta = (\theta_j)$ with $\theta_j = log(0.5p)$ for $j = 1, \ldots, 5$ and $\theta_j = 0$ otherwise. To describe different types of correlations among the components, we generated three general correlation structures: Independent design where covariates are independent from each other, Toeplitz design where $\Sigma = (\rho^{|i-j|})$ with $\rho = 0.1, 0.3, 0.5, 0.7, 0.9$, and Block design with 5 blocks, where the intra-block correlations are 0.1, 0.3, 0.5, 0.7, 0.9, and the inter-block correlation is 0.09.

2. Obtain the covariate matrix $X = (x_{ij})$ by the transformation $x_{ij} = \frac{exp(w_{ij})}{\sum_{k=1}^{p} exp(w_{ik})}$, and the $n \times p$ log-ratio matrix $z = log(X)$, which follows a logistic normal distribution (Aitchison, 1982).

3. Generate the responses y according to the model $y = Z\beta^* + \varepsilon$, $\sum_{j=1}^{p} \beta_j^* = 0$, where $\varepsilon \sim N(0, 0.5^2)$, and $\beta^* = (1, -0.8, 0.6, 0, 0, -1.5, -0.5, 1.2, 0, \ldots, 0)^T$, indicating that only 6 features are real signals.

4. Repeat steps 1-3 for 100 times to obtain 100 simulated datasets for each simulation setting, and apply the desired feature selection algorithm with 10-fold cross-validation on the 100 simulated datasets. Specifically, each simulated dataset is separated into training and test sets in the ratio of 8 : 2, 10-fold cross-validation is applied to the training set (80% of the data) for parameter tuning and variable selection, and then model prediction (i.e. MSE) is evaluated on the test set (20% of the

data). Hence, stability is measured according to Nogueira's definition based on the 100 subsets of selected features. Average MSE is calculated as the mean of the MSEs across the 100 simulated datasets, and the average false positive or false negative rate denotes the mean of the false positive or false negative rates across the 100 simulated datasets.

In summary, a total of 176 simulation scenarios were generated, with 16 for Independent design, 80 for Toeplitz or Block design, and 100 replicated datasets were simulated for each simulation setting, resulting in 17,600 simulated datasets in total.

4.4 Simulation results

Given that the true numbers of false positive and false negative features are known in simulations, we can utilize their relationships with MSE and Stability to compare the reliability of MSE and Stability in evaluating feature selection methods. In theory, we would expect to see a positive correlation between MSE and false positive rate or false negative rate, while a negative correlation between Stability and false positive or false negative rates. This is because when the real signals are harder to select (i.e. increasing false positive or false negative rates), a feature selection method would perform worse (i.e. increasing MSE or decreasing Stability). The first column in Figure 4.1 shows the relationship between MSE and false positive rate in three correlation designs, and the second column in Figure 4.1 shows the relationship between Stability and false positive rate. In contrast to the random pattern in MSE vs. false positive rate (Figure 4.1 A-C-E), where drastic increase in false positive rate could lead to little change in MSE (e.g. random forests), or big drop in MSE corresponds to little change in false positive rate (e.g. elastic net), we see a clear negative correlation pattern between Stability and false positive rate (Figure 4.1 B-D-F). Regarding false negative rate, we also observe a random pattern in MSE and a meaningful negative correlation relationship in Stability (Figure 4.4). These results suggest that Stability is a more reliable evaluation criterion than MSE due to its closer reflection of the ground truth in the simulations (i.e. false positive & false negative rates), and this is true irrespective of feature selection method used, features-to-sample size ratio (p/n) or correlation structure among the features.

Using the more reliable criterion Stability, we now investigate the best feature selection method

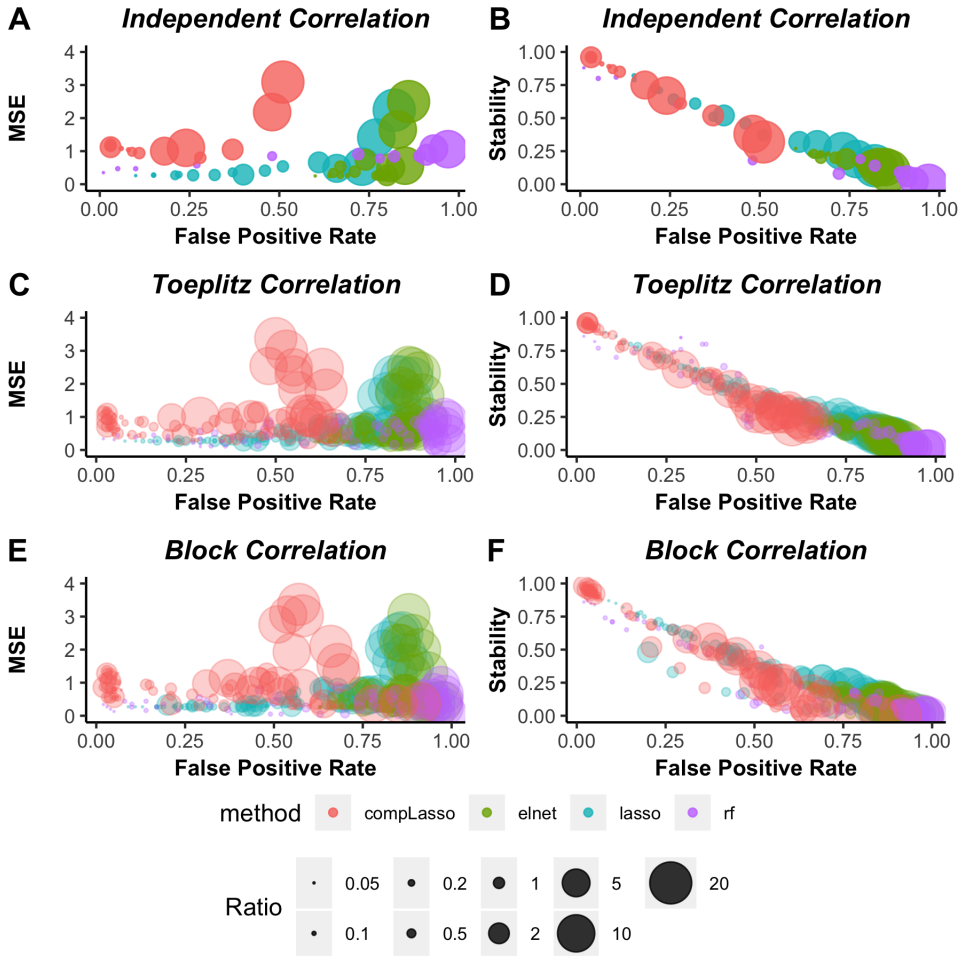

Figure 4.1: Comparing the relationship between MSE and False Positive Rate vs. Stability and False Positive Rate in three correlation structures.

in different simulation scenarios. Based on Stability, compositional lasso has the highest stability in "easier" correlation settings (Toeplitz 0.1 – 0.7 in Figure 4.5, represented by Toeplitz 0.5 in Figure 4.2 A due to their similar results; Block 0.9-0.3 in Figure 4.6, represented by Block 0.5 in Figure 4.2 C) for all combinations of n (number of samples) and p (number of features). Across all "easier" correlation scenarios, compositional lasso has an average stability of 0.76 with its minimum at 0.21 and its maximum close to 1 (0.97), while the 2nd best method Lasso has an average stability of only 0.44 with the range from 0.09 to 0.89, and the average stabilities of random forests and Elastic Net hit as low as 0.24 and 0.17 respectively. In "extreme" correlation settings (Toeplitz 0.9 in Figure 4.2 B or Block 0.1 in Figure 4.2 D), compositional lasso no longer maintains the highest stability across all scenarios, but it still has the highest average stability of 0.42 in Toeplitz 0.9 (surpassing the 2nd best Lasso by 0.09), and the second highest average stability in Block 0.1 (only 0.03 lower than the winner Lasso). Regarding specific scenarios in "extreme" correlation settings, compositional lasso, lasso or random forests can be the best in different combinations of p and n. For example, in both Toeplitz 0.9 and Block 0.1, with small p (when p = 50 or 100), random forests has highest stability (≥ 0.8) when n is largest ($n = 1000$), but Lasso or compositional lasso surpasses random forest when n is smaller than 1000, although all methods have poor stability (≤ 0.4) when $n \leq 100$. This indicates that best feature selection method based on Stability depends on the correlation structure among features, the number of samples and the number of features in each particular dataset; thus there is no single omnibus best, i.e., most stable, feature selection method.

