UNIVERSITY OF CALIFORNIA SAN DIEGO

Semiparametric Regression Models for Between- and Within-subject Attributes: Applications to High-Dimensional Data, Asymptotic Efficiency and Beyond

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

in

Biostatistics

by

Jinyuan Liu

Committee in charge:

Professor Xin Tu, Chair
Professor Lin Liu
Professor Loki Natarajan
Professor Tanya T Nguyen
Professor Bernd Schnabl
Professor Xinlian Zhang

2022

The Dissertation of Jinyuan Liu is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2022

EPIGRAPH

Clear thinking becomes clear writing;

one can't exist without the other.

William Zinsser

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

Standing at this point, there are so many things to be grateful for. As people always say, life is short, you live just once, carpe diem. I am glad that I seized the moment and am approaching the finishing line of the five years commitment.

First, I would like to express my deepest gratitude to my parents for their selfless support and love, being raised in such a considerate environment is my luck. I appreciate that they respect every decision I made, cheer me on at each step, and encourage me to never give up. Their kindness and patience have empowered me to chase my goals and passion. I love them so much!

Next, I want to acknowledge my advisor, also the chair of my committee, Prof. Tu, for his remarkable mentoring. He genuinely cares about us students, not only academically but also in our life. Through working together with him for over six years, I have learned a lot and will keep learning more from him. I still remember our very first conversation when I expressed my interest in working with him on a project on the social network, it leads me to the world of conducting research! Rooted in the strong theoretical ground, Prof. Tu is always so successful in generating innovative insights and solving real hassles for collaborators effectively. His life-long learning attitude and persistence always inspire me to keep challenging myself. I am sincerely grateful to have such a great role model!

I also want to thank Xinlian for her generous help and support! After working with her for more than two years, I am still impressed by her sharp ideas in identifying problems and solving them efficiently. As someone who also just went through the faculty search, I am so grateful to receive her endless support and encouragement during my own process. This dissertation and my research become more concrete with her great insights into both the scientific and mathematical sides!

Next, I would like to express my gratitude to my thesis committee members for their substantial support in forming this dissertation. As my academic advisor, Dr. Liu provides more

than just academic guidance but also moral support for all the small achievements I made along the way. I also greatly appreciate Dr. Natarajan for her innate compassion and kindness for us students, which made our program more like a real family. I really enjoy working with Tanya, her working attitude and scientific understanding of microbiome always motivate me to keep going. I also want to thank Dr. Schnabl for his support. It was such a great experience to work with him and his amazing groups, such as Sonja, Yi, and Huikuan.

I would like to acknowledge Dr. De Gruttola for his remarkable support of my faculty search. As I always said, how lucky I am to be able to work with and learn from all those great minds! His perseverance and efforts to step in the investigators' shoes guide me to be a more solid biostatistician.

Working as a GSR in the Center for Healthy Aging with Dr. Jeste's group for almost five years, I was granted the opportunity to work closely with outstanding investigators such as Ellen, Tanya, Rebecca, etc. I not only improved the essential skills to be a biostatistician but was also strongly motivated by their working attitude and productivity. I also want to thank Dr. Jeste and Paula for their help and support during my faculty search!

I would like to acknowledge everyone in our division and program. As graduate students, we received endless support from the faculty such as Dr. Messer, Dr. Vaida, etc., as well as the help and encouragement from our program coordinators Sarah and Stella. As the second cohort, it is a pleasure to grow together with our program. It feels more like a big family with its rapid growth recently!

I also want to acknowledge my peers, especially those that I worked closely with. I would like to thank Tuo for his help in checking the theoretical ground in many parts of my research. I really enjoy our weekly seminars discussing topics such as semiparametric efficiency. The discussions can be intense sometimes, but we all benefit from such great learning experiences. It has already become something that I am looking forward to every week! I also want to thank Morris for his help in facilitating the code. I am always impressed by his

organized manner in many aspects! I would like to thank Kathy for her help in assembling the R package so efficiently. I also enjoy those conversations with Brian, Senere, Matt, Ruifeng, etc. I really miss those moments when we could meet in person before the pandemic!

I also want to thank those who showed up in my life: Dan, Zhe, Lin, Antai, Chen, Qinxin, Yang, Shiyuan, Henry, Lu, etc., for encouraging me during difficult times! I am genuinely grateful for being surrounded by the cheerful people who support my passions and enthusiasm!

Time flies, five years just passed so quickly, and I will undertake new challenges in the next phase of my life journey. I would like to acknowledge all the help and encouragement I received during my job interviews, such as from Dr. Vandekar, Dr. Schildcrout, Dr. Breheny, Dr. Sun, Dr. Genton, Dr. Ombao, Dr. Goff, etc. They are all great senior faculty but empower our junior researchers to think out of the box, defend our research, and claim our ground!

I want to conclude by thanking myself for not giving up during hard times. Although there is still a long way to go, celebrating tiny or big moments is essential! As people say, do what you love, and you will never have to work another day!

In addition, Chapter 2, in full, is a reprint of the material as it appears in Biometrics 2021. The dissertation author was the primary investigator and author of this paper. The co-authors include Zhang, X, Chen, T, Wu, T, Lin, T, Jiang, L, Lang, S, Liu, L, Natarajan, L, Tu, JX, Kosciolek, T, Morton, J, Nguyen, TT, Schnabl, B, Knight, R, Feng, C, Zhong, Y, and Tu, XM.

Chapter 3, in part is currently being prepared for submission for publication of the material. The dissertation author was the primary investigator and author of this material. The co-authors include Zhang, X., Zhong, Y., Lin, T., Chen, T., Wu, T., Nguyen, T.T, Jeste, D. V. and Tu, XM.

Chapter 4, in part is currently being prepared for submission for publication of the material. The dissertation author was the primary investigator and author of this material. The co-authors include Lin, T., Zhang, X., Chen, T. and Tu, XM.

VITA

2015	Bachelor of Science, Nanjing University of Posts and Telecommunications
2017	Mater of Arts, University of Rochester
2022	Doctor of Philosophy, University of California, San Diego

FIELDS OF STUDY

Biostatistics

PUBLICATIONS

Publications in statistical method

1. **Liu, J**, Zhang, X, Chen, T, Wu, T, Lin, T, Jiang, L, Lang, S, Liu, L, Natarajan, L, Tu, JX, Kosciolek, T, Morton, J, Nguyen, TT, Schnabl, B, Knight, R, Feng, C, Zhong, Y, Tu, XM. (2021). A semiparametric model for between-subject attributes: Applications to beta-diversity of microbiome data. *Biometrics.* 2021; 1– 13. https://doi.org/10.1111/biom.13487.

2. Lin, T, Chen, T, **Liu, J**, Tu, XM. (2021). Extending the Mann-Whitney-Wilcoxon rank sum test to survey data for comparing mean ranks. *Statistics in Medicine.* 2021; 40: 1705–1717. https://doi.org/10.1002/sim.8865.

3. J. Kowalski, S. Hao, T. Chen, Y. Liang, **J. Liu**, L. Ge, C. Feng, X. M. Tu. (2018). Modern variable selection for longitudinal semi-parametric models with missing data. *Journal of Applied Statistics.* 2018; 45:14, 2548-2562.

Publications in statistical application

1. Liu, C., Zhang, X., Nguyen, T.T., **Liu, J.**, Wu, T., Lee, E.E. and Tu, X.M. (2022). Partial least squares regression and principal component analysis: similarity and differences between two popular variable reduction approaches. *General Psychiatry*, 2022;35:e100662.

2. Aslam, S., **Liu, J.**, Sigler, R., Syed, R. R., Tu, X. M., Little, S. J., De Gruttola, V. (2022). COVID-19 vaccination is protective of clinical disease in solid organ transplant recipients. *Transplant Infectious Disease: an official journal of the Transplantation Society,* 10.1111/tid.13788.

3. Michael L. Thomas, Barton W. Palmer, Ellen E. Lee, **Jinyuan Liu**, Rebecca Daly, Xin M. Tu, Dilip V. Jeste (2021). Abbreviated San Diego Wisdom Scale (SD-WISE-7) and Jeste-Thomas Wisdom Index (JTWI). *International Psychogeriatrics*, 1-10.

4. Lee, E.E., Govind, T., Ramsey, M., Wu, T.C., Daly, R., **Liu, J**, Tu, X.M., Paulus, M.P., Thomas, M.L. and Jeste, D.V. (2021). Compassion toward others and self-compassion predict mental and physical well-being: a 5-year longitudinal study of 1090 community-dwelling adults across the lifespan. *Translational Psychiatry* 11(1), pp.1-9.

5. Nguyen, T. T., Zhang, X., Wu, T. C., **Liu, J**, Le, C., Tu, X. M., *et al.* (2021). Association of Loneliness and Wisdom With Gut Microbial Diversity and Composition: An Exploratory Study. *Frontiers in psychiatry*, 12, 395.

6. Jeste, D.V., Thomas, M.L., Liu, J., Daly, R.E., Tu, X.M., Treichler, E.B., Palmer, B.W. and Lee, E.E., (2021). Is spirituality a component of wisdom? Study of 1,786 adults using expanded San Diego Wisdom Scale (Jeste-Thomas Wisdom Index). *Journal of Psychiatric Research*, 132, pp.174-181.

7. Jeste, D.V., Di Somma, S., Lee, E.E., Nguyen, T.T., Scalcione, M., Biaggi, A., Daly, R.,**Liu, J**, Tu, X., Ziedonis, D. and Glorioso, D., (2020). Study of loneliness and wisdom in 482 middle-aged and oldest-old adults: a comparison between people in Cilento, Italy and San Diego, USA. *Aging & Mental Health*, pp.1-11.

8. Lee, E.E., Bangen, K.J., Avanzino, J.A., Hou, B., Ramsey, M., Eglit, G., **Liu, J**, Tu, X.M., Paulus, M. and Jeste, D.V., (2020). Outcomes of randomized clinical trials of interventions to enhance social, emotional, and spiritual components of wisdom: a systematic review and meta-analysis. *JAMA psychiatry*, 77(9), pp.925-935.

9. Lang, S., Martin, A., Farowski, F., Wisplinghoff, H., Vehreschild, M.J., **Liu, J**, Krawczyk, M., Nowag, A., Kretzschmar, A., Herweg, J. and Schnabl, B., (2020). High protein intake is associated with histological disease activity in patients with NAFLD. *Hepatology communications*, 4(5), pp.681-695.

10. Yi Duan, Cristina Llorente, Katharina Brandl, Huikuan Chu, Lu Jiang, Sonja Lang, Manolito Torralba, Yan Shao, **Jinyuan Liu**, *et al.* (2019). Bacteriophage targeting of gut bacterium attenuates alcoholic liver disease. *Nature*, 575(7783), 505-511.

11. Chu H, Duan Y, Lang S, Jiang L, Wang Y, Llorente C, **Liu, J**, *et al.* (2019). The Candida albicans exotoxin Candidalysin promotes alcohol-associated liver disease. *J Hepatol.*, 2019 Oct 10.

12. Lang S, Duan Y, **Liu, J**, Torralba MG, Kuelbs C, et al. (2019). Intestinal Fungal Dysbiosis and Systemic Immune Response to Fungi in Patients With Alcoholic Hepatitis. *Hepatology,* 2019 Jun 22.

13. Lee EE, Sears DD, **Liu, J**, Jin H, Tu XM, Eyler LT, Jeste DV. (2019). A novel biomarker of cardiometabolic pathology in schizophrenia? *J Psychiatr Res.*, 2019 Oct; 117:31-37.

14. Martin A, Eglit GML, Maldonado Y, Daly R, **Liu, J**, Tu X, Jeste DV. (2019). Attitude Toward Own Aging Among Older Adults: Implications for Cancer Prevention. *Gerontologist,* 2019 05 17; 59(Suppl 1):S38-S49.

15. Soontornniyomkij V, Lee EE, Jin H, Martin AS, Daly RE, **Liu, J**, Tu XM, Eyler LT, Jeste DV. (2019). Clinical Correlates of Insulin Resistance in Chronic Schizophrenia: Relationship to Negative Symptoms. *Front Psychiatry,* 2019; 10:251.

16. Jeste DV, Glorioso D, Lee EE, Daly R, Graham S, **Liu, J**, Paredes AM, Nebeker C, Tu XM, Twamley EW, Van Patten R, Yamada Y, Depp C, Kim HC. (2019). Study of Independent Living Residents of a Continuing Care Senior Housing Community: Sociodemographic and Clinical Associations of Cognitive, Physical, and Mental Health. *Am J Geriatr Psychiatry.* 2019 Sep; 27(9):895-907.