How will results differ if we use MSE as the evaluation criterion? Using the extreme correlation settings (Toeplitz 0.9 and Block 0.1) as examples, random forests has lowest MSEs for all combinations of p and n (Figure 4.3 A-B). However, Figure 4.3 C-D unveils that random forests has highest false negative rates in all scenarios of Toeplitz 0.9 and Block 0.1, and its false negative rates can reach as high as the maximum 1, indicating that random forests fails to pick up any real signal despite its low prediction error. Moreover, Figure 4.3 E-F show that random forests can have highest false positive rates when p is as large as 500 or 1000. All these highlight the danger of choosing inappropriate feature selection method based on MSE, where the merit of high predictive power masks high errors in false positives and false negatives. On the other hand, the method with lowest false positive rates (compositional lasso) (Figure 4.3 E-F) was rather found to have the worst performance by MSE (Figure 4.3 A-B), suggesting another pitfall of missing

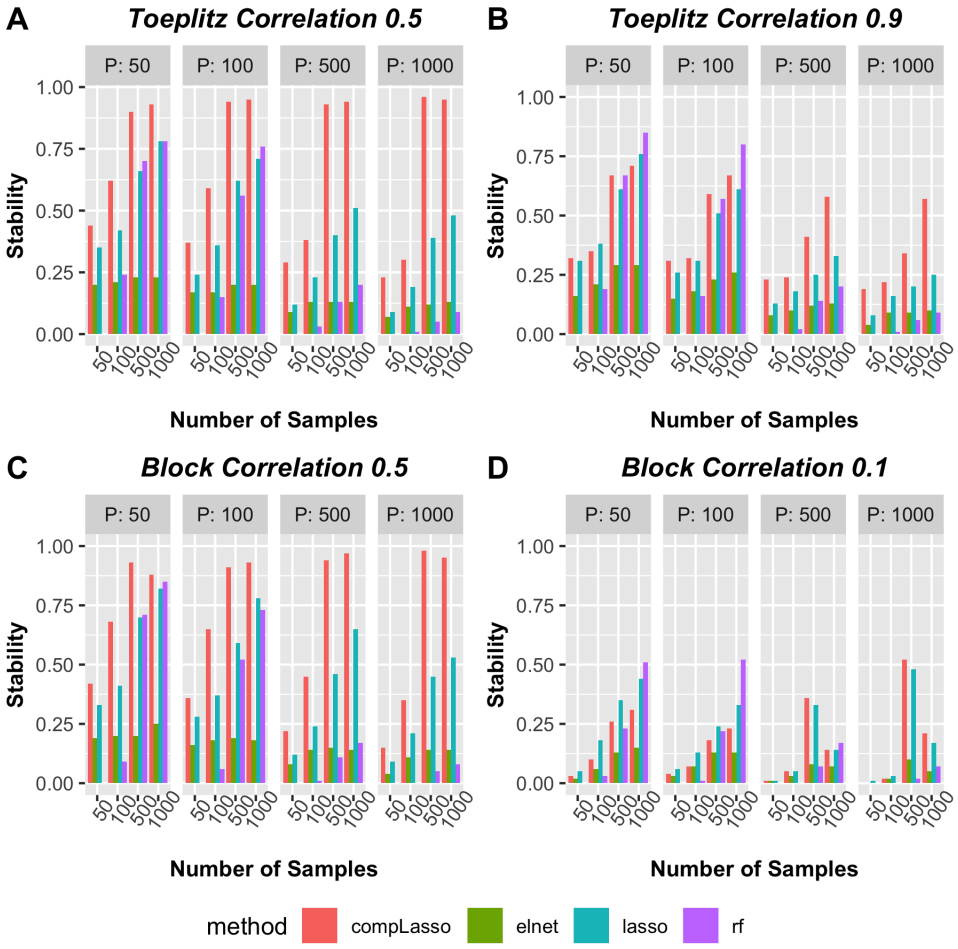

Figure 4.2: Method comparisons based on Stability in representative correlation structures.

Table 4.1: Hypothesis testing using Bootstrap to compare compositional lasso (CL) with random forests (RF) based on Stability or MSE using two simulation scenarios.

Example (N = 100 & P = 1000)	Estimated mean difference (CL – RF) in Stability index with 95% CI	Estimated mean difference (CL – RF) in MSE with 95% CI
Toeplitz 0.5	0.22 (0.19, 0.28)*	0.23 (-0.62, 1.36)
Block 0.5	0.23 (0.17, 0.29)*	0.44 (-0.27, 1.57)

the optimal method when using MSE as the evaluation criterion.

The use of point estimates alone to compare feature selection methods, without incorporating variability in these estimates, could be misleading. Hence, as a next step, we evaluate reliability of MSE and Stability across methods using a hypothesis testing framework. This is demonstrated with the cases of $n = 100$ & $p = 1000$ for Toeplitz 0.5 and Block 0.5, where compositional lasso is found to be the best feature selection method based on Stability, while random forests is the best based on MSE. We use bootstrap to construct 95% confidence intervals to compare compositional lasso vs. random forests based on Stability or MSE. For each simulated data (100 in total for Toeplitz 0.5 or Block 0.5), we generate 100 bootstrapped datasets and apply feature selection methods to each bootstrapped dataset. Then for each simulated data, Stability is calculated based on the 100 subsets of selected features from the bootstrapped replicates, and the variance of Stability is measured as its variability across the 100 simulated data. Since MSE can be obtained for each simulated data without bootstrapping, we use the variability of MSE across the 100 simulated data as its variance. Based on the 95% CI for the difference in Stability between compositional lasso and random forest methods (Table 4.1), we see that compositional lasso is better than random forest in terms of Stability index, and not statistically inferior to random forests in terms of MSE despite its inferior raw value. This suggests that Stability has higher precision (i.e. lower variance). Conversely, MSE has higher variance, which results in wider confidence intervals and its failure to differentiate methods.

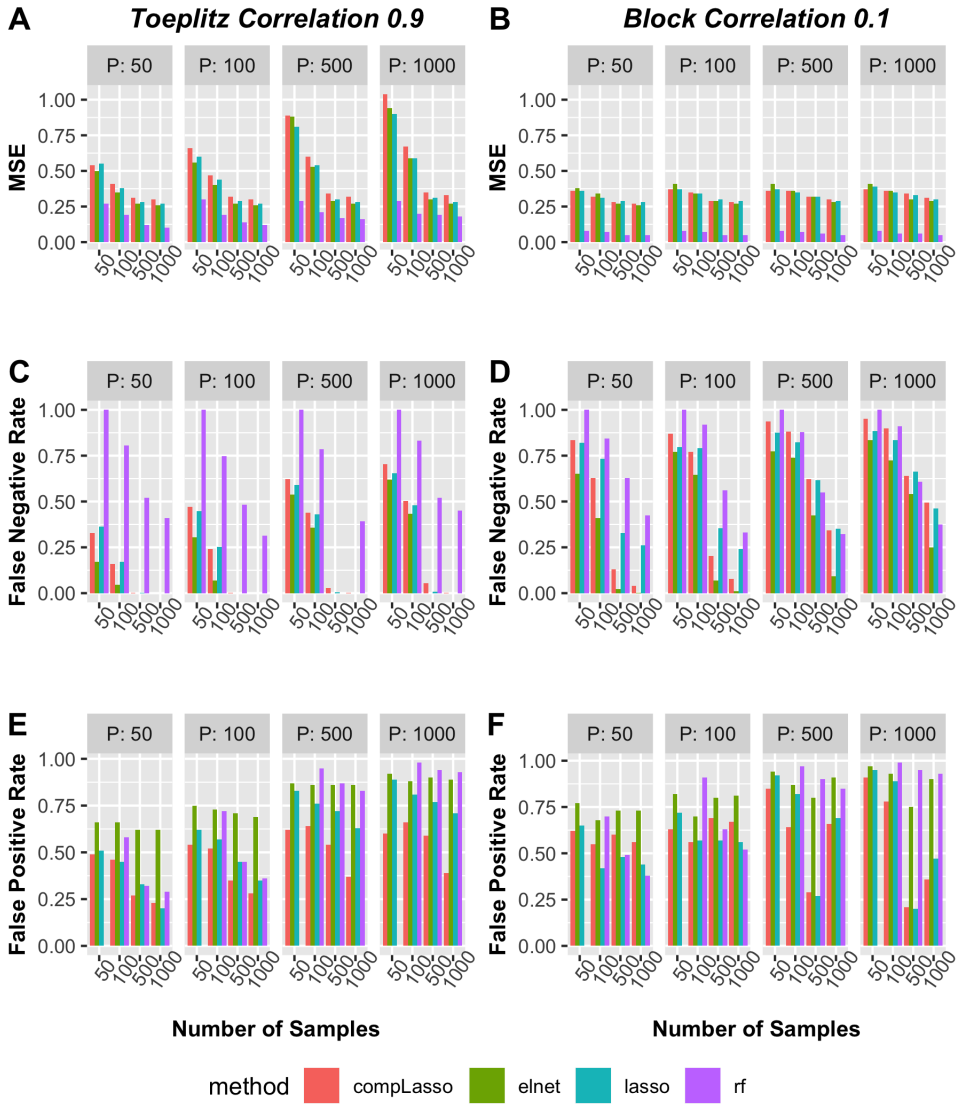

Figure 4.3: Method comparisons based on MSE in extreme correlation structures.

4.5 Experimental microbiome data applications

To compare the reliability of MSE and Stability in choosing feature selection methods in microbiome data applications, two experimental microbiome datasets were chosen to cover common sample types (human gut and environmental soil samples) and the scenarios of $p \approx n$ and $p >> n$ (where p is the number of features and n is the number of samples). The human gut dataset represents a cross-sectional study of 98 healthy volunteers to investigate the connections between long-term dietary patterns and gut microbiome composition (Wu et al., 2011), and we are interested in identifying a subset of important features associated with BMI, which is a widely-used gauge of human body fat and associated with the risk of diseases. The soil dataset contains 88 samples collected from a wide array of ecosystem types in North and South America (Lauber et al., 2009), and we are interested in discovering microbial features associated with the pH gradient, as pH was reported to be a strong driver behind fluctuations in the soil microbial communities (Morton et al., 2017). Prior to our feature selection analysis, the same filtering procedures were applied to the microbiome count data from these two datasets, where only the microbes with a taxonomy assignment at least to genus level or lower were retained for interpretation, and microbes present in fewer than 1% of the total samples were removed. Moreover, the count data were transformed into compositional data after replacing any zeroes by the maximum rounding error 0.5 (Lin et al., 2014).

Comparisons of feature selection methods in these two microbiome datasets are shown in Table 4.2, which are consistent with simulation results, where the best method chosen by MSE or Stability in each dataset can be drastically different. Based on MSE, random forests is the best in the BMI Gut dataset, while being the worst based on Stability. Similarly, in the pH Soil dataset, random forests is the second best method according to MSE, yet the worst in terms of Stability. If we use Stability as the evaluation criterion, then Elastic Net is the best in the BMI Gut and compositional lasso is the best in the pH Soil, yet both methods would be the worst if MSE was used as the evaluation criterion. One important note is that the Stability values in these two experimental microbiome datasets are low: none of the feature selection method exceeds a stability of 0.4, indicating the challenging task of feature selection in real microbiome applications. However, this possibility of low Stability values was already reflected in our simulated scenarios of "extreme" correlation scenarios. Another important note, which might be counter-intuitive, is

Table 4.2: Method comparisons based on Stability Index and MSE in experimental microbiome datasets.

Dataset	$n * p$ (p/n)	MSE (lower is better)	Stability (higher is better)
BMI Gut	98 * 87 (0.9)	Random forests (4.99) Compositional lasso (21.59) Lasso (24.07) Elastic Net (25.33)	Elastic Net (0.23) Compositional lasso (0.22) Lasso (0.14) Random forests (0.02)
pH Soil	89 * 2183 (24.5)	Elastic Net (0.23) Random forests (0.26) Lasso (0.34) Compositional lasso (0.46)	Compositional lasso (0.39) Lasso (0.31) Elastic Net (0.16) Random forests (0.04)

that the dataset with a high p/n ratio (pH Soil) has higher stabilities than the dataset with p/n ratio close to 1 (i.e. similar p & n values) (BMI Gut). This might be explained by the clearer microbial signals in environmental samples than in human gut samples, but it also highlights the impact of the dataset itself, whose characteristics cannot be easily summarized with the numbers of p and n, on feature selection results. Correlation structures between features as considered in our simulations could play an important role, and there may be many other unmeasured factors involved as well.

Apart from the comparisons based on point estimates, we can further compare MSE and Stability with hypothesis testing using nested bootstrap (Wainer and Cawley, 2018). The outer bootstrap generates 100 bootstrapped replicates of the experimental microbiome datasets, and the inner bootstrap generates 100 bootstrapped dataset for each bootstrapped replicate from the outer bootstrap. Feature selections are performed on each inner bootstrapped dataset with 10-fold cross-validation after a 80:20 split of training and test sets. The variance of Stability is calculated based on the Stability values across the outer bootstrap replicates, and the variance of MSE is calculated across both inner and outer bootstrap replicates, since MSE is available for each bootstrap replicate while Stability has to be estimated based on feature selection results across multiple bootstrap replicates. Using the datasets of BMI Gut and pH Soil, Table 4.3 confirms with simulation results that raw value difference in MSE does not indicate statistical difference, yet difference in Stability does help to differentiate methods due to its higher precision. A comparison between the observed difference in Table 4.2 and the estimated mean difference from bootstrap in Table 4.3 further confirms this discovery. Compared to the estimated mean differences between compositional

lasso and random forests based on stability (0.27 in the BMI Gut and 0.36 in the pH Soil), the observed differences (0.2 in the BMI Gut and 0.35 in the pH Soil) differ by 26% in the BMI Gut and 3% in the pH Soil. However, this difference is much more drastic based on MSE. Compared to the estimated mean differences between compositional lasso and random forests based on MSE (16.6 in the BMI Gut and 0.2 in the pH Soil), the observed differences (11.8 in the BMI Gut and 0.08 in the pH Soil) have huge differences of 41% and 160% in each dataset respectively. Hence, Stability is consistently shown to be more appropriate than MSE in experimental data applications as in simulations.

Table 4.3: Hypothesis testing using Bootstrap to compare compositional lasso (CL) with random forests (RF) based on Stability or MSE using two experimental microbiome datasets.

Dataset	Estimated mean difference (CL – RF) in Stability index with 95% CI	Estimated mean difference (CL – RF) in MSE with 95% CI
BMI Gut	0.27 (0.17, 0.34)*	11.8 (-2.1, 41.2)
pH Soil	0.36 (0.28, 0.44)*	0.08 (-0.28, 0.95)

4.6 Discussion

Reproducibility is imperative for any scientific discovery, but there is a growing alarm about irreproducible research results. According to a survery by Nature Publishing Group of 1,576 researchers in 2016, more than 70% of researchers have tried and failed to reproduce another scientist's experiments, and more than half have failed to reproduce their own experiments (Baker, 2016). This "reproducibility crisis" in science affects microbiome research as much as any other areas, and microbiome researchers have long struggled to make their research reproducible (Schloss, 2018). Great efforts have been made towards setting protocols and standards for microbiome data collection and processing (Thompson et al., 2017), but more could be achieved using statistical techniques for reproducible data analysis. Microbiome research findings rely on statistical analysis of high-dimensional data, and feature selection is an indispensable component for discovering biologically relevant microbes. In this article, we focus on discovering a reproducible criterion for evaluating feature selection methods rather than developing a better feature selection method. We question the common practice of evaluating feature selection methods based on

overall performance of model prediction (Knights et al., 2011), such as Mean Squared Error (MSE), as we detect a stark contrast between prediction accuracy and reproducible feature selection. Instead, we propose to use a reproducibility criterion such as Nogueira's Stability measurement (Nogueira et al., 2017) for identifying the optimal feature selection method.