17. Lee EE, Martin AS, Kaufmann CN, **Liu, J**, Paredes AM, Nebeker C, Tu XM, Twamley EW, Van Patten R, Yamada Y, Depp C and Jeste, D. V. (2019). Comparison of schizophrenia outpatients in residential care facilities with those living with someone: Study of mental and physical health, cognitive functioning, and biomarkers of aging. *Psychiatry Res.* 2019 05; 275:162-168.

18. Meier, A., Yang, J., **Liu, J**, Beitler, J.R., Tu, X.M., Owens, R.L., Sundararajan, R.L., Malhotra, A. and Sell, R.E. (2019). Female physician leadership during cardiopulmonary resuscitation is associated with improved patient outcomes. *Critical Care Medicine*, 2019 01; 47(1): e8-e13.

19. Lee, E.E., Depp, C., Palmer, B.W., Glorioso, D., Daly, R., **Liu, J**, Tu, X.M., Kim, H., Tarr, T., Yamada, Y., Jeste, D.V. (2018). High prevalence of loneliness in community-dwelling adults across the lifespan: Can wisdom be an antidote? *Int Psychogeriatr.*, 2018 Dec 18; 1-16.

20. Wang H, Wang B, Tu XM, **Liu J**, Feng C. (2018). Sample sizes based on three popular indices of risks. *General Psychiatry*, 2018; 31:e100011.

21. Lee EE, **Liu J**, Tu X, Palmer BW, Eyler LT, Jeste DV. (2018). A widening longevity gap between people with schizophrenia and general population: A literature review and call for action. *Schizophr Res.*, 2018 Jun;196:9-13.

22. Zhang X, Lyu J, Tu J, **Liu J**, Lu X. (2017). Sample Size Calculations for Comparing Groups with Binary Outcomes. *Shanghai Archives of Psychiatry*, October 2017 29(5):316-324.

23. **Jinyuan Liu**, Wan Tang, Guanqin Chen, Yin Lu, Changing Feng, Xin M. Tu. (2016). Correlation and agreement: overview and clarification of competing concepts and measures *Shanghai Archives of Psychiatry*, 2016, Vol. 28, No. 2.

ABSTRACT OF THE DISSERTATION

Semiparametric Regression Models for Between- and Within-subject Attributes: Applications to High-Dimensional Data, Asymptotic Efficiency and Beyond

by

Jinyuan Liu

Doctor of Philosophy in Biostatistics

University of California San Diego, 2022

Professor Xin Tu, Chair

Breakthroughs in innovative technologies such as next-generation sequencing and wearable devices are producing flourishing data to facilitate deriving scientific insights. But they also provoke substantial challenges in both statistical analyses and interpretations by bringing forth data that is sparse and astronomically high-dimensional. Directly modeling such raw data is not only laborious due to the untenable model assumption, but also suffers from multiple testing and low power. Therefore, an emerging alternative is to first reduce the data dimension at the outset by comparing two subjects' sequences using a dissimilarity/distance metric, termed "between-subject attributes", for a pair. We refer to their classical counterparts that concern only

one individual as "within-subject attributes."

The method development of this dissertation is motivated by analyzing data from microbiome studies. To drive insights into disease mechanisms, the human microbiome is interrogated using high-throughput sequencing (e.g., 16s sequencing of gut microbiota). This procedure generates taxonomic sequence counts (for each subject) that are sparse and high-dimensional. Due to their inherent non-normality, the concept of diversity is introduced to summarize the microbial community. In fact, in most applications, researchers start with the community-level analysis of the microbiome diversity instead of directly tackling the individual-level raw data that may suffer from weak signals. Specifically, a popular diversity metric that naturally encompasses a between-subject nature is the Beta-diversity, defined by pairwise distances of taxonomic counts between two individuals.

In the first part of this dissertation, we extend the mainstream ANOVA-based diversity analyses tool to a semiparametric regression by modeling the Beta-diversity as the response (dependent variable) in the functional response model (FRM). Its superiority over the existing approach is demonstrated both in scalability and statistical power.

The versatility of the between-subject attributes is by no means confined to one discipline. In fact, they serve as effective dimension reductions in various disciplines. In the second part, by extending the classical generalized linear model (GLM) from within- to between-subject attributes, we present a unified GLM-type semiparametric regression framework for distance metrics, then unravel the intricate connections in high-dimensional sequences from microbiome and wearables. This timely solution provides robust inference about relationships between pairwise distances that are of interest in a mounting number of applications.

Despite the growing implementations of this new paradigm, the efficiency of their estimators has not yet been studied carefully. But this is of fundamental importance for semiparametric models due to the efficiency loss at the price of the minimum model assumptions we posit. In the third part, leveraging the Hilbert-Space-based semiparametric efficiency theory,

we show that estimators from a class of U-statistics-based generalized estimating equation (UGEE) achieve the semiparametric efficiency bound. This semiparametric efficiency allows sensitive signal detection in practice.

Therefore, such a distance-based semiparametric regression framework for between-subject attributes harmonizes efficiency and robustness, which will be propelling growing applications in biomedical, psychosocial, and related research to inform appropriate knowledge discovery and decision-making.

Chapter 1

Introduction

Breakthroughs in innovative technologies are generating flourishing data to facilitate data-driven understandings. At the same time, data with astronomical dimensions also provoke substantial challenges to statistical analyses. For example, it is daunting to find a few individual culprits that are fully responsible for a disease of interest. Additionally, suitable data interpretations are needed to inform appropriate knowledge discovery and decision-making.

Rooted in the high-dimensional data from real applications, the main objectives of this dissertation are to develop new statistical methodologies to 1) overcome the challenges of analyzing high-dimensional data through effective dimension reductions; 2) unify a framework to fill the vital gaps in quantifying their effects by addressing the inherent correlations properly; 3) ground the implementations of the new method from a rigorous theoretical perspective.

To achieve these objectives, we first adopt pairwise distances to reduce data dimension and propose a unified semiparametric regression framework for distance metrics that merit their interest. Harmonizing robustness and efficiency, this new distance-based paradigm provides a timely tool to tackle high-dimensional data and derive scientific insights.

Our method development is motivated by analyzing data from microbiome studies. Fueled by technological revolutions such as next-generation sequencing, high-throughput data play an increasing role in biomedical and other burgeoning research areas. Nowadays, the human microbiome can be interrogated using high-throughput sequencing (e.g., 16s sequencing

of gut microbiota). This procedure generates taxonomic sequence counts (for each subject) that are sparse and astronomically high-dimensional. Directly modeling such sequences is not only challenging due to untenable model assumptions but also suffers from multiple testing and lower power. In most applications, researchers are more interested in the role of the microbiome as a community but lack handy analytical techniques. Hence, an emerging alternative has been adopted to reduce the data dimension at the outset first by comparing two subjects' genome sequences with dissimilarity/distance metrics, which we term the "between-subject attribute" for a pair. Its classical counterpart concerning only one individual is termed the "within-subject attribute." For instance, a popular metric in the microbiome that naturally encompasses such a between-subject nature is the Beta-diversity. Defined by pairwise distances of taxonomic counts between individuals, it is recognized as a key indicator of human health (Durack and Lynch, 2019a). To derive insights into disease mechanisms, various attempts have been made to examine relationships between Beta-diversity and other phenotypic outcomes. Permutational Multivariate Analysis of Variance (PERMANOVA) (McArdle and Anderson, 2001) is one of the mainstream approaches with the ANOVA as a premise. However, it is inflexible to adjust for covariates while also computationally demanding, due to the reliance on permutation for the inference. Therefore, we propose to address such limitations by modeling Beta-diversity in a regression through a class of semiparametric functional response models (FRM), which is uniquely positioned to model between-subject attributes. This timely solution not only disentangles sources of variation carried by Beta-diversity but also generates interpretable results on both the direction and size of the effects, potentially shedding light on the disease onset, progression, and treatment. We also develop a novel approach to non-parametrically simulate life-like Beta-diversity outcomes to help demonstrate the performance of its asymptotic inference.

While the versatility of the between-subject attributes is by no means confined to one discipline, they serve as effective dimension reductions in various biomedical research (Moon

et al., 2017). For example, in this burgeoning digital era for disease diagnosis and prevention, data generated from mHealth studies can facilitate personalized interventions to improve patient care. While to date, the analyses of such data are still in an embryonic stage dominated by descriptive statistics. Even if modern methods such as (multilevel) functional principal component analysis (FPCA) (Di et al., 2009), or penalized multi-band learning (Li et al., 2021) have been proposed, they may be subjected to information loss due to the selected principal components (PCs) or penalization. In some studies, in addition to inspecting population mean function over time (with functional PCs), investigators also aim to capture the variability or heterogeneity of activities among subgroups formed by clinical traits, such as the disease status.

Migrating the notion of between-subject attributes to the longitudinal sequences collected from wearables, we found that pairwise distances compressed from high-dimensional sequences can naturally capture the between-subject variability. For example, the mean of squared Euclidean distance pertains to the variance. Hence, they further motivate us to expand the previous work. By extending the classical generalized linear models (GLM) from within- to between-subject attributes, we present a unified GLM-type semiparametric framework based on distances that accommodate different data types. We also illustrate how to construct such between-subject distances extensively, with motivating examples from the human microbiome and mHealth that we encounter. By modeling several pairwise distances simultaneously, the proposed approach can potentially shed light on a variety of fields. For instance, in microbiome studies, to explore the complex interplay of diet, microbiome, and metabolome in disease phenotypes; in mHealth and epidemiology studies, to connect disease status with sleep quality, physical activity, and even social networks; in genetic studies, to investigate interactions between genetic and environmental factors (G$\times$E interaction) in disease development, etc.

Taken together, this new framework shatters barriers of the predominant paradigm to unify a new class of distance-based semiparametric regression, confronting the challenges of modeling high-dimensional data. Its semiparametric nature grants it robustness to model

misspecifications by relaxing the stringent parametric assumptions. Despite their growing applications, the efficiency of parameter estimators has not yet been studied carefully. But this is of fundamental importance for semiparametric models due to the efficiency loss at the price of the minimum model assumptions we posit. To further ground its implementations, our next goal is to find the estimator(s) with the smallest asymptotic variance under weaker (semiparametric) assumptions, namely, the "semiparametric efficient estimator(s)."

By leveraging the Hilbert-Space-based semiparametric efficiency theory, we first extend regular asymptotic linear estimators and influence functions from the classical within- to between-subject attributes. In the Hilbert space, we introduce two inner products to identify the asymptotic variances and connect them by establishing a "dual orthogonality" property. We show that estimators from the U-statistics-based generalized estimating equation (UGEE) deliver the smallest asymptotic variance for the semiparametric FRM. In a nutshell, akin to GEE for semiparametric GLM, UGEE estimators are asymptotically efficient for the semiparametric FRM, rendering the least efficiency loss to allow sensitive signal detection in practice. With blooming implementations of between-subject attributes as effective dimension reduction tools, the efficiency of UGEE estimators will propel growing applications of the distance-based semiparametric FRM (and other related models) for high-dimensional data.

In summary, the major contributions of this unified paradigm for between-subject attributes include reducing the astronomical data dimensions effectively, harmonizing robustness and efficiency in statistical modeling to inform scientific insights, and being theoretically grounded in statistical inference to accelerate blooming applications in biomedical, psychosocial, and related research. This building block also provides a premise for extensions to longitudinal data and causal effects in the future.

The dissertation is organized as follows. In Chapter 2, we first propose the class of semiparametric functional response models (FRM) to model Beta-diversity. We then present a unified GLM-type semiparametric regression framework designated for distance metrics accom-

modating different data types in Chapter 3. To rigorously ground the application of this new paradigm, in Chapter 4, we leverage the Hilbert-Space-based efficiency theory to demonstrate that estimators from the UGEE deliver the smallest asymptotic variance for the semiparametric FRM. In Chapter 5, we provide concluding remarks and several future directions.

Chapter 2

A Semiparametric Model for Between-Subject Attributes: Applications to Beta-diversity of Microbiome Data

2.1 Introduction

This methodological development is motivated by the problem to test associations between the microbiome diversity and clinical variables. The human microbiome refers to all microorganisms on or in the human body, their genes, and surrounding environmental conditions (National Academies of Sciences and Medicine, 2018). In recent years, a preponderance of microbiome studies have implicated the role of the human microbiome in the pathogenesis of complex diseases, including diabetes, alcoholic liver disease, and even cancers (Lang et al., 2020; Holmes et al., 2011). Therefore, identifying potential biological or clinical variables associated with the microbiome and defining their relationships not only enlighten the inherent disease mechanisms but also enhance modulating microbiome compositions for therapeutic purposes.