In both our simulations and experimental microbiome data applications, we have shown that Stability is a preferred evaluation criterion over MSE for feature selection, because of its closer reflection of the ground truth (false positive and false negative rates) in simulations, and its better capacity to differentiate methods due to its higher precision. Hence, if the goal is to identify the underlying true biological signal, we propose to use a reproducibility criterion like Stability instead of a prediction criterion like MSE to choose feature selection algorithms for microbiome data applications. MSE is better suited for problems where prediction accuracy alone is the focus.

The strength of our work lies in the comparisons of widely used microbiome feature selection methods using extensive simulations, and experimental microbiome datasets covering various sample types and data characteristics. The comparisons are further confirmed with non-parametric hypothesis testing using bootstrap. Although Nogueira et al. were able to derive the asymptotical normal distribution of Stability (Nogueira et al., 2017), their independent assumption for two-sample test might not be realistic due to the fact that two feature selection methods are applied to the same dataset. Hence our non-parametric hypothesis testing is an extension of their two-sample test for Stability. However, our current usage of bootstrap, especially the nested bootstrap approach for experimental microbiome data applications, is computationally expensive; further theoretical development on hypothesis testing for reproducibility can be done to facilitate more efficient method comparisons based on Stability. Last but not least, although our paper is focused on microbiome data, we do expect the superiority of reproducibility criteria over prediction accuracy criteria in feature selection to apply in other types of datasets as well. We thus recommend that researchers use stability as an evaluation criterion while performing feature selection in order to yield reproducible results.

4.7 Acknowledgements

Chapter 4, in full, has been submitted for publication and is presented as it may appear in “Jiang, L.; Haiminen, N.; Carrieri, A.; Huang, S.; Vázquez-Baeza, Y.; Parida, L.; Kim, H.; Swafford, A.D.; Knight, R.; Natarajan, L. *Utilizing Stability Criteria in Choosing Feature Selection Methods Yields Reproducible Results in Microbiome Data.* Biometrics." The dissertation author was the primary investigator and author of this work.

We gratefully acknowledge supports from IBM Research through the AI Horizons Network, and UC San Diego AI for Healthy Living program in partnership with the UC San Diego Center for Microbiome Innovation. This work was also supported in part by CRISP, one of six centers in JUMP, a Semiconductor Research Corporation (SRC) program sponsored by DARPA. LN was partially supported by NIDDK 1R01DK110541-01A1.

4.8 Supplementary Materials

The code that implements the methodology, simulations and experimental microbiome data applications is available at the Github repository https://github.com/knightlab-analyses/stability-analyses. Also, three supplementary figures for simulation results are provided.

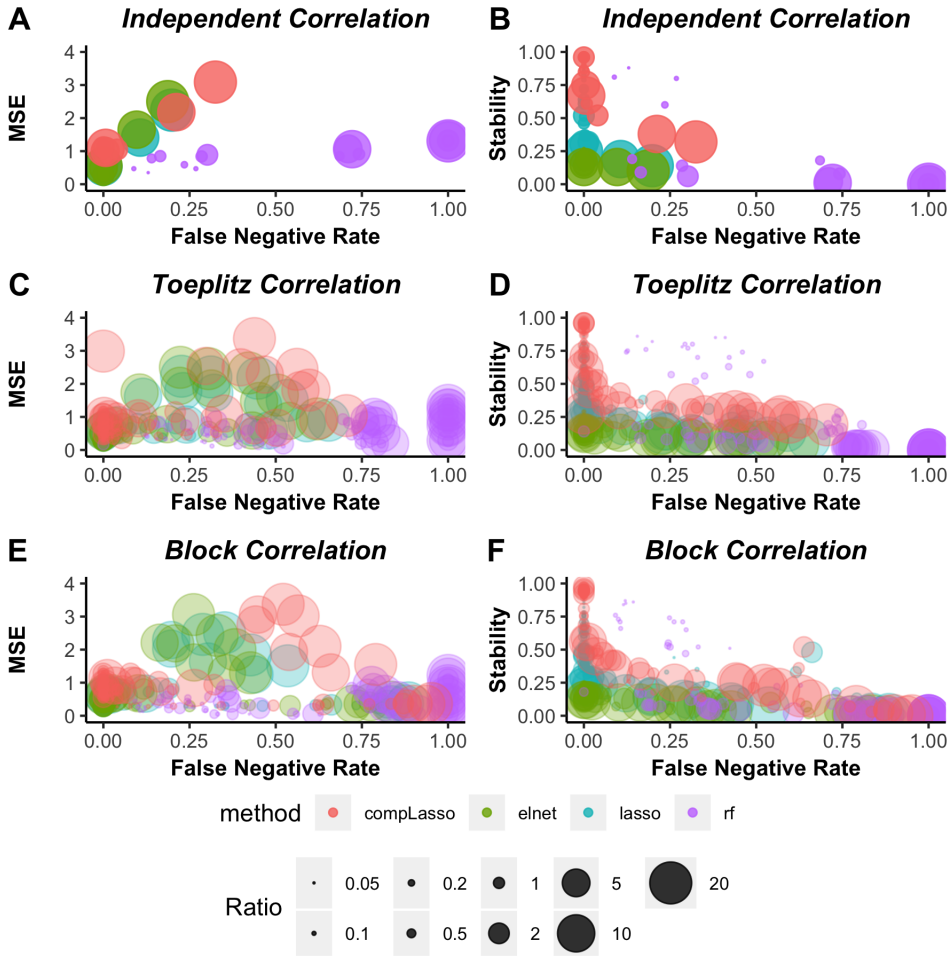

Figure 4.4: Compare the relationship between MSE and False Negative Rate vs. Stability and False Negative Rate in three correlation structures.

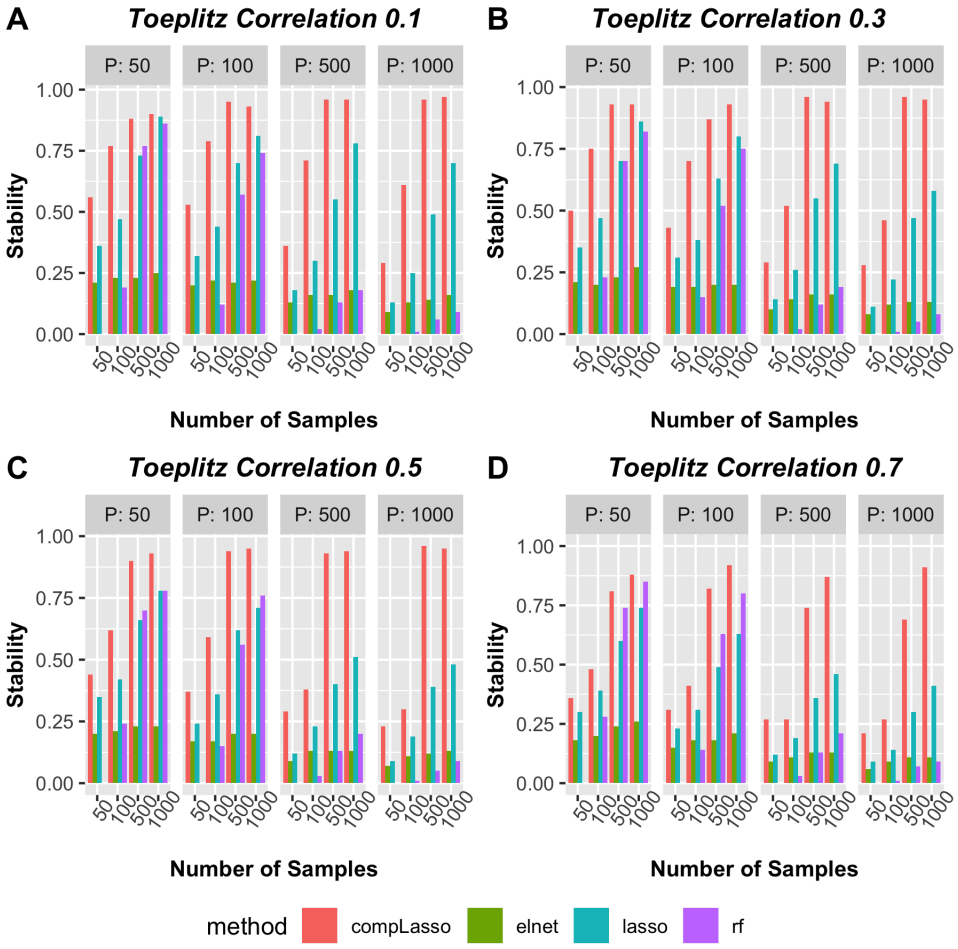

Figure 4.5: Method comparisons based on Stability in easier Toeplitz correlation structures.