Fueled by the technological advancement of next-generation sequencing, the human microbiome can be interrogated using high-throughput sequencing. For example, one strategy amplifies and sequences the bacterial 16S ribosomal RNA gene (16S rRNA) for species identification. These sequences are further clustered into nearly identical Operational Taxonomic

Units (OTUs) and compared with reference databases to produce OTU counts profiles based on taxonomic assignments.

The OTU counts are often sparse and high-dimensional. Direct analysis of such data with limited samples raises several statistical challenges, including modeling the skewed and over-dispersed count data with a preponderance of zeros. Since the sequencing depth varies, OTU counts are usually normalized into proportions within each subject to form the OTU relative abundance. They can be further summarized at the microbial community level using diversity metrics, including the "within-subject" Alpha-diversity and "between-subject" Beta-diversity. Unlike Alpha-diversity that consists of individual outcomes, or within-subject attributes, Beta-diversity considers the number of shared taxa between subjects, thus representing their differences in OTU abundance profiles. Each Beta-diversity outcome is a pairwise distance between two subjects, or between-subject attribute. The two major categories of statistical analyses for the microbiome, i.e., the "individual" level effect of a single OTU and the "community" level effect of microbiome composition with summary statistics of diversity, complement each other.

Notably, a variety of disorders are shown to be associated with the loss of gut microbial diversity (Durack and Lynch, 2019b). One common approach to evaluate such associations using Beta-diversity is the Permutational Multivariate Analysis of Variance Using Distance Matrices (PERMANOVA) (McArdle and Anderson, 2001). This approach partitions the Beta-diversity into within- and between-group variations and implements a permutation test based on pseudo-F statistics for inference. A major limitation is the difficulty to discern the sources of variation when the null hypothesis is rejected. Also, it is unsuitable for between-subject covariates in some applications, such as a dissimilarity measure describing the difference between subjects' metabolites abundance profile. Additionally, it requires a large number of permutations to ensure stable results (Dubitzky et al., 2013). All these limitations severely circumscribe its applications in practice.

In this chapter, we propose a new approach to address the aforementioned limitations of PERMANOVA by utilizing the functional response models (FRM) (Kowalski and Tu, 2008a), which are uniquely positioned to address between-subject attributes defining the Beta-diversity in the current context. We provide a brief overview of the Beta-diversity and PERMANOVA first.

2.2 Beta-diversity and PERMANOVA

2.2.1 Beta-diversity Measures

Beta-diversity captures within- and between-group differences by comparing individuals' distributions of taxonomic units. For example, the Bray-Curtis distance (Sørensen, 1948) is a quantitative measure based on OTU relative abundance. For a pair of subjects i and j, the Bray-Curtis distance is defined by $BC_{ij} = 1 - \frac{2C_{ij}}{S_i+S_j}$, where C_{ij} indicates the sum of the OTU relative abundance that the pair has in common and S_i $\left(S_j\right)$ denotes the total number of OTU relative abundance for the ith (jth) subject. This measure ranges from 0 to 1, with 0 (1) indicating exactly the same (completely different) taxonomic abundances. As Beta-diversity incorporates taxa information into distances, its size is determined by the number of subjects rather than that of taxonomic units for the high-dimensional OTUs.

Unlike the Euclidean distance, most Beta-diversity measures calculate weighted relative differences, where each species' contribution is weighted by the sum of the species' abundance in the two subjects being compared (Roberts, 2017). Some forms such as the Unifrac can additionally account for the phylogenetic distances (Lozupone and Knight, 2005). Hence, non-Euclidean Beta-diversity measures are widely adopted as the basis of statistical analyses to detect a wider range of biologically relevant changes in the microbiome (Legendre and Gallagher, 2001).

2.2.2 PERMANOVA

Consider a sample of n subjects with microbiome profiles (counts) defined by m OTUs. Let $\mathbf{y}_i$ denote an $m \times 1$ column vector of OTU relative abundance (after normalization) and $\mathbf{x}_i$ a vector of explanatory variables such as the status of a disease for the ith subject. Let $d_{\mathbf{i}} = d(\mathbf{y}_{i_1}, \mathbf{y}_{i_2})$ denote a Beta-diversity outcome for a pair of subjects $\mathbf{i} = (i_1, i_2) \in C_2^n$, where C_q^n denotes the set of q-combinations $(i_1, \ldots, i_q)$ from the integer set $\{1, \ldots, n\}$. We are interested in testing the association between the Beta-diversity $d_{\mathbf{i}}$ and some clinical variables such as the status of a disease or, more generally, a continuous explanatory variable such as bilirubin, an indication of liver disease progression.

If $\mathbf{x}_i$ is a categorical variable for groups, PERMANOVA can be used to compare Beta-diversity across different groups, which adopts a pseudo-F statistic for inference (McArdle and Anderson, 2001). We provide details and formulas in the Supporting Information.

PERMANOVA has several limitations. First, it does not provide coefficient estimators for explanatory variables, which hinders generating interpretable results on both the direction and size of the effects, or discerning sources of differences. Second, it describes relationships of Beta-diversity (a between-subject attribute) with within-subject attributes only, not between-subject attributes such as metabolites abundance profile. Also, it requires a large number of permutations for stable results and thus carries more overheads in terms of the computational burden. Additionally, it is quite difficult to extend PERMANOVA to longitudinal studies (with missing data) that are potentially valuable given the dynamic and highly personalized nature of the microbiome.

2.3 Functional Response Models for Beta-diversity

The aforementioned limitations of PERMANOVA result from a lack of ability to model between-subject attributes under the predominant statistical paradigm. With a few exceptions such as the Mann-Whitney-Wilcoxon rank-sum test (Wu et al., 2014a; Lin et al., 2021), all

popular statistical models focus on relationships between variables from the same subject, or within-subject attributes. As Beta-diversity measures the difference between a pair of subjects' OTUs, conventional statistical models are not amenable to modeling such between-subject attributes. In this section, we develop a regression framework to model Beta-diversity by utilizing a class of functional response models (FRM).

2.3.1 Functional Response Models for Between-subject Attributes

Consider a class of semiparametric functional response models (FRM):

$$E\left\{\mathbf{f}\left(\mathbf{y}_{i_1},\ldots,\mathbf{y}_{i_q}\right) \mid \mathbf{x}_{i_1},\ldots,\mathbf{x}_{i_q}\right\} = \mathbf{h}\left(\mathbf{x}_{i_1},\ldots,\mathbf{x}_{i_q};\theta\right), \qquad (2.1)$$
$$\left(i_1,\ldots,i_q\right) \in C_q^n, \quad 1 \le q, \quad 1 \le i \le n,$$

where $\mathbf{y}_i = (y_{i1},\ldots,y_{im})^\top \in \mathbb{R}^m$ denotes the response vector from the ith subject, $\mathbf{f}(\cdot)$ is some vector-valued function, $\mathbf{h}(\cdot)$ is some vector-valued smooth function (e.g., with continuous derivatives up to the second order), θ is a vector of parameters, q is some positive integer. The FRM in (2.1) extends the semiparametric generalized linear models (GLM) from within- to between-subject attributes (Kowalski and Tu, 2008a). For example, when $q = 1$ and $f(y_i) = y_i$, (2.1) immediately reduces to the restricted moment GLM. When $q = 2$ and set

$$f_{\mathbf{i}} = d\left(\mathbf{y}_{i_1},\mathbf{y}_{i_2}\right), \quad h_{\mathbf{i}}(\theta) = E\left\{d\left(\mathbf{y}_{i_1},\mathbf{y}_{i_2}\right)\right\} = \theta, \quad (i_1,i_2) \in C_2^n, \qquad (2.2)$$

the FRM in (2.1) models the Beta-diversity distance $d\left(\mathbf{y}_{i_1},\mathbf{y}_{i_2}\right)$ and provides inference about the mean distance θ.

2.3.2 Functional Response Models for Beta-diversity with Covariates

Group Comparison

We start by comparing Beta-diversity across multiple groups. Consider K groups with n_k denoting the sample size of the kth group ($1 \leq k \leq K$), $n = \sum_{k=1}^{K} n_k$ denoting the total sample size of all K groups combined. Let x_i denote a categorical variable indicating group membership for subject i ($1 \leq x_i \leq K$, $1 \leq i \leq n$).

For each pair, we observe their OTU relative abundance outcomes $\mathbf{y_i} = \{\mathbf{y}_{i_1}, \mathbf{y}_{i_2}\}$ ($\mathbf{i} = (i_1, i_2) \in C_2^n$), along with the pairwise group indicators $\mathbf{x_i} = \{x_{i_1}, x_{i_2}\}$ ($1 \leq x_{i_1}, x_{i_2} \leq K$). Denote all combinations of $\mathbf{x_i}$ with a vector $\delta(\mathbf{x_i}) \in \mathbb{R}^{K+C_2^K}$ through a one-hot encoding function $\delta : \{1,...,K\} \times \{1,...,K\} \mapsto \{0,1\}^{K+C_2^K}$ such that for its $\mathbf{k}^{th}$ ($\mathbf{k} = \{k_1, k_2\}$) entry:

$$\delta_{\mathbf{k}}(\mathbf{x_i}) = \begin{cases} 1 & \text{if } \mathbf{x_i} = \{x_{i_1}, x_{i_2}\} = \{k_1, k_2\} \ = \mathbf{k} \\ 0 & \text{otherwise} \end{cases}, \quad \mathbf{i} = (i_1, i_2) \in C_2^n, \tag{2.3}$$

$$\delta(\mathbf{x_i}) = \left(\delta_{11}(x_{\mathbf{i}}), \ldots, \delta_{(K-1)K}(x_{\mathbf{i}}), \delta_{KK}(x_{\mathbf{i}})\right)^{\top}, \quad 1 \leq k_1 \leq k_2 \leq K.$$

Let $f(\mathbf{y_i}) = d(\mathbf{y}_{i_1}, \mathbf{y}_{i_2})$ and define an FRM:

$$E\{f(\mathbf{y_i}) \mid \delta(\mathbf{x_i})\} = \exp\left\{\sum_{1 \leq k_1 \leq k_2 \leq K} \tau_{k_1 k_2} \delta_{k_1 k_2}(\mathbf{x_i})\right\} = \exp\left\{\theta^{\top} \delta(\mathbf{x_i})\right\}, \tag{2.4}$$

where $\exp(\cdot)$ ensures that the right side of the equation is positive as $f(\mathbf{y_i}) \geq 0$. The FRM above is determined by the parameter vector $\theta = \left(\tau_{11}, \ldots, \tau_{(K-1)K}, \tau_{KK}\right)^{\top}$.

Unlike conventional analysis for within-subject attributes, models for between-subject attributes involve more complex parameters and interpretations. For the FRM in (2.4), $\exp(\tau_{kk})$ is the mean of $f(\mathbf{y_i})$ when both subjects of the $\mathbf{i}^{th}$ pair are from group k, and $\exp(\tau_{k_1 k_2})$ is the mean of $f(\mathbf{y_i})$ when one (the other) is from group k_1 (k_2). Thus, in addition to group means as in conventional within-subject analysis, we now have (1) within-group means $\exp(\tau_{kk})$ and

(2) between-group means $\exp(\tau_{k_1k_2})$. For two groups k_1 and k_2 with the same or similar OTU distributions, their within- and between-group means are usually similar. However, if they have different OTU distributions, they may still have similar within-group means (this can occur, for example, if OTUs' have similar variability within each group), but the between-group means $\exp(\tau_{k_1k_2})$ can be different from within-group means $\exp(\tau_{k_1k_1})$ or $\exp(\tau_{k_2k_2})$.

Thus, under the FRM in (2.4), we are interested in three types of null hypotheses to describe group differences in Beta-diversity:

$$\begin{aligned}
&\text{(1) Within-group}: && H_{01}: \tau_{kk} = \tau_{k'k'} \quad \text{for any } (k,k'), 1 \le k < k' \le K \\
& && H_{a1}: \tau_{kk} \neq \tau_{k'k'} \quad \text{for some } (k,k') \\
&\text{(2) Between-group}: && H_{02}: \tau_{kl} = \tau_{k'l'} \quad \text{for any } (k,l,k',l'), 1 \le k,k' < l,l' \le K \\
& && H_{a2}: \tau_{kl} \neq \tau_{k'l'} \quad \text{for some } (k,l,k',l') \\
&\text{(3) Within- vs between-group}: && H_{03}: \tau_{kk} = \tau_{k'l'} \quad \text{for any } (k,k',l'), 1 \le k \le K, 1 \le k' < l' \le K \\
& && H_{a3}: \tau_{kk} \neq \tau_{k'l'} \quad \text{for some } (k,k',l')
\end{aligned} \tag{2.5}$$

Hypotheses (2) and (3) are unique to between-subject attributes, each revealing different aspects. For example, if the patterns of OTU distribution are "flipped" across two groups, the difference of Beta-diversity could be detected by the "within- vs. between-" instead of the "within-" type of hypothesis.

For PERMANOVA, if we obtain an insignificant pseudo-F statistic, we conclude with not enough evidence to reject the null. But, if this test is significant, it is unclear if the difference occurs in within-group or between-group means or both. By partitioning sources of variation and building formal hypotheses to depict the underlying differences of microbiome diversity across groups, a formal regression model for between-subject attributes in (2.4) allows for discerning sources of differences, potentially leading to more in-depth scientific findings.

All three types of hypotheses in (2.5) are readily tested using linear contrasts: $H_0: \mathbf{C}\theta = \mathbf{0}$ vs. $H_a: \mathbf{C}\theta \neq \mathbf{0}$, where $\mathbf{C}$ is a matrix of known constants. For example, when comparing

Beta-diversity for three groups, we may use the following $\mathbf{C}$ matrices to test the hypotheses in (2.5):

$$K=3,\ \theta\ =(\tau_{11},\tau_{22},\tau_{33},\tau_{12},\tau_{13},\tau_{23})^{\top},\quad (a):\mathbf{C}_1=\left(\begin{array}{ccc}\mathbf{1}_2, & (-1)\cdot\mathbf{I}_2, & \mathbf{0}_{2\times 3}\end{array}\right);\quad (2.6)$$
$$(b):\mathbf{C}_2=\left(\begin{array}{ccc}\mathbf{0}_{2\times 3}, & \mathbf{1}_2, & (-1)\cdot\mathbf{I}_2\end{array}\right);\quad (c):\mathbf{C}_3=\left(\begin{array}{cc}\mathbf{1}_5, & (-1)\cdot\mathbf{I}_5\end{array}\right),$$

where $\mathbf{1}_n$ denotes a $n\times 1$ column vector of 1's, and $\mathbf{I}_n$ denotes the $n\times n$ identity matrix.

Covariates for Confounders

As most human population studies of microbiome are observational due to cost, logistic, and difficulties in experimental control, it is crucial to control for potential confounders that may impact group differences, such as demographics (ethnicity, genetic background), biometrics (medications, diet), molecular measures (microbial metabolites, gene expression), and environmental exposures (National Academies of Sciences and Medicine, 2018). A more substantial improvement over PERMANOVA is FRM's ease to control for a broader range of confounders, including between-subject attributes such as metabolites abundance profiles. This is achieved by leveraging the regression feature of FRM to include either within- or between-subject covariates.

As a motivating example for including a within-subject covariate, consider a linear regression relating a continuous variable z_i to a continuous response y_i: $y_i=\eta_0+\eta_1 z_i+\varepsilon_i$, $\varepsilon_i\sim\left(0,\sigma^2\right)$, $1\le i\le n$, where $\left(0,\sigma^2\right)$ denotes some continuous distribution with mean zero and variance σ^2. Now consider the squared difference, $f(y_{\mathbf{i}})=(y_{i_1}-y_{i_2})^2$. It follows that

$$E\left\{f(y_{\mathbf{i}})\mid z_{i_1},z_{i_2}\right\}=E\left(\varepsilon_{i_1}-\varepsilon_{i_2}\right)^2+\eta_1^2\left(z_{i_1}-z_{i_2}\right)^2=2\sigma^2+\eta_1^2\left(z_{i_1}-z_{i_2}\right)^2. \qquad (2.7)$$

Although Beta-diversity is more complex, we use the same rationale to control for covariates by adding $(z_{i_1}-z_{i_2})^2$, or a more general non-negative transformation $g(\mathbf{z}_{\mathbf{i}})$ of $\mathbf{z}_{\mathbf{i}}=$

$\{z_{i_1}, z_{i_2}\}$ to the FRM in (2.4):

$$E\{f(y_{\mathbf{i}}) \mid \delta(\mathbf{x}_{\mathbf{i}}), \mathbf{z}_{\mathbf{i}}\} = \exp\left\{ \sum_{1 \le k_1 \le k_2 \le K} \tau_{k_1 k_2} \delta_{k_1 k_2}(\mathbf{x}_{\mathbf{i}}) + \xi_1 g(\mathbf{z}_{\mathbf{i}}) \right\}, \ \mathbf{i} = (i_1, i_2) \in C_2^n. \quad (2.8)$$

For a categorical covariate, we can define a series of indicators akin to (3.8), i.e. for the $\mathbf{i}^{th}$ pair, we observe the pairwise indicators $\mathbf{x}_{l\mathbf{i}} = \{x_{li_1}, x_{li_2}\}$ $(1 \le x_{li_1}, x_{li_2} \le K_l)$ for the l^{th} $(1 \le l \le p)$ categorical covariate with K_l levels. We one-hot encode those p categorical covariates into $\delta(\mathbf{x}_{\mathbf{i}}) \in \mathbb{R}^{1+\sum_{l=1}^{p}(K_l + C_2^{K_l} - 1)}$, with the encoding function defined similarly as in (3.8), but designating a referent to obtain a similar form as in conventional regression.

Specifically, for the l^{th} categorical covariate, we define $\delta_l : \{1, ..., K_l\} \times \{1, ..., K_l\} \mapsto \{0,1\}^{K_l + C_2^{K_l} - 1}$ (excluding the case where $k_{l1} = k_{l2} = 1$) such that for the $\mathbf{k}_l^{th}$ $(\mathbf{k}_l = \{k_{l1}, k_{l2}\})$ entry of $\delta_l(\mathbf{x}_{l\mathbf{i}})$:

$$\delta_{l\mathbf{k}}(\mathbf{x}_{l\mathbf{i}}) = \begin{cases} 1 & \text{if } \mathbf{x}_{l\mathbf{i}} = \{x_{li_1}, x_{li_2}\} = \{k_{l1}, k_{l2}\} \ = \mathbf{k}_l \\ 0 & \text{otherwise} \end{cases}, \quad (2.9)$$

$$\delta_l(\mathbf{x}_{l\mathbf{i}}) = \left(\delta_{l12}(\mathbf{x}_{l\mathbf{i}}), \ldots, \delta_{l(K-1)K}(\mathbf{x}_{l\mathbf{i}}), \delta_{lKK}(\mathbf{x}_{l\mathbf{i}})\right)^\top, \quad 1 \le l \le p,$$

$$\delta(\mathbf{x}_{\mathbf{i}}) = \left(1, \delta_1(\mathbf{x}_{1\mathbf{i}})^\top, \ldots, \delta_l(\mathbf{x}_{l\mathbf{i}})^\top, ..., \delta_p(\mathbf{x}_{p\mathbf{i}})^\top\right)^\top,$$

$$\mathbf{i} = (i_1, i_2) \in C_2^n, \quad 1 \le k_{l1} \le k_{l2} \le K_l, \quad 1 = k_{l1} \ne k_{l2}.$$

Thus, with p categorical covariates (including one for diagnostic groups), x_{li} $(1 \le l \le p)$, and q continuous covariates, z_{mi} $(1 \le m \le q)$ for subject i, we can, after designating the first group as the referent by including an intercept β_0, express the FRM as:

$$\begin{aligned} E\{f(\mathbf{y}_{\mathbf{i}}) \mid \mathbf{x}_{\mathbf{i}}, \mathbf{z}_{\mathbf{i}}\} &= \exp\left\{ \beta_0 + \sum_{l=1}^{p} \left(\sum_{1 \le k_{l1} \le k_{l2} \le K_l}^{1 = k_{l1} \ne k_{l2}} \beta_{lk_1k_2} \delta_{lk_1k_2}(\mathbf{x}_{l\mathbf{i}}) \right) + \sum_{m=1}^{q} \xi_m g_m(\mathbf{z}_{m\mathbf{i}}) \right\}, \\ &= \exp\left\{ \beta^\top \delta(\mathbf{x}_{\mathbf{i}}) + \xi^\top \mathbf{g}(\mathbf{z}_{\mathbf{i}}) \right\}, \end{aligned} \quad (2.10)$$

where $\mathbf{x}_{l\mathbf{i}} = \{x_{li_1}, x_{li_2}\}$, $\mathbf{z}_{m\mathbf{i}} = \{z_{mi_1}, z_{mi_2}\}$, $\mathbf{g}(\mathbf{z}_{\mathbf{i}}) = \left(g_1(\mathbf{z}_{1\mathbf{i}}), \ldots, g_q\left(\mathbf{z}_{q\mathbf{i}}\right)\right)^\top$ and K_l denotes the levels of category of the l^{th} categorical variable x_{li} $(1 \leq l \leq p)$. The FRM above is parameterized by a vector $\theta \in \mathbb{R}^{1+\sum_{l=1}^{p}(K_l + C_2^{K_l} - 1)+q}$:

$$\begin{aligned} \beta_l &= \left(\beta_{l12\ldots}, \beta_{l(K_l-1)K_l}, \beta_{lK_lK_l}\right)^\top, \quad \beta = \left(\beta_0, \beta_1^\top, \ldots, \beta_p^\top\right)^\top, \\ \xi &= \left(\xi_1, \ldots, \xi_q\right)^\top, \quad \theta = \left(\beta^\top, \xi^\top\right)^\top. \end{aligned} \tag{2.11}$$

Akin to (2.4), the parameters for the covariates possess more complex interpretations. For a continuous covariate $\mathbf{z}_{m\mathbf{i}}$, ξ_m represents change in the mean of $\log\{f(\mathbf{y}_{\mathbf{i}})\}$ per unit change in $g_m(\mathbf{z}_{m\mathbf{i}})$. For a categorical one, say gender, we now have male-male, female-female, or male-female pairs. If we set male-male as the referent, coefficients for female-female and male-female pairs represent differences in the log of mean Beta-diversity when comparing the respective gender pair to the referent.

We illustrate this model with a relatively simple log-linear form in (2.10), yet the applicability of FRM is far beyond the assumed simple relationship. Like any regression model such as the GLM, more complex relationships such as higher-order terms and interactions can be specified as deemed appropriate. The FRM in (2.10) looks like a conventional (log-linear) regression model, except that $\mathbf{i}$ indexes pairs of, rather than, individual, subjects. This critical difference precludes applications of standard inference methods for regression models as we discuss next.

Inference

As the response function $f_{\mathbf{i}} = f(\mathbf{y}_{\mathbf{i}})$ of the FRM-based regression for Beta-diversity in (2.10) involves pairs of subjects, inferences about θ must address the interlocking dependence of $f_{\mathbf{i}}$'s. Since this type of dependence structure is not addressed by standard methods such as the Generalized Estimating Equations (GEE), we develop inferences using a class of U-statistics based Generalized Estimating Equations (UGEE).

U-statistics based Generalized Estimating Equations

Let

$$S_{\mathbf{i}} = f_{\mathbf{i}} - h_{\mathbf{i}},\ \mathbf{D}_{\mathbf{i}} = \frac{\partial}{\partial \theta} h_{\mathbf{i}},\ V_{\mathbf{i}} = Var\left(f_{\mathbf{i}} \mid \mathbf{x}_{\mathbf{i}}, \mathbf{z}_{\mathbf{i}}\right),\ \mathbf{i} = (i_1, i_2) \in C_2^n, \tag{2.12}$$

in practice, $V_{\mathbf{i}}$ is generally unknown and substituted by a working variance such as $V_{\mathbf{i}}(h_{\mathbf{i}}) = h_{\mathbf{i}}$, as the form of FRM is similar to log-linear models for within-subject attributes. Thus, define the UGEE:

$$\mathbf{U}_n(\theta) = \sum_{\mathbf{i} \in C_2^n} \mathbf{U}_{n,\mathbf{i}} = \sum_{\mathbf{i} \in C_2^n} \mathbf{D}_{\mathbf{i}} V_{\mathbf{i}}^{-1} S_{\mathbf{i}} = \mathbf{0}, \tag{2.13}$$

where the estimates $\widehat{\theta}$ are obtained through the Newton-Raphson method (see the Supporting Information for details).

Although similar in appearance, the UGEE above is not a sum of independent variables as in GEE (Tang, He, and Tu, 2012a). Standard asymptotic methods such as the central limit theorem cannot be applied directly, but the theory of U-statistics is useful for addressing such interlocking dependence. For ease of reference, we summarize the asymptotic properties in the theorem below and provide a sketch of proof in the Supporting Information.