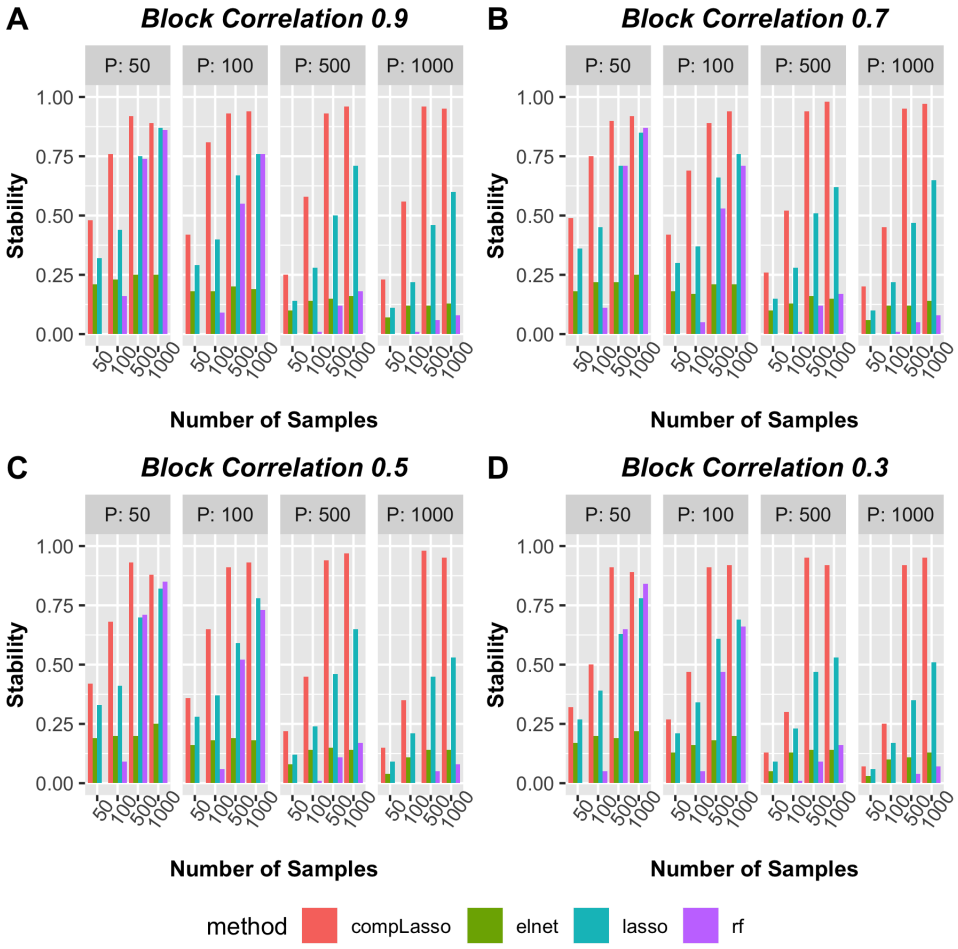

Figure 4.6: Method comparisons based on Stability in easier Block correlation structures.

Chapter 5

Conclusions and Future Work

This dissertation presents novel longitudinal methods and reproducibility criterion for feature selection in microbiome data analysis. Our methodological innovation not only brings in statistical novelty, but also provide down-to-earth methods to perform automatic model selection and model diagnostics, thus empowering users to check their model validity and detect potential gains or needs for model adjustment by using our developed methods. Therefore, we expect our work to be not only interesting to the statistical audience, but also highly applicable and valuable to the microbiome community.

In Chapter 2 we introduce Bayesian sparse functional principal components analysis (SFPCA), a Bayesian extension to the existing SFPCA methodology, which allows for efficient model selection, via leave-one-out cross-validation (LOO) with Pareto-smoothed importance sampling (PSIS), and visual model diagnostics tools such as PSIS-LOO diagnostic plots and graphical posterior predictive checks. Our Bayesian approach not only enables users to perform careful examination of their SFPCA applications to real microbiome data analysis, but also provide flexible modeling for users to incorporate alternative prior distributions to accommodate their specific needs. Moreover, our method is feasible to be extended to model multiple temporal measurements simultaneously, which was successfully developed in Chapter 3. Despite all these merits, limitations do exist with our current approach, which in fact suggests new directions for future work. The first limitation is that our model does not directly incorporate confounding factors into the model, but rather does a post-hoc analysis by investigating the relationship between

estimated FPC scores with hypothesized covariates. Given our flexible Bayesian approach, it is highly likely to incorporate these covariates into the model directly and then adjust for these covariates if they are confounders, or perform a two-sample testing for comparing the mean functions and FPC scores of covariates of interest if the group effect is to be investigated. The second limitation lies in the relative computational inefficiency of Bayesian modeling compared to frequentist approaches. Despite the fact that users might be willing to sacrifice efficiency in terms of valid modeling with our Bayesian approach, we still believe that more efficient computation should be an important goal, and this could be possibly achieved by utilizing the GPU-based parallel computation support in Stan (Češnovar et al., 2019).

In Chapter 3 we present multivariate SFPCA, which extends our earlier Bayesian SFPCA method to model multiple temporal measurements simultaneously, and infer the temporal associations among interested measurements based on estimation of mutual information. Our greatest methodological novelty lies in efficient covariance matrix estimation, where we utilize Cholesky decomposition to estimate covariance matrix efficiently and guarantee it to be positive semi-definite under the constrained form of covariance structure. This approach is not only more computationally efficient, but also achieves higher estimation accuracy than the existing methods. Moreover, we utilized the concept of mutual information to define the marginal and conditional temporal associations, which provides a meaningful and interpretable measure to quantify temporal associations. Last but not least, our Bayesian implementation in Stan enables the usage of PSIS-LOO for efficient model selection, and visual model diagnostic methods to evaluate the validity of mSFPCA models and highlight potential outliers. Although theoretically there are no limitations on the number of temporal measurements our multivariate SFPCA can model simultaneously, computational cost does grow exponentially with the increasing number of temporal measurements and sample size. The computational efficiency is unlikely to be easily improved as in the case of SFPCA by GPU acceleration. But this is possible to be achieved by migrating our code from R and Stan into Julia, a new programming language for data and analytics that combines the functionality of quantitative environments such as R and Python with the speed of production programming languages like Java and C++. Since Julia provides parallel and distributed computing capabilities, and literally infinite scalability, we do expect to overcome the computational hurdle with our Julia collaborators.

In Chapter 4 we focus on discovering a reproducible criterion for evaluating feature selection

methods rather than developing a better feature selection method. We question the common practice of evaluating feature selection methods based on overall performance of model prediction, such as Mean Squared Error (MSE), and propose to use a reproducibility criterion such as Stability for identifying the optimal feature selection method. In both simulations and microbiome applications, we find that Stability is a preferred feature selection criterion over MSE because it better quantifies the reproducibility of the feature selection method. The strength of our work lies in the comparisons of widely used microbiome feature selection methods using extensive simulations, and experimental microbiome datasets covering various sample types and data characteristics. The comparisons are further confirmed with non-parametric hypothesis testing using bootstrap. This non-parametric hypothesis testing is a novelty of our work, yet also a limitation, as our current usage of bootstrap, especially the nested bootstrap approach for experimental microbiome data applications, is computationally expensive; further theoretical development on hypothesis testing for reproducibility can be done to facilitate more efficient method comparisons based on Stability. Further work would also be needed to explore the additional biological insights brought by selected features using best method found by Stability instead of MSE. Last but not least, although our work is focused on microbiome data, we do expect the superiority of reproducibility criteria over prediction accuracy criteria in feature selection to apply in other types of datasets as well, and we thus recommend researchers to use stability as evaluation criterion while performing feature selection in order to yield reproducible results.

In conclusion, our novel longitudinal methods and reproducibility criterion for feature selection provide the state-of-art solutions for microbiome researchers to detect biologically meaningful signals in the complex microbiome data, while at the same time enable them to validate their applications with graphical diagnostics tools. Moreover, we believe that the general frameworks presented in our three methods can also be applied to address challenges in other types of datasets, beyond the scope of microbiome data analysis.

Bibliography

Ahmed, N. A. and Gokhale, D. (1989). Entropy expressions and their estimators for multivariate distributions. *IEEE Transactions on Information Theory* **35,** 688–692.

Aitchison, J. (1982). The statistical analysis of compositional data. *Journal of the Royal Statistical Society: Series B (Methodological)* **44,** 139–160.

Altmann, A., Toloşi, L., Sander, O., and Lengauer, T. (2010). Permutation importance: a corrected feature importance measure. *Bioinformatics* **26,** 1340–1347.

Arellano-Valle, R. B., Contreras-Reyes, J. E., and Genton, M. G. (2013). Shannon entropy and mutual information for multivariate skew-elliptical distributions. *Scandinavian Journal of Statistics* **40,** 42–62.

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility.

Ballen, K., Ahn, K. W., Chen, M., Abdel-Azim, H., Ahmed, I., Aljurf, M., Antin, J., Bhatt, A. S., Boeckh, M., Chen, G., et al. (2016). Infection rates among acute leukemia patients receiving alternative donor hematopoietic cell transplantation. *Biology of Blood and Marrow Transplantation* **22,** 1636–1645.

Barnard, J., McCulloch, R., and Meng, X.-L. (2000). Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. *Statistica Sinica* pages 1281–1311.

Belk, A., Xu, Z. Z., Carter, D. O., Lynne, A., Bucheli, S., Knight, R., and Metcalf, J. L. (2018). Microbiome data accurately predicts the postmortem interval using random forest regression models. *Genes* **9,** 104.

Bodein, A., Chapleur, O., Droit, A., and Lê Cao, K.-A. (2019). A generic multivariate framework for the integration of microbiome longitudinal studies with other data types. *Frontiers in genetics* **10,**.

Boulesteix, A.-L. and Slawski, M. (2009). Stability and aggregation of ranked gene lists. *Briefings in bioinformatics* **10,** 556–568.

Bouslimani, A., da Silva, R., Kosciolek, T., Janssen, S., Callewaert, C., Amir, A., Dorrestein, K., Melnik, A. V., Zaramela, L. S., Kim, J.-N., et al. (2019). The impact of skin care products on skin chemistry and microbiome dynamics. *BMC biology* **17,** 1–20.

Breiman, L. (2001). Random forests. *Machine learning* **45,** 5–32.

Caporaso, J. G., Lauber, C. L., Costello, E. K., Berg-Lyons, D., Gonzalez, A., Stombaugh, J., Knights, D., Gajer, P., Ravel, J., Fierer, N., et al. (2011). Moving pictures of the human microbiome. *Genome biology* **12,** 1–8.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of statistical software* **76,**.

Češnovar, R., Bronder, S., Sluga, D., Demšar, J., Ciglarič, T., Talts, S., and Štrumbelj, E. (2019). Gpu-based parallel computation support for stan. *arXiv preprint arXiv:1907.01063* .

Chiou, J.-M., Chen, Y.-T., and Yang, Y.-F. (2014). Multivariate functional principal component analysis: A normalization approach. *Statistica Sinica* pages 1571–1596.

Clemente, J. C., Ursell, L. K., Parfrey, L. W., and Knight, R. (2012). The impact of the gut microbiota on human health: an integrative view. *Cell* **148,** 1258–1270.

Cover, T. M. (1999). *Elements of information theory*. John Wiley & Sons.

David, L. A., Maurice, C. F., Carmody, R. N., Gootenberg, D. B., Button, J. E., Wolfe, B. E., Ling, A. V., Devlin, A. S., Varma, Y., Fischbach, M. A., et al. (2014). Diet rapidly and reproducibly alters the human gut microbiome. *Nature* **505,** 559–563.

Dethlefsen, L. and Relman, D. A. (2011). Incomplete recovery and individualized responses of the human distal gut microbiota to repeated antibiotic perturbation. *Proceedings of the National Academy of Sciences* **108,** 4554–4561.

Di, C., Crainiceanu, C. M., and Jank, W. S. (2014). Multilevel sparse functional principal component analysis. *Stat* **3,** 126–143.

Di, C.-Z., Crainiceanu, C. M., Caffo, B. S., and Punjabi, N. M. (2009). Multilevel functional principal component analysis. *The annals of applied statistics* **3,** 458.

Ding, T. and Schloss, P. D. (2014). Dynamics and associations of microbial community types across the human body. *Nature* **509,** 357–360.

Dominguez-Bello, M. G., Blaser, M. J., Ley, R. E., and Knight, R. (2011). Development of the human gastrointestinal microbiota and insights from high-throughput sequencing. *Gastroenterology* **140,** 1713–1719.

Dominguez-Bello, M. G., Costello, E. K., Contreras, M., Magris, M., Hidalgo, G., Fierer, N., and Knight, R. (2010). Delivery mode shapes the acquisition and structure of the initial microbiota across multiple body habitats in newborns. *Proceedings of the National Academy of Sciences* **107,** 11971–11975.

Dominguez-Bello, M. G., De Jesus-Laboy, K. M., Shen, N., Cox, L. M., Amir, A., Gonzalez, A., Bokulich, N. A., Song, S. J., Hoashi, M., Rivera-Vinas, J. I., et al. (2016). Partial restoration of the microbiota of cesarean-born infants via vaginal microbial transfer. *Nature medicine* **22,** 250.

Dominguez-Bello, M. G., Godoy-Vitorino, F., Knight, R., and Blaser, M. J. (2019). Role of the microbiome in human development. *Gut* **68,** 1108–1114.

Duvallet, C., Gibbons, S. M., Gurry, T., Irizarry, R. A., and Alm, E. J. (2017). Meta-analysis of gut microbiome studies identifies disease-specific and shared responses. *Nature communications* **8,** 1–10.

Efron, B. and Tibshirani, R. J. (1994). *An introduction to the bootstrap*. CRC press.

Fierer, N., Hamady, M., Lauber, C. L., and Knight, R. (2008). The influence of sex, handedness, and washing on the diversity of hand surface bacteria. *Proceedings of the National Academy of Sciences* **105,** 17994–17999.

Flores, G. E., Caporaso, J. G., Henley, J. B., Rideout, J. R., Domogala, D., Chase, J., Leff, J. W., Vázquez-Baeza, Y., Gonzalez, A., Knight, R., et al. (2014). Temporal variability is a personalized feature of the human microbiome. *Genome biology* **15,** 531.

Frati, F., Salvatori, C., Incorvaia, C., Bellucci, A., Di Cara, G., Marcucci, F., and Esposito, S. (2019). The role of the microbiome in asthma: The gut–lung axis. *International Journal of Molecular Sciences* **20,** 123.

Gabry, J., Simpson, D., Vehtari, A., Betancourt, M., and Gelman, A. (2019). Visualization in bayesian workflow. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **182,** 389–402.

Gelfand, A. E., Dey, D. K., and Chang, H. (1992). Model determination using predictive distributions with implementation via sampling-based methods. Technical report, STANFORD UNIV CA DEPT OF STATISTICS.

Gelman, A., Meng, X.-L., and Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica sinica* pages 733–760.

Gentle, J. E. (2012). *Numerical linear algebra for applications in statistics*. Springer Science & Business Media.

Gibson, T. E. and Gerber, G. K. (2018). Robust and scalable models of microbiome dynamics. *arXiv preprint arXiv:1805.04591* .

Gill, S. R., Pop, M., DeBoy, R. T., Eckburg, P. B., Turnbaugh, P. J., Samuel, B. S., Gordon, J. I., Relman, D. A., Fraser-Liggett, C. M., and Nelson, K. E. (2006). Metagenomic analysis of the human distal gut microbiome. *science* **312,** 1355–1359.

Goh, W. W. B. and Wong, L. (2016). Evaluating feature-selection stability in next-generation proteomics. *Journal of bioinformatics and computational biology* **14,** 1650029.

Halfvarson, J., Brislawn, C. J., Lamendella, R., Vázquez-Baeza, Y., Walters, W. A., Bramer, L. M., D'amato, M., Bonfiglio, F., McDonald, D., Gonzalez, A., et al. (2017). Dynamics of the human gut microbiome in inflammatory bowel disease. *Nature microbiology* **2,** 1–7.