Theorem 2.1. Let

$$\begin{aligned} &\mathbf{v}_{i_1} = E\left(\mathbf{U}_{n,\mathbf{i}} |\ \mathbf{y}_{i_1}, \mathbf{x}_{i_1}, \mathbf{z}_{i_1}\right), \quad \mathbf{B} = E\left(\mathbf{D}_{\mathbf{i}} V_{\mathbf{i}}^{-1} \mathbf{D}_{\mathbf{i}}^{\top}\right), \\ &\Sigma_U = 4Var\left(\mathbf{v}_{i_1}\right), \quad \Sigma_\theta = \mathbf{B}^{-1} \Sigma_U \mathbf{B}^{-1}, \quad \mathbf{i} = (i_1, i_2) \in C_2^n. \end{aligned} \tag{2.14}$$

Then under mild regularity conditions,

(a) $\widehat{\theta}$ is consistent and asymptotically normal:

$$\sqrt{n}\left(\widehat{\theta} - \theta\right) \to_d N\left(\mathbf{0}, \Sigma_\theta\right), \tag{2.15}$$

where $\rightarrow_d$ denotes convergence in distribution.

(b) A consistent estimate of Σ_θ is obtained by substituting consistent estimates of θ and moments of the respective quantities in Σ_θ.

Theorem 1 above is readily applied to test any linear hypotheses concerning θ, such as the linear contrasts in (2.6). Under the null, the Wald statistic has an asymptotic χ^2 distribution:

$$W_n = n\left(\mathbf{C}\widehat{\theta}\right)^\top \left(\mathbf{C}\widehat{\Sigma}_\theta \mathbf{C}^\top\right)^{-1} \left(\mathbf{C}\widehat{\theta}\right) \rightarrow_d \chi^2_s, \tag{2.16}$$

where s is the rank of $\mathbf{C}$ and χ^2_s denotes a (central) χ^2 distribution with s degrees of freedom. For example, in testing the within-group difference H_{01} in (2.6), $W_n \rightarrow_d \chi^2_2$ under H_{01}.

The Score Test

As Wald-type tests are typically anti-conservative, score statistics may be used as an alternative to reduce such bias, especially for small to moderate samples (Kennedy, 2003). To develop a score statistic based on the UGEE in (6.9), let $\theta = \left(\theta^\top_{(1)}, \theta^\top_{(2)}\right)^\top$, where $\theta_{(2)}$ is the parameter of interest, $\theta_{(1)} \in \mathbb{R}^p$, $\theta_{(2)} \in \mathbb{R}^q$. Consider testing the null $H_0 : \theta_{(2)} = \theta_{(20)}$, with $\theta_{(20)}$ a vector of known constants. We have the partition:

$$\mathbf{D}_\mathbf{i} = \left(\frac{\partial h(\theta)}{\partial \theta_{(1)}}, \frac{\partial h(\theta)}{\partial \theta_{(2)}}\right)^\top = \left(\mathbf{D}_{\mathbf{i}(1)}, \mathbf{D}_{\mathbf{i}(2)}\right)^\top, \quad \mathbf{U}_n(\theta) = \left(\mathbf{U}_{n(1)}(\theta), \mathbf{U}_{n(2)}(\theta)\right)^\top, \tag{2.17}$$

let $\widetilde{\theta}_{(1)}$ denote the estimate of $\theta_{(1)}$ from solving the following reduced estimating equation given $\theta_{(2)} = \theta_{(20)}$:

$$\mathbf{U}_{n(1)}\left(\theta_{(1)}, \theta_{(20)}\right) = \binom{n}{2}^{-1} \sum_{\mathbf{i} \in C^n_2} \mathbf{D}_{\mathbf{i}(1)} V_\mathbf{i}^{-1} S_\mathbf{i} = \mathbf{0}. \tag{2.18}$$

To define the score statistic, let

$$\widetilde{\theta} = \left(\widetilde{\theta}_{(1)}, \theta_{(20)}\right)^{\top}, \quad \mathbf{B} = E\left(\mathbf{D_i} V_{\mathbf{i}}^{-1} \mathbf{D_i}^{\top}\right) = \begin{pmatrix} \mathbf{B}_{11} & \mathbf{B}_{12} \\ \mathbf{B}_{12}^{\top} & \mathbf{B}_{22} \end{pmatrix}, \tag{2.19}$$

$$\mathbf{G} = \begin{pmatrix} -\mathbf{B}_{21}\mathbf{B}_{11}^{-1}, & \mathbf{I}_q \end{pmatrix}, \quad \Sigma_{(2)} = \mathbf{G}\Sigma_U \mathbf{G}^{\top},$$

where $\mathbf{I}_q$ denotes the $q \times q$ identity matrix, $\mathbf{B}_{11} \in \mathbb{R}^{p\times p}$, $\mathbf{B}_{12} \in \mathbb{R}^{p\times q}$, and $\mathbf{B}_{22} \in \mathbb{R}^{q\times q}$ denote the respective submatrices from partitioning the matrix $\mathbf{B} \in \mathbb{R}^{(p+q)\times(p+q)}$, and Σ_U is defined in (2.14). Let

$$\widetilde{\mathbf{U}}_{n(2)} = \mathbf{U}_{n(2)}\left(\widetilde{\theta}_{(1)}, \theta_{(20)}\right), \quad \widetilde{\Sigma}_{(2)}^{-1} = \Sigma_{(2)}^{-1}\left(\widetilde{\theta}_{(1)}, \theta_{(20)}\right), \tag{2.20}$$

i.e., the quantities of $\mathbf{U}_{n(2)}$ and $\Sigma_{(2)}$ with θ substituted by $\widetilde{\theta}$. The theorem below summarizes the asymptotic properties of the score statistic.

Theorem 2.2. Under mild regularity conditions and $H_0 : \theta_{(2)} = \theta_{(20)}$, the score test statistic $S_n\left(\widetilde{\theta}_{(1)}, \theta_{(20)}\right)$ has an asymptotic χ_q^2 distribution with q degrees of freedom, i.e.,

$$S_n\left(\widetilde{\theta}_{(1)}, \theta_{(20)}\right) = n\widetilde{\mathbf{U}}_{n(2)}^{\top}\widetilde{\Sigma}_{(2)}^{-1}\widetilde{\mathbf{U}}_{n(2)} \to_d \chi_q^2. \tag{2.21}$$

A sketch of proof is provided in the Supporting Information.

2.4 Applications

We first investigated the performance of this FRM approach and compared it with the PERMANOVA, then applied it to a study on alcoholic liver disease (ALD). For Monte Carlo (MC) simulations, we set $M = 1{,}000$ for MC iterations, two-sided type I error rate $\alpha = 0.05$, and sample size (per group) $n_k = 50$, 100, 500 ($k = 1, 2$) for two groups. All analyses were performed within the R software platform (Team, 2017), with code optimized using Rcpp (Eddelbuettel et al., 2011) for run-time improvement, which is available as Supporting

Information.

2.4.1 Simulation Study

Beta-diversity is a feature summarization for the high-dimensional and zero-inflated counts of taxonomic units extracted from sequence data. Hence, our approach is to first generate those taxonomic abundances, and then compute Beta-diversity distances from the normalized taxonomic abundances. Also, as microbial abundances for each taxonomic unit are usually not independent, common approaches to generate taxonomic abundances from parametric distributions fail to produce life-like microbiome data (Zhang et al., 2017). We thus develop an approach to generate data that resemble real taxonomic abundances based on their empirical cumulative distribution function (eCDF) and copula (See the Supporting Information for details). As this procedure does not involve analytical distributional models, population-level characteristics such as the mean are estimated by Monte Carlo simulation with a large MC size of 5,000.

Simulation Settings

We generated Beta-diversity outcomes from eCDFs of OTU counts in a study on alcoholic liver disease (Lang et al., 2020). Chronic alcohol consumption increases intestinal permeability and changes the intestinal microbiota composition, which contributes to the progression of alcohol-related liver disease (ALD). In this study, $n = 85$ subjects including 59 alcoholic hepatitis (AH) patients, 15 alcohol user disorder (AUD) patients, and 11 healthy controls (HC) were enrolled. Fungal ITS sequencing and analysis were conducted using the Illumina MiSeq V3 platform specific for the fungal ITS1 region, resulting in $p = 81$ detected genera. Beta-diversity were computed from the OTU relative abundance vector $\mathbf{Y}_{85\times 81} = [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_{85}]^\top$. For space consideration, we reported results using the Bray-Curtis distance.

Shown in the left-most panel of Figure 1 are eCDFs of Beta-diversity in the three diagnostic groups. The eCDFs are considerably different between the AH and HC as well as

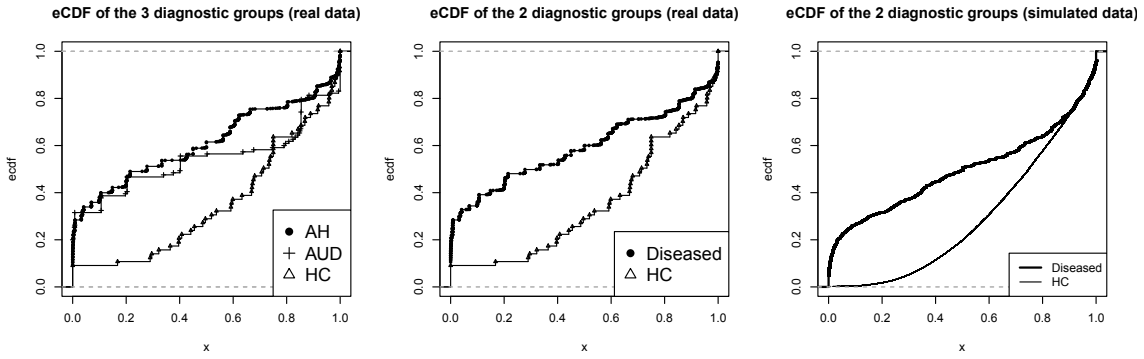

Figure 2.1. Empirical cumulative distribution functions (eCDF) of OTU relative abundances for (1) real data of alcoholic hepatitis (AH) patients, alcohol user disorder (AUD) patients, and non-alcoholic controls (HC) (left) (2) real data of combined diseased (AH and AUD patients) group and non-alcoholic controls (HC) (middle), and (3) simulated data of combined diseased (AH and AUD patients) group and non-alcoholic controls (HC) (right).

AUD and HC group, but less so between the AUD and AH. To illustrate, we combined the AH and AUD patients and simulated OTUs from this combined disease (D) and HC group. Shown in the center of Figure 1 are the eCDFs of observed Beta-diversity for the D and HC group, and in the right-most panel are those of the simulated Beta-diversity for a sample size of $n_k = 500$, which are nearly identical to their original counterparts. The Supporting Figure 1 provides Principal Coordinates Analysis (PCoA) plot, a popular visualizing tool for Beta-diversity (Kruskal and Wish, 1978), which also reveals similar patterns.

To assess whether the data generating procedure retains the important feature of zero-inflated OTUs, we evaluated the average percentage of zero counts in real (93.93%) and simulated OTUs, which are 93.34% ($sd = .004$) for $n_k = 50$; 93.55% ($sd = .003$) for $n_k = 100$; and 94.10% ($sd = .001$) for $n_k = 500$, indicating that the simulated OTUs do reflect the zero-inflated nature of the real OTUs.

Group Comparison

We first considered group comparisons without any covariate, where the FRM parameterized with an intercept is given by:

$$E\left\{f\left(\mathbf{y_i}\right) \mid \mathbf{x_i}\right\} = h\left(\mathbf{x_i}, \theta\right) = \exp\left\{\beta_0 + \beta_{22}\delta_{22}\left(x_{\mathbf{i}}\right) + \beta_{12}\delta_{12}\left(x_{\mathbf{i}}\right)\right\}, \tag{2.22}$$
$$\mathbf{i} = (i_1, i_2) \in C_2^n, \quad \theta = (\beta_0, \beta_{22}, \beta_{12})^\top,$$

where $n = n_1 + n_2$ with n_k denoting the sample size of group k and $f\left(\mathbf{y_i}\right) = d_{i_1,i_2}$ denoting the Beta-diversity outcome for pair $\mathbf{i} = (i_1, i_2) \in C_2^n$. The three types of hypotheses are:

$$\begin{aligned}
\text{Within-group}: &\quad H_{01}: \beta_{22} = 0, \quad \text{vs.} \quad H_{a1}: \beta_{22} \neq 0, \\
\text{Between-group}: &\quad H_{02}: \beta_{12} = 0, \quad \text{vs.} \quad H_{a2}: \beta_{12} \neq 0, \\
\text{Within- vs. between- group}: &\quad H_{03}: \beta_{22} = \beta_{12}, \quad \text{vs.} \quad H_{a3}: \beta_{22} \neq \beta_{12}.
\end{aligned} \tag{2.23}$$

To assess the performance of the proposed approach for varying sample sizes, we simulated OTUs from a single group based on the eCDF of group D using the copula approach. In this case, all three null hypotheses in (2.23) hold.

Let $\widehat{\theta}^{(m)}$ denote the estimator of θ and $\widehat{\Sigma}_\theta^{(m)}$ the asymptotic variance from the mth MC iteration, $\widehat{\theta}$ and $\widehat{\Sigma}_\theta^{(asymp)}$ denote the sample mean of $\widehat{\theta}^{(m)}$ and $\widehat{\Sigma}_\theta^{(m)}$, respectively, and let $\widehat{\Sigma}_\theta^{(emp)}$ denote the sample variance of $\widehat{\theta}^{(m)}$. Let $W_n^{(m)}$ denote the Wald statistic in (2.16) for testing a hypothesis at the mth MC iteration. The type I error rate based on the asymptotic variance is given by $\widehat{\alpha}^W = (1/M)\sum_{m=1}^M I\left(W_n^{(m)} \geq q_{s,0.95}\right)$, where $q_{s,0.95}$ denotes the 95^{th} percentile of a central χ^2 distribution with s degrees of freedom. The score type I error rate $\widehat{\alpha}^s$ was computed similarly by replacing $W_n^{(m)}$ with the score statistic in (2.21) at the mth iteration.

We assess the asymptotic performance by comparing asymptotic and empirical standard errors from $\widehat{\Sigma}_\theta^{(asymp)}$ and $\widehat{\Sigma}_\theta^{(emp)}$, and by comparing $\widehat{\alpha}^W$ ($\widehat{\alpha}^s$) and $\alpha = 0.05$.

Table 2.1. MC estimates, standard errors (asymptotic and empirical) for FRM under the null hypotheses, averaged over MC $M = 1,000$ iterations.

Under Null Hypotheses			
Parameter	Est.	Std. err	
		Asymptotic	Empirical
	$n_k = 50$		
β_0	-.438	.091	.093
β_{22}	.003	.128	.133
β_{12}	.004	.066	.068
	$n_k = 100$		
β_0	-.452	.066	.065
β_{22}	.0003	.093	.096
β_{12}	.002	.048	.049
	$n_k = 500$		
β_0	-.458	.030	.031
β_{22}	.0007	.043	.043
β_{12}	.0006	.021	.021

Shown in Table 2.1 are estimates (Est.) of θ, asymptotic and empirical standard errors. $\widehat{\beta}_{22}$ and $\widehat{\beta}_{12}$ were quite close to 0 (true value). The true $\beta_0 = -0.4595$ was obtained by the sample mean of Beta-diversity for a large MC sample size of $5,000$. The estimated $\widehat{\beta}_0$'s were close to the truth for all three sample sizes. The asymptotic standard errors were close to their empirical counterparts. As expected, discrepancies became smaller as the sample size increased. But estimates and asymptotic standard errors of θ were still good for $n_k = 50$.

Shown in Table 2.2 are type I errors of FRM for the three nulls in (2.23) and PERMANOVA for the overall group difference. For the FRM, although exhibiting a small upward bias for $n_k = 50$, the Wald type I errors were close to $\alpha = 0.05$ in all three cases. The score tests worked well to reduce bias for $n_k = 50$ and 100 with nearly identical type I errors as the Wald for large sample sizes. PERMANOVA also performed well, albeit with a small downward bias for $n_k = 50$ and 100, which often occurs for small sample sizes (Hemerik et al., 2018).

Table 2.2. Comparison of type I error rates between FRM (based on Wald and Score tests) and PERMANOVA (based on permutation).

	FRM: Type of Hypothesis			PERMANOVA
Sample size n_k	Within-: $H_{01}: \beta_{22} = 0$	Between- $H_{02}: \beta_{12} = 0$	Within- vs. Between- $H_{03}: \beta_{22} = \beta_{12}$	
		Type I Error Rates (Wald)		
50	.045	.081	.087	
100	.046	.063	.071	
500	.047	.053	.057	
		Type I Error Rates (Score)		Type I Error Rates
50	.038	.048	.054	.043
100	.044	.047	.054	.048
500	.047	.051	.053	.051

Group Comparison Accounting for Covariates

We illustrate with one continuous and one binary covariate, with the same two diagnostic groups as in (2.22), the FRM becomes:

$$E\left\{f\left(\mathbf{y_i}\right) \mid \mathbf{x_i}, z_\mathbf{i}\right\} = h\left(\mathbf{x_i}, z_\mathbf{i}; \theta\right) = \exp\left(\mathbf{u_i}^\top \theta\right), \tag{2.24}$$

$$\mathbf{u_i}^\top \theta = \beta_0 + \beta_{22}^d \delta_{22}^d\left(x_\mathbf{i}^d\right) + \beta_{12}^d \delta_{12}^d\left(x_\mathbf{i}^d\right) + \beta_{22}^g \delta_{22}^g\left(x_\mathbf{i}^g\right) + \beta_{12}^g \delta_{12}^g\left(x_\mathbf{i}^g\right) + \xi^a g^a\left(z_\mathbf{i}^a\right),$$

$$\theta = \left(\beta_0, \beta_{22}^d, \beta_{12}^d, \beta_{22}^g, \beta_{12}^g, \xi^a\right)^\top, \quad \mathbf{i} = (i_1, i_2) \in C_2^n,$$

where $x_\mathbf{i}^d$, $x_\mathbf{i}^g$ and $z_\mathbf{i}^a$ denote the diagnostic group, binary and continuous covariates for each pair $\mathbf{i} \in C_2^n$. In addition to the three null hypotheses comparing diagnostic groups, two new hypotheses can be tested with $H_{04a}: \xi^a = 0$ for the continuous and $H_{04b}: \beta_{22}^g = \beta_{12}^g = 0$ for the binary covariate. Simulation details are provided in the Supporting Information.

Shown in Table 2.3 are estimates and results for testing the nulls. Again, all estimates were close to their respective true values, and asymptotic standard errors were close to their

Table 2.3. MC estimates, standard errors (asymptotic and empirical), and type I error rates (Wald and Score) of FRM controlling for covariates under the null hypotheses, averaged over MC $M = 1,000$ iterations.

Categorical Covariate: Gender (β^g), Continuous Covariate: Age (ξ^a)					
Parameter	Est.	Std. err		Type I Error	
		Asymptotic	Empirical	Wald	Score
		$n_k = 50$			
β_0	-.442	.127	.135	.087	.048
β_{22}^d	.003	.130	.139	.059	.055
β_{12}^d	.004	.068	.072	.074	.045
β_{22}^g	.497	.129	.133	.047	.039
β_{12}^g	.501	.066	.069	.084	.056
ξ^a	.500	.098	.097	.050	.037
		$n_k = 100$			
β_0	-.456	.085	.083	.057	.046
β_{22}^d	.0005	.094	.097	.060	.055
β_{12}^d	.002	.048	.049	.076	.059
β_{22}^g	.502	.094	.094	.046	.044
β_{12}^g	.502	.048	.048	.064	.046
ξ^a	.500	.056	.055	.048	.044
		$n_k = 500$			
β_0	-.456	.039	.041	.057	.056
β_{22}^d	.0003	.043	.044	.050	.050
β_{12}^d	.0004	.022	.022	.049	.046
β_{22}^g	.498	.043	.045	.055	.056
β_{12}^g	.499	.021	.022	.061	.057
ξ^a	.500	.029	.029	.049	.050

empirical counterparts. Wald and score type I errors were also close to the nominal value, albeit a bit inflated for the Wald with $n_k = 50$. The gaps between Wald and score type I errors became negligible with large sample sizes.

Power Comparison with the Existing Approach

We then compared the power and computational time of the proposed FRM with PERMANOVA to highlight its advantages.

Specifically, we compared hypotheses: (1) "Between-group" difference with PERMANOVA and (2) "Within-group" difference with 'betadisper' function in *'vegan'* (Oksanen et al., 2013) as a proxy, since PERMANOVA does not directly test this hypothesis. Since it is not straightforward for PERMANOVA to test (3) "Within- vs. Between-group" difference, we did not include this comparison. The simulation details are provided in the Supporting Information. Both permutation-based PERMANOVA and 'betadisper' were conducted with the number of permutations set to 99, 299, 499, and 999, respectively.

Shown in Table 2.4 are group size, effect size, power, and elapsed time (of one iteration) for comparison. In detecting between-group differences (i.e., location), FRM outperformed PERMANOVA in both power and scalability. Not only did FRM attain much higher power, but it also required far less computing time. For within-group differences (i.e., dispersion), FRM still surpassed 'betadisper' in scalability and achieved slightly higher power. For both PERMANOVA and 'betadisper', the computational time increased dramatically with the increased number of permutations.

2.4.2 Real Data Analyses

We also applied the proposed FRM to the alcoholic liver disease study (Lang et al., 2020) to compare Beta-diversity among the original three diagnostic groups. Our goal was to identify the association between the microbiome diversity and diagnostic groups, controlling

Table 2.4. Comparisons of power and computational time between FRM and PERMANOVA as well as 'betadisper', with the number of permutations set to 99, 299, 499, and 999 for both permutation-based approaches.

"Between-group" difference (location): FRM vs. PERMANOVA											
n_k	Eff. Size	Power					Time for one iteration (s)				
		FRM	PERMANOVA (#)				FRM	PERMANOVA (#)			
			99	299	499	999		99	299	499	999
50	.322	.637	.152	.168	.172	.176	.009	.017	.051	.079	.180
100	.346	.905	.383	.423	.431	.441	.024	.078	.238	.408	.878
200	.346	.994	.892	.927	.922	.921	.108	.332	1.051	1.929	3.642

"Within-group" difference (dispersion): FRM vs. 'betadisper'											
n_k	Eff. Size	Power					Time for one iteration (s)				
		FRM	Betadisper (#)				FRM	Betadisper (#)			
			99	299	499	999		99	299	499	999
50	.352	.698	.662	.698	.697	.691	.009	.015	.040	.062	.121
100	.366	.956	.914	.922	.928	.925	.024	.015	.041	.064	.126
200	.362	1.000	.996	1.000	.999	.998	.108	.020	.049	.075	.153

for demographics. The FRM for diagnostic groups and two covariates of gender and age is:

$$E\left\{f\left(\mathbf{y_i}\right) \mid \mathbf{x_i}, z_\mathbf{i}\right\} = h\left(\mathbf{x_i}, z_\mathbf{i}; \theta\right) = \exp\left(\mathbf{u_i}^\top \theta\right), \tag{2.25}$$

$$\mathbf{u_i} = \left(1, \delta_{22}^d\left(x_\mathbf{i}^d\right), \delta_{33}^d\left(x_\mathbf{i}^d\right), \delta_{12}^d\left(x_\mathbf{i}^d\right), \delta_{13}^d\left(x_\mathbf{i}^d\right), \delta_{23}^d\left(x_\mathbf{i}^d\right), \delta_{22}^g\left(x_\mathbf{i}^g\right),$$

$$\delta_{12}^g\left(x_\mathbf{i}^g\right), g^a\left(z_\mathbf{i}^a\right)\Big)^\top$$

$$\mathbf{i} = (i_1, i_2) \in C_2^n, \quad \theta = \left(\beta_0, \beta_{22}^d, \beta_{33}^d, \beta_{12}^d, \beta_{13}^d, \beta_{23}^d, \beta_{22}^g, \beta_{12}^g, \xi^a\right)^\top,$$

where β_0 represents the log of mean within-group Beta-diversity for the reference AH group, β_{kk}^d represent the log of mean within-group Beta-diversity differences for AUD $(k=2)$ and HC $(k=3)$ with the AH $(k=1)$, and β_{kl}^d represent the log of mean differences of the respective between-group Beta-diversity of AH and AUD $\left(\beta_{12}^d\right)$, AH and HC $\left(\beta_{13}^d\right)$, AUD and HC $\left(\beta_{23}^d\right)$ compared with the AH, β_{22}^g $\left(\beta_{12}^g\right)$ represents the log of mean difference of Beta-diversity comparing female-female (male-female) and the reference male-male pairs, and ξ^a represents

the change in the log of mean Beta-diversity per unit increase in age difference (measured by Euclidean distance). Given the relatively small sample sizes for AUD and HC, we report both Wald and score results, as well as Bootstrap results (based on $5,000$ Bootstrap samples) to assess the accuracy of asymptotic results.

The top of Table 5 shows estimates, standard errors (asymptotic under "A. SE" and Bootstrap under "B. SE"), test statistics and p-values (Wald under "W. p", score under "S. p", Bootstrap Wald under "B.W. p" and Bootstrap score under "B.S. p") for the nulls. All Bootstrap standard errors were close to their asymptotic counterparts. For each hypothesis, the test results were consistent, except for a noticeable discrepancy of the score test for β_{33}^d due to the small sample size of HC group ($n_3 = 11$).