Hall, P. and Hosseini-Nasab, M. (2006). On properties of functional principal components analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **68,** 109–126.

Holleran, G., Scaldaferri, F., Ianiro, G., Lopetuso, L., Mc, D. N., Mele, M., Gasbarrini, A., and Cammarota, G. (2018). Fecal microbiota transplantation for the treatment of patients with ulcerative colitis and other gastrointestinal conditions beyond clostridium difficile infection: an update. *Drugs of today (Barcelona, Spain: 1998)* **54,** 123–136.

Huttenhower, C., Gevers, D., Knight, R., Abubucker, S., Badger, J. H., Chinwalla, A. T., Creasy, H. H., Earl, A. M., FitzGerald, M. G., Fulton, R. S., et al. (2012). Structure, function and diversity of the healthy human microbiome. *nature* **486,** 207.

iHMP Consortium (2014). The integrative human microbiome project: dynamic analysis of microbiome-host omics profiles during periods of human health and disease. *Cell host & microbe* **16,** 276.

James, G. M., Hastie, T. J., and Sugar, C. A. (2000). Principal component models for sparse functional data. *Biometrika* **87,** 587–602.

Jiang, D., Armour, C. R., Hu, C., Mei, M., Tian, C., Sharpton, T. J., and Jiang, Y. (2019). Microbiome multi-omics network analysis: statistical considerations, limitations, and opportunities. *Frontiers in genetics* **10,**.

Jiang, L., Zhong, Y., Elrod, C., Natarajan, L., Knight, R., and Thompson, W. K. (2020). Bayestime: Bayesian functional principal components for sparse longitudinal data. *Journal of Computational and Graphical Statistics* page submitted.

Joe, H. (1989). Relative entropy measures of multivariate dependence. *Journal of the American Statistical Association* **84,** 157–164.

Kalousis, A., Prados, J., and Hilario, M. (2005). Stability of feature selection algorithms. In *Fifth IEEE International Conference on Data Mining (ICDM'05)*, pages 8–pp. IEEE.

Kalousis, A., Prados, J., and Hilario, M. (2007). Stability of feature selection algorithms: a study on high-dimensional spaces. *Knowledge and information systems* **12,** 95–116.

Kidziński, Ł. and Hastie, T. (2018). Longitudinal data analysis using matrix completion. *arXiv preprint arXiv:1809.08771* .

Knights, D., Costello, E. K., and Knight, R. (2011). Supervised classification of human microbiota. *FEMS microbiology reviews* **35,** 343–359.

Knights, D., Parfrey, L. W., Zaneveld, J., Lozupone, C., and Knight, R. (2011). Human-associated microbial signatures: examining their predictive value. *Cell host & microbe* **10,** 292–296.

Koenig, J. E., Spor, A., Scalfone, N., Fricker, A. D., Stombaugh, J., Knight, R., Angenent, L. T., and Ley, R. E. (2011). Succession of microbial consortia in the developing infant gut microbiome. *Proceedings of the National Academy of Sciences* **108,** 4578–4585.

Kostic, A. D., Gevers, D., Siljander, H., Vatanen, T., Hyötyläinen, T., Hämäläinen, A.-M., Peet, A., Tillmann, V., Pöhö, P., Mattila, I., et al. (2015). The dynamics of the human infant gut microbiome in development and in progression toward type 1 diabetes. *Cell host & microbe* **17,** 260–273.

Kuczynski, J., Costello, E. K., Nemergut, D. R., Zaneveld, J., Lauber, C. L., Knights, D., Koren, O., Fierer, N., Kelley, S. T., Ley, R. E., et al. (2010). Direct sequencing of the human microbiome readily reveals community differences. *Genome biology* **11,** 210.

Kurtz, Z. D., Müller, C. L., Miraldi, E. R., Littman, D. R., Blaser, M. J., and Bonneau, R. A. (2015). Sparse and compositionally robust inference of microbial ecological networks. *PLoS Comput Biol* **11,** e1004226.

Lauber, C. L., Hamady, M., Knight, R., and Fierer, N. (2009). Pyrosequencing-based assessment of soil ph as a predictor of soil bacterial community structure at the continental scale. *Applied and environmental microbiology* **75,** 5111–5120.

Lee, H. W., Lawton, C., Na, Y. J., and Yoon, S. (2013). Robustness of chemometrics-based feature selection methods in early cancer detection and biomarker discovery. *Statistical applications in genetics and molecular biology* **12,** 207–223.

Li, H. (2015). Microbiome, metagenomics, and high-dimensional compositional data analysis. *Annual Review of Statistics and Its Application* **2,** 73–94.

Li, Y., Hsing, T., et al. (2010). Uniform convergence rates for nonparametric regression and principal component analysis in functional/longitudinal data. *The Annals of Statistics* **38,** 3321–3351.

Liechty, J. C., Liechty, M. W., and Müller, P. (2004). Bayesian correlation estimation. *Biometrika* **91,** 1–14.

Lin, W., Shi, P., Feng, R., and Li, H. (2014). Variable selection in regression with compositional covariates. *Biometrika* **101,** 785–797.

Liu, Y., Tang, S., Fernandez-Lozano, C., Munteanu, C. R., Pazos, A., Yu, Y.-z., Tan, Z., and González-Díaz, H. (2017). Experimental study and random forest prediction model of microbiome cell surface hydrophobicity. *Expert Systems with Applications* **72,** 306–316.

Lloyd-Price, J., Arze, C., Ananthakrishnan, A. N., Schirmer, M., Avila-Pacheco, J., Poon, T. W., Andrews, E., Ajami, N. J., Bonham, K. S., Brislawn, C. J., et al. (2019). Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature* **569,** 655–662.

McDonald, D., Hyde, E., Debelius, J. W., Morton, J. T., Gonzalez, A., Ackermann, G., Aksenov, A. A., Behsaz, B., Brennan, C., Chen, Y., et al. (2018). American gut: an open platform for citizen science microbiome research. *Msystems* **3,** e00031–18.

Methé, B. A., Nelson, K. E., Pop, M., Creasy, H. H., Giglio, M. G., Huttenhower, C., Gevers, D., Petrosino, J. F., Abubucker, S., Badger, J. H., et al. (2012). A framework for human microbiome research. *nature* **486,** 215.

Morton, J. T., Aksenov, A. A., Nothias, L. F., Foulds, J. R., Quinn, R. A., Badri, M. H., Swenson, T. L., Van Goethem, M. W., Northen, T. R., Vazquez-Baeza, Y., et al. (2019). Learning representations of microbe–metabolite interactions. *Nature methods* **16,** 1306–1314.

Morton, J. T., Sanders, J., Quinn, R. A., McDonald, D., Gonzalez, A., Vázquez-Baeza, Y., Navas-Molina, J. A., Song, S. J., Metcalf, J. L., Hyde, E. R., et al. (2017). Balance trees reveal microbial niche differentiation. *MSystems* **2,**.

Namkung, J. (2020). Machine learning methods for microbiome studies. *Journal of Microbiology* **58,** 206–216.

Nash, J. (1990). The cholesky decomposition. *Compact numerical methods for computers: Linear algebra and function minimisation* **2,**.

Nogueira, S. (2018). *Quantifying the stability of feature selection*. The University of Manchester (United Kingdom).

Nogueira, S., Sechidis, K., and Brown, G. (2017). On the stability of feature selection algorithms. *The Journal of Machine Learning Research* **18,** 6345–6398.

O'Keefe, S. J., Li, J. V., Lahti, L., Ou, J., Carbonero, F., Mohammed, K., Posma, J. M., Kinross, J., Wahl, E., Ruder, E., et al. (2015). Fat, fibre and cancer risk in african americans and rural africans. *Nature communications* **6,** 1–14.

Palmer, C., Bik, E. M., DiGiulio, D. B., Relman, D. A., and Brown, P. O. (2007). Development of the human infant intestinal microbiota. *PLoS biol* **5,** e177.

Pannaraj, P. S., Li, F., Cerini, C., Bender, J. M., Yang, S., Rollie, A., Adisetiyo, H., Zabih, S., Lincez, P. J., Bittinger, K., et al. (2017). Association between breast milk bacterial communities and establishment and development of the infant gut microbiome. *JAMA pediatrics* **171,** 647–654.