AUD had no significant within-group difference in mean diversity compared with the AH ($\widehat{\beta}_{22}^d = .226$, p-values range $[.419, .662]$), but HC had a significantly higher within-group diversity than the AH from Wald test ($\widehat{\beta}_{33}^d = .572$, W. p $= .002$), which is consistent with Figure 1. While the score test for β_{33}^d revealed that more evidence needed to be collected to reject the null (S. p $= .130$), this discrepancy may be due to the small sample size of HC. However, after Bootstrapping, both Wald and score were consistently significant for β_{33}^d (B.W. p $= .007$, B.S. p $< .0001$). All the above results reveal the scientific finding that alcoholic liver disease is associated with reduced microbial diversity. For covariates, age had a positive effect with $\widehat{\xi}^a = .006$, both female-female ($\widehat{\beta}_{22}^g = .125$) and male-female ($\widehat{\beta}_{12}^g = .072$) pairs had higher mean diversity than male-male pairs. None of the covariates were significant.

The bottom of Table 2.5 includes statistics and p-values. The null of no within-group difference ($H_{01} : \beta_{22}^d = \beta_{33}^d = 0$) was rejected consistently by Wald (W. p $= .007$) and two bootstrap tests (B.W. p $= .017$, B.S. p $< .0001$), while the score test was close to being significant with S. p $= .071$, suggesting a larger sample size may be needed to confirm significance. The null of no between-group difference ($H_{02} : \beta_{12}^d = \beta_{13}^d = \beta_{23}^d$) across the three groups was rejected by all tests with the p-values ranging in $(.0001, .001]$.

Table 2.5. Estimates, asymptotic standard errors (A. SE), Bootstrap standard errors (B. SE) based on $B = 5,000$ Bootstrap samples, Wald statistics, Score statistics, Wald p-valules (W. p), Score p-values (S. p), Bootstrap Wald p-valules (B.W. p) and Bootstrap Score p-valules (B.S. p) for the real study data using FRM, including covariates.

Categorical Covariate: Gender (β^g), Continuous Covariate: Age (ξ^a)									
Param.	Est.	Std. err		Statistic		p-value			
		A. SE	B. SE	Wald	Score	W. p	S. p	B.W. p	B.S. p
β_0	-1.04	.215	.23	23.49	13.63	$<$.001	.0002	$<$.001	$<$.001
β_{22}^d	.226	.302	.290	.560	.442	.454	.506	.419	.662
β_{33}^d	.572	.186	.201	.416	2.29	.002	.130	.007	$<$.001
β_{12}^d	.114	.193	.174	.350	.331	.554	.565	.519	.674
β_{13}^d	.634	.173	.183	13.41	7.46	$<$.001	.006	.002	$<$.001
β_{23}^d	.672	.180	.190	14.00	5.41	$<$.001	.020	.001	$<$.001
β_{22}^g	.125	.189	.175	.436	.399	.509	.528	.477	.613
β_{12}^g	.072	.121	.111	.357	.356	.550	.551	.511	.583
ξ^a	.006	.005	.005	1.72	1.48	.189	.224	.184	.348

	Hypothesis	Statistic		p-value			
		Wald	Score	W. p	S. p	B.W. p	B.S. p
Within	$H_{01} : \beta_{22}^d = \beta_{33}^d = 0$	9.86	5.30	.007	.071	.017	$<$.001
Between	$H_{02} : \beta_{12}^d = \beta_{13}^d = \beta_{23}^d$	19.01	28.48	$<$.001	$<$.001	.001	$<$.001
Within vs	$H_{03}^{(1)} : \beta_{12}^d = 0$	.350	.331	.554	.565	.519	.674
Between	$H_{03}^{(2)} : \beta_{13}^d = 0$	13.41	7.46	$<$.001	.006	.002	$<$.001
	$H_{03}^{(3)} : \beta_{23}^d = 0$	14.00	5.41	$<$.001	.020	.001	$<$.001
	$H_{04a} : \xi^a = 0$	1.72	1.48	.189	.224	.184	.613
Covariate	$H_{04b}^{(1)} : \beta_{22}^g = 0$	.436	.399	.509	.528	.477	.583
	$H_{04b}^{(2)} : \beta_{12}^g = 0$	.357	.356	.550	.551	.511	.348
	$H_{04b} : \beta_{22}^g = \beta_{12}^g = 0$	.621	.241	.733	.886	.732	1.00

The between- vs. within-group differences were significant for between-group variability of D-HC and within-group variability of AH-AH pairs: with p-values ranging in $(.0001, .006]$ for $H_{03}^{(2)} : \beta_{13}^{d} = 0$ (AH-HC vs. AH-AH) and $(.0001, .020]$ for $H_{03}^{(3)} : \beta_{23}^{d} = 0$ (AUD-HC vs. AH-AH). However, there was no evidence to reject $H_{03}^{(1)} : \beta_{12}^{d} = 0$ concerning the between-group variability of AUD-AH vs. within-group variability of AH-AH pairs. There was no significant difference across the three gender pair groups (p-values range in $[.732, 1]$).

The results above were not corrected for multiple comparisons. We also provide FDR corrected results in the Supporting Information by applying the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995) to control the family-wise FDR at 5%, where major conclusions remained unchanged except for $H_{03}^{(3)} : \beta_{23}^{d} = 0$ (AUD-HC vs. AH-AH), the score test p-value (S. p) was .020 before and .060 after correction.

In summary, both within- and between-group hypotheses detected group differences, driven by the fact that the HC group was rather distinct from the two disease groups. While the within- vs. between-group hypotheses enabled a more comprehensive comparison, the difference between AH-AH and AUD-AUD pairs was not as pronounced, yet any pair involving one subject from HC was significantly different from AH-AH pairs. These specific conclusions underscore the advantages of partitioning the sources of variation under the FRM.

2.5 Discussion

We developed a new approach to model Beta-diversity utilizing the functional response models (FRM). Unlike conventional approaches such as the PERMANOVA, the proposed FRM can disentangle information carried by Beta-diversity flexibly with the unique interpretations of "mean within-group diversity" for each group and the "mean between-group diversity" between any two groups. This regression approach also provides coefficient estimators for explanatory variables, generating interpretable results on both the direction and size of the effects and leading to more in-depth scientific findings.

In addition, the proposed approach carries far fewer overheads than PERMANOVA in terms of the computational burden. Also, the semiparametric nature of the model enables valid inferences without any parametric assumption on the correlated and non-negative Beta-diversity. Lastly, the approach to simulate life-like OTUs and Beta-diversity allows one to relate simulation study results directly to the performance of the proposed and other statistical models for such data in real studies.

Comparing with other methods for multivariate responses to improve inference of the mean response such as the covariance regression model (Hoff and Niu, 2012), the proposed approach aims to directly model the relationships between Beta-diversity, a complex yet biologically meaningful between-subject attribute, and a set of explanatory variables, which can be within-, between-subject or both, as deemed appropriate by content experts. Also, FRM's ability to control for between-subject confounders, such as a dissimilarity measure comparing subjects' metabolites abundance profile, makes it particularly useful in certain circumstances involving such confounders. Given some recent discussions (Morton et al., 2019) regarding the confounding of sequencing depth, one potential issue in most compositional data analysis is the stochastic nature of sampling reads due to technical variation, yielding a potential confounding effect. If this is the case in some applications, we can alleviate it by modeling Beta-diversity from the absolute abundance (instead of relative abundance) and including the sampling depth as an offset term in the proposed model.

In practice, we suggest conducting both score and Wald tests in applying the proposed model. If the sample size for some groups is relatively small (for example, $n_k < 50$), an additional Bootstrap procedure is recommended. One major limitation of the approach is that it only applies to cross-sectional data. Currently, leveraging semiparametric regression models for longitudinal data, we are working on extending the approach to facilitate analyses of such data.

Chapter 2, in full, is a reprint of the material as it appears in Biometrics 2021. The dissertation author was the primary investigator and author of this paper. The co-authors include

Zhang, X, Chen, T, Wu, T, Lin, T, Jiang, L, Lang, S, Liu, L, Natarajan, L, Tu, JX, Kosciolek, T, Morton, J, Nguyen, TT, Schnabl, B, Knight, R, Feng, C, Zhong, Y, and Tu, XM.

Chapter 3

A Distance-based Semiparametric Regression Framework for Between-subject attributes: Applications to High-dimensional Sequences of Microbiome and Wearables

3.1 Introduction

Breakthroughs in innovative technologies such as next-generation sequencing are producing flourishing high-throughput data, which have evolved into the center stage of biomedical and other burgeoning research areas. This procedure brings forth data that is sparse and astronomically high-dimensional. For example, sequenced genomes range from hundreds to tens of thousands in human microbiome studies to determine their role in the pathogenesis of complex diseases (Nguyen et al., 2021a). Other pioneering technologies also originate high-dimensional sequences such as real-time measurements from wearables, which monitor patients' health objectively to a minute-by-minute frequency. Nevertheless, the sheer volume of such data also brings enormous statistical challenges, including modeling the astronomical dimensionality and the hard-to-track within- and between-subject variability.

Directly modeling such raw data is not only laborious due to the untenable model assumption and high dimensionality, but also suffers from multiple testing and low power. Although regularizations such as the least absolute shrinkage and selection operator (lasso) may

be applied (Tibshirani, 2011), it is quite daunting to find a few individual genomes that are fully responsible for the phenotype of a disease. In many situations, such individual culprits may simply not exist on the conceptual grounds (Chabris et al., 2013). Moreover, in most applications, researchers are intrigued by the overall effect of high-dimensional outcomes but lack a handy analytical tool. Therefore, an emerging alternative is to first reduce the data dimension at the outset by comparing two subjects' sequences using a dissimilarity/distance metric, termed "between-subject attributes", for a pair. We refer to their classical counterparts that concern only one individual as "within-subject attributes." Such distances are gaining popularity as effective dimension reductions and have been widely applied in various biomedical research such as single-cell RNA sequencing (Moon et al., 2017).

For statistical modeling, the fundamental framework pertains to the generalized linear model (GLM) for within-subject attributes. It encompasses non-normal and noncontinuous responses (dependent variables) to present a unified paradigm for different response types (Nelder and Wedderburn, 1972; Agresti, 2003a). By further removing the distributional assumption, the semiparametric GLM enables valid inference where the correct parametric inference is difficult, such as handling heteroscedastic error terms or over-dispersed counts. Despite its broad applicability, this GLM paradigm predominantly focuses on the relationships within the same subject from the raw data. But in the growing applications, of major interest are outcomes defined by a pair of subjects or the between-subject attributes.

Modeling such between-subject attributes has only recently become a central focus with the impetus from high-dimensional data (Liu et al., 2021). For example, methods have been proposed to compare the means of between-subject outcomes across different groups in microbiome diversity, such as the PERMANOVA (Aldous et al., 2012). Additionally, Mantel's test (Mantel, 1967) has been developed to determine the correlation between two matrices, extensions including multiple regression on distance matrices (MRM) can handle more than two matrices (Lichstein, 2007). However, limitations of these existing approaches include (1)

lacking a unified framework to accommodate different modeling goals; (2) inflexibility to adjust link functions confronting different data types; (3) tediousness and uncleared documentation to include factors as explanatory variables with the existing R package (Goslee and Urban, 2007); (4) the prevailing permutation-based inference is not only computational burdensome but also extremely difficult to implement, especially for distances entailing a dendrogram/tree structure.

Essentially, we aim to address these limitations by extending the classical GLM from within- to between-subject attributes to present a unified GLM-type regression framework for distance metrics. Akin to other analyses of distances, this timely solution provides inference about relationships between pairwise distances, instead of the raw data. We illustrate how to construct such between-subject distances with the motivating data from the human microbiome and mHealth.

3.2 Motivating Data

3.2.1 Human Microbiome

The human microbiota consists of the 10-100 trillion symbiotic microbial cells harbored by each person, primarily bacteria in the gut; the human microbiome consists of the genes these cells harbor (Turnbaugh et al., 2007). Recent studies have linked dysfunctions of the human microbiota to complex diseases ranging from diabetes to cancers, and even psychiatric disorders (Cho and Blaser, 2012; Nguyen et al., 2019).

Fueled by technological advances such as next-generation sequencing, the human microbiome can be interrogated using high-throughput sequencing (e.g., 16s sequencing of gut microbiota) to drive insights into disease mechanisms. This procedure generates taxonomic sequence counts (for each subject) that are sparse and astronomically high-dimensional (e.g., in our data application, the dimension $m = 12,131$). Due to their additional sparsity and non-normality, "diversity" is introduced to further summarize the microbial at a community level. In fact, most researchers start with the community-level microbiome diversity analysis instead of

directly tackling the raw data that often suffers from weak signals.