Peng, J. and Paul, D. (2009). A geometric approach to maximum likelihood estimation of the functional principal components from sparse longitudinal data. *Journal of Computational and Graphical Statistics* **18,** 995–1015.

Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K. S., Manichanh, C., Nielsen, T., Pons, N., Levenez, F., Yamada, T., et al. (2010). A human gut microbial gene catalogue established by metagenomic sequencing. *nature* **464,** 59–65.

Ramsay, J. and Silverman, B. W. (1997). Functional data analysis (springer series in statistics).

Ramsay, J. O. and Silverman, B. W. (2007). *Applied functional data analysis: methods and case studies*. Springer.

Ranjan, R., Rani, A., Metwally, A., McGee, H. S., and Perkins, D. L. (2016). Analysis of the microbiome: Advantages of whole genome shotgun versus 16s amplicon sequencing. *Biochemical and biophysical research communications* **469,** 967–977.

Rice, J. A. and Silverman, B. W. (1991). Estimating the mean and covariance structure nonparametrically when the data are curves. *Journal of the Royal Statistical Society: Series B (Methodological)* **53,** 233–243.

Ridenhour, B. J., Brooker, S. L., Williams, J. E., Van Leuven, J. T., Miller, A. W., Dearing, M. D., and Remien, C. H. (2017). Modeling time-series data from microbial communities. *The ISME journal* **11,** 2526–2537.

Santo, D., Loncar-Turukalo, T., Stres, B., Crnojevic, V., and Brdar, S. (2019). Clustering and classification of human microbiome data: Evaluating the impact of different settings in bioinformatics workflows. In *2019 IEEE 19th International Conference on Bioinformatics and Bioengineering (BIBE)*, pages 838–845. IEEE.

Sberro, H., Fremin, B. J., Zlitni, S., Edfors, F., Greenfield, N., Snyder, M. P., Pavlopoulos, G. A., Kyrpides, N. C., and Bhatt, A. S. (2019). Large-scale analyses of human microbiomes reveal thousands of small, novel genes. *Cell* **178,** 1245–1259.

Schloss, P. D. (2018). Identifying and overcoming threats to reproducibility, replicability, robustness, and generalizability in microbiome research. *MBio* **9,**.

Sender, R., Fuchs, S., and Milo, R. (2016). Revised estimates for the number of human and bacteria cells in the body. *PLoS biology* **14,** e1002533.

Sharon, G., Cruz, N. J., Kang, D.-W., Gandal, M. J., Wang, B., Kim, Y.-M., Zink, E. M., Casey, C. P., Taylor, B. C., Lane, C. J., et al. (2019). Human gut microbiota from autism spectrum disorder promote behavioral symptoms in mice. *Cell* **177,** 1600–1618.

Shendure, J. and Ji, H. (2008). Next-generation dna sequencing. *Nature biotechnology* **26,** 1135–1145.

Shenhav, L., Furman, O., Briscoe, L., Thompson, M., Silverman, J. D., Mizrahi, I., and Halperin, E. (2019). Modeling the temporal dynamics of the gut microbial community in adults and infants. *PLoS computational biology* **15,** e1006960.

Shi, P., Zhang, A., Li, H., et al. (2016). Regression analysis for microbiome compositional data. *The Annals of Applied Statistics* **10,** 1019–1040.

Silverman, J. D., Durand, H. K., Bloom, R. J., Mukherjee, S., and David, L. A. (2018). Dynamic linear models guide design and analysis of microbiota studies within artificial human guts. *Microbiome* **6,** 1–20.

Silverman, J. D., Roche, K., Holmes, Z. C., David, L. A., and Mukherjee, S. (2019). Bayesian multinomial logistic normal models through marginally latent matrix-t processes. *arXiv preprint arXiv:1903.11695* .

Silverman, J. D., Washburne, A. D., Mukherjee, S., and David, L. A. (2017). A phylogenetic transform enhances analysis of compositional microbiota data. *Elife* **6,** e21887.

Smarr, L., Hyde, E. R., McDonald, D., Sandborn, W. J., and Knight, R. (2017). Tracking human gut microbiome changes resulting from a colonoscopy. *Methods of information in medicine* **56,** 442–447.

Sommer, M. O., Dantas, G., and Church, G. M. (2009). Functional characterization of the antibiotic resistance reservoir in the human microflora. *science* **325,** 1128–1131.

Statnikov, A., Henaff, M., Narendra, V., Konganti, K., Li, Z., Yang, L., Pei, Z., Blaser, M. J., Aliferis, C. F., and Alekseyenko, A. V. (2013). A comprehensive evaluation of multicategory classification methods for microbiomic data. *Microbiome* **1,** 11.

Stewart, C. J., Ajami, N. J., O'Brien, J. L., Hutchinson, D. S., Smith, D. P., Wong, M. C., Ross, M. C., Lloyd, R. E., Doddapaneni, H., Metcalf, G. A., et al. (2018). Temporal development of the gut microbiome in early childhood from the teddy study. *Nature* **562,** 583–588.

Sze, M. A. and Schloss, P. D. (2016). Looking for a signal in the noise: revisiting obesity and the microbiome. *MBio* **7,** e01018–16.

Thompson, L. R., Sanders, J. G., McDonald, D., Amir, A., Ladau, J., Locey, K. J., Prill, R. J., Tripathi, A., Gibbons, S. M., Ackermann, G., et al. (2017). A communal catalogue reveals earth's multiscale microbial diversity. *Nature* **551,** 457–463.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* **58,** 267–288.

Turnbaugh, P. J., Ley, R. E., Hamady, M., Fraser-Liggett, C. M., Knight, R., and Gordon, J. I. (2007). The human microbiome project. *Nature* **449,** 804–810.

Turnbaugh, P. J., Ridaura, V. K., Faith, J. J., Rey, F. E., Knight, R., and Gordon, J. I. (2009). The effect of diet on the human gut microbiome: a metagenomic analysis in humanized gnotobiotic mice. *Science translational medicine* **1,** 6ra14–6ra14.

Vatanen, T., Franzosa, E. A., Schwager, R., Tripathi, S., Arthur, T. D., Vehik, K., Lernmark, Å., Hagopian, W. A., Rewers, M. J., She, J.-X., et al. (2018). The human gut microbiome in early-onset type 1 diabetes from the teddy study. *Nature* **562,** 589–594.

Vehtari, A., Gelman, A., and Gabry, J. (2017). Practical bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and computing* **27,** 1413–1432.

Vehtari, A., Simpson, D., Gelman, A., Yao, Y., and Gabry, J. (2015). Pareto smoothed importance sampling. *arXiv preprint arXiv:1507.02646* .

Wainer, J. and Cawley, G. (2018). Nested cross-validation when selecting classifiers is overzealous for most practical applications. *arXiv preprint arXiv:1809.09446* .

Weingarden, A., González, A., Vázquez-Baeza, Y., Weiss, S., Humphry, G., Berg-Lyons, D., Knights, D., Unno, T., Bobr, A., Kang, J., et al. (2015). Dynamic changes in short-and long-term bacterial composition following fecal microbiota transplantation for recurrent clostridium difficile infection. *Microbiome* **3,** 1–8.

Whitman, W. B., Coleman, D. C., and Wiebe, W. J. (1998). Prokaryotes: the unseen majority. *Proceedings of the National Academy of Sciences* **95,** 6578–6583.

Wu, G. D., Chen, J., Hoffmann, C., Bittinger, K., Chen, Y.-Y., Keilbaugh, S. A., Bewtra, M., Knights, D., Walters, W. A., Knight, R., et al. (2011). Linking long-term dietary patterns with gut microbial enterotypes. *Science* **334,** 105–108.

Yao, F., Müller, H.-G., and Wang, J.-L. (2005). Functional data analysis for sparse longitudinal data. *Journal of the American statistical association* **100,** 577–590.

Yatsunenko, T., Rey, F. E., Manary, M. J., Trehan, I., Dominguez-Bello, M. G., Contreras, M., Magris, M., Hidalgo, G., Baldassano, R. N., Anokhin, A. P., et al. (2012). Human gut microbiome viewed across age and geography. *nature* **486,** 222–227.

Zhang, J. and Stephens, M. A. (2009). A new and efficient estimation method for the generalized pareto distribution. *Technometrics* **51,** 316–325.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)* **67,** 301–320.