This biologically-relevant microbiome diversity is a critical indicator of human health (Durack and Lynch, 2019a). For example, a popular metric that naturally encompasses a between-subject nature is the Beta-diversity, defined by pairwise distances of taxonomic counts. For example, consider a sample of n subjects, let $\mathbf{x}_i \in \mathbb{R}^m$ denote a column vector of relative abundance (proportions) of taxonomic units for the i-th subject, the Aitchison Beta-diversity (Aitchison, 1989) between any pair $(i_1, i_2) \in C_2^n$ is

$$d_A(\mathbf{x}_{i_1}, \mathbf{x}_{i_2}) = \left[\sum_{k=1}^{m} \left\{ \log \frac{x_{i_1 k}}{g(\mathbf{x}_{i_1})} - \log \frac{x_{i_2 k}}{g(\mathbf{x}_{i_2})} \right\}^2 \right]^{1/2}, \; g(\mathbf{x}_i) = \left(\prod_{k=1}^{m} x_{ik} \right)^{1/m}, \tag{3.1}$$

where C_q^n denotes the set of q-combinations $\left(i_1, \ldots, i_q\right)$ from the integer set $\{1, \ldots, n\}$, $g(\mathbf{x}_i)$ is the geometric mean of $\mathbf{x}_i$.

By integrating information from high-dimensional omics data for each individual, Beta-diversity represents a totality measure of dissimilarity between two subjects across all (or a proportion of) the sequenced genomes, which merits its interest.

3.2.2 mHealth Studies

The versatility of the between-subject attributes is by no means confined to one discipline; in fact, the notion of pairwise distances can be migrated to the blooming real-time longitudinal sequences collected from wearables.

In the digital era for disease diagnosis, treatment, and prevention, valuable data generated from mHealth studies can facilitate personalized interventions to improve patient care. While to date, the data analyses are still in an embryonic stage dominated by descriptive statistics. Even if modern methods such as (multilevel) functional principal component analysis (FPCA) (Di et al., 2009) or penalized multi-band learning (Li et al., 2021) have been proposed, they may be subjected to information loss due to the selected principal components or penalization.

In some studies, in addition to inspecting population mean function over time (as

achieved with functional PCs), investigators also aim to capture the variability or heterogeneity of activities among subgroups. Pairwise distances of the high-dimensional sequences can naturally capture the between-subject variability. For example, the mean of squared Euclidean distance pertains to the variance. Hence, they can potentially help unravel the intricate connections among sleep, physical activity, and mental traits.

3.3 Semiparametric Regression for Distances

3.3.1 Functional Response Models

Consider a study with n subjects, let y_i denote a response, $\mathbf{x}_i \in \mathbb{R}^p$ a column vector of explanatory variables for the i-th subject. As a motivating example, the semiparametric GLM characterizing the relationship between y_i and $\mathbf{x}_i$ is:

$$E(y_i \mid \mathbf{x}_i) = h(\mathbf{x}_i; \beta), \quad 1 \leq i \leq n, \tag{3.2}$$

where $h(\cdot)$ is the inverse of some link functions (Tang et al., 2012a). Compared with the classical parametric GLM, (3.2) is more flexible as it removes the distributional assumption on y_i thus yields valid inference even when the data deviate from such an assumption.

Under suitable regularity conditions, the "sandwich" estimators from the generalized estimating equations (GEE) are consistent and asymptotically normal. They have also been shown to achieve the semiparametric efficiency bound (Tsiatis, 2007). This semiparametric efficiency allows a sensitive signal-detection in practice while simultaneously harmonizing robustness to model misspecification.

However, one limitation is that it does not apply to the between-subject, or pairwise, distances that are of interest in a mounting number of applications. Hence, we adopt an alternative paradigm involving a functional response of multiple (q) subjects, and develop the

semiparametric framework of functional response models (FRM) (Kowalski and Tu, 2008a):

$$E\left\{\mathbf{f}\left(y_{i_1},\ldots,y_{i_q}\right) \mid \mathbf{x}_{i_1},\ldots,\mathbf{x}_{i_q}\right\} = \mathbf{h}\left(\mathbf{x}_{i_1},\ldots,\mathbf{x}_{i_q};\beta\right), \tag{3.3}$$
$$\left(i_1,\ldots,i_q\right) \in C_q^n, \quad i_1 < \ldots < i_q, \quad q \geq 1,$$

where $\mathbf{f}(\cdot)$ is some vector-valued function, $\mathbf{h}(\cdot)$ is some vector-valued smooth function (e.g. with continuous derivatives up to the second-order), β is a vector of parameters, q is a positive integer. Akin to (3.2), the above is also semiparametric without any distributional assumption on the response function $\mathbf{f}\left(y_{i_1},\ldots,y_{i_q}\right)$.

We now readily implement (3.3) to model the pairwise distances. For notational consistency, we use i to index a subject and $\mathbf{i} = (i_1, i_2) \in C_2^n$ to index a pair in what follows.

For a sample of size n, we observe raw data $(\mathbf{y}_i^\top, \mathbf{x}_i^\top)$, where $\mathbf{y}_i(\mathbf{x}_i) \in \mathbb{R}^s(\mathbb{R}^l)$ $(s, l \geq 1)$ is a column vector of multivariate response (explanatory variable) for the i-th subject. To address the high-dimensionality of $\mathbf{y}_i(\mathbf{x}_i)$, we can, at the outset, compact their information by constructing respective pairwise distances. Specifically, we create $d_{\mathbf{i}}^y = d\left(\mathbf{y}_{i_1}, \mathbf{y}_{i_2}\right)$ termed the "pairwise response", and $d_{\mathbf{i}}^x = d\left(\mathbf{x}_{i_1}, \mathbf{x}_{i_2}\right)$ as the "pairwise explanatory variable". Now with $d_{\mathbf{i}}^y$ and $d_{\mathbf{i}}^x$ the new modeling units, we take $q = 2$ in (3.3) and characterize the relationship between $d_{\mathbf{i}}^y$ and $d_{\mathbf{i}}^x$ with

$$E\left(d_{\mathbf{i}}^y \mid d_{\mathbf{i}}^x\right) = h\left(d_{\mathbf{i}}^x;\beta\right), \quad \mathbf{i} \in C_2^n. \tag{3.4}$$

Now (3.4) extends the classical GLM from within- to between-subject attributes using pairwise distances that are generally tricky to model. It not only achieves effective dimension-reduction but also establishes a complementing angle to reveal unexplored scientific findings, especially for data entailing an intrinsic between-subject nature.

Note again that the primary interest is to delineate the relationships between distance metrics $(d_{\mathbf{i}}^y, d_{\mathbf{i}}^x)$ that receive growing attentions, instead of between the raw data $(\mathbf{y}_i^\top, \mathbf{x}_i^\top)$. Next, we illustrate how to construct such between-subject distances. The applicability of this paradigm

far surpasses the limited settings we demonstrate, but our goal is to unify this framework to enlighten further constructive implementations.

3.3.2 Constructing Pairwise Response $d_{\mathbf{i}}^{y}$

Univariate Outcome y_i

Continuous Data

Consider a motivating linear regression for y_i and x_i that are both continuous. By defining "difference indices" $d_{\mathbf{i}}^{y} = y_{i_1} - y_{i_2}$, $d_{\mathbf{i}}^{x} = x_{i_1} - x_{i_2}$, we obtain $E\left(y_{i_1} - y_{i_2} \mid x_{i_1}, x_{i_2}\right) = \beta\left(x_{i_1} - x_{i_2}\right), \quad (i_1, i_2) \in C_2^n$.

In real data applications with intrinsic between-subject explanatory variables $d_{\mathbf{i}}^{x}$, such a construction evaluates the association between $d_{\mathbf{i}}^{y}$ and $d_{\mathbf{i}}^{x}$, in a form of "differential response." For instance, let $d_{\mathbf{i}}^{x} = d(\mathbf{x}_{i_1}, \mathbf{x}_{i_2})$ denote the microbiome Beta-diversity for the $\mathbf{i}$-th pair, if we are interested in its effect on $d_{\mathbf{i}}^{y} = y_{i_1} - y_{i_2}$ (such as BMI difference), we have

$$E\left(d_{\mathbf{i}}^{y} \mid d_{\mathbf{i}}^{x}\right) = \beta d_{\mathbf{i}}^{x}, \quad \mathbf{i} = (i_1, i_2) \in C_2^n. \tag{3.5}$$

One glitch is that while the Beta-diversity $d_{\mathbf{i}}^{x}$ is non-negative, $d_{\mathbf{i}}^{y}$ here can be positive or negative. This is readily fixed by setting $d_{\mathbf{i}}^{x}$ to $d_{\mathbf{i}}^{x} sign(\mathbf{i})$, where $sign(\mathbf{i})$ denotes the sign function with $sign(\mathbf{i}) = 1$ if $i_1 - i_2 > 0$, $sign(\mathbf{i}) = -1$ if $i_1 - i_2 < 0$, and $sign(\mathbf{i}) = 0$ otherwise. For conciseness, we continue to denote $d_{\mathbf{i}}^{x} sign(\mathbf{i})$ by $d_{\mathbf{i}}^{x}$ in what follows.

Now for (3.5), $|\beta|$ represents the differential response $d_{\mathbf{i}}^{y}$ per unit difference in the Beta-diversity $d_{\mathbf{i}}^{x}$ for the $\mathbf{i}$-th pair, its sign does not yield a meaningful interpretation. Although positing the sign function is not the only way to handle this, its advantage surfaces when including additional continuous explanatory variables (see Section 3.3.3).

Binary Data

Let y_i denote a binary response such as a disease indicator, where $y_i = 1\ (0)$ if diseased (otherwise). For the $\mathbf{i}$-th pair, four pairwise responses exist: $y_{i_1} = y_{i_2} = 1$ (both diseased),

$y_{i_1} = y_{i_2} = 0$ (both healthy), and $y_{i_1} = 0(1)$, $y_{i_2} = 1(0)$ (discordant). It is easily deduced from the independence (among individuals) that only discordant response pairs are associated with the differences in their explanatory variables. Hence, by only considering this subset of pairs, we construct a new binary pairwise response $d^y_{\mathbf{i}} = 1$ if $y_{i_1} = 1$, $y_{i_2} = 0$, and $d^y_{\mathbf{i}} = 0$ otherwise. Then define an FRM that relates $d^x_{\mathbf{i}}$ to $d^y_{\mathbf{i}}$:

$$E\left(d^y_{\mathbf{i}} \mid d^x_{\mathbf{i}}\right) = h\left(\beta d^x_{\mathbf{i}}\right) = \text{expit}\left(\beta d^x_{\mathbf{i}}\right), \quad \mathbf{i} = (i_1, i_2) \in C^n_2, \tag{3.6}$$

where $\text{expit}(\rho) = \exp(\rho)/\{1+\exp(\rho)\}$, $|\beta|$ is the log odds ratio of $d^y_{\mathbf{i}} = 1$ to $d^y_{\mathbf{i}} = 0$ per unit difference in $d^x_{\mathbf{i}}$. Akin to the continuous case, its sign has no meaningful interpretation.

The reason for considering the subset in (3.6) is similar to modeling paired binary outcomes in McNemar's test and conditional logistic regression (Agresti, 2003a). In essence, (3.6) is motivated by $d^x_{\mathbf{i}}$ indexing distances/differences, pairs with concordant responses provide no information for the relationship of interest. But under more general settings outside of the difference representations (e.g., $d^x_{\mathbf{i}} = x_{i_1} + x_{i_2}$), the FRM can be adapted case-by-case.

Count Data

The semiparametric GLM for a count y_i adopts the log-linear form to ensure that $E(y_i \mid x_i)$ falls in the appropriate range between 0 and ∞. Now consider the difference index $d^y_{\mathbf{i}} = y_{i_1} - y_{i_2}$ with the range $(-\infty, \infty)$, unlike the original y_i, here $d^y_{\mathbf{i}}$ no longer has the range restriction. Therefore, we can continue to model $d^y_{\mathbf{i}}$ with the identity link as in the continuous case with (3.5).

Multivariate Outcome $\mathbf{y}_i$

One merit of extending from within- to between-subject regression is that the distanced-based FRM handles multivariate outcomes $\mathbf{y}_i$ in a neat but inclusive way.

In mHealth studies, multiple intensive longitudinal measurements are recorded, along with some clinical phenotypes. Consider a sample of size n, for each subject i, we mon-

itor P explanatory variables over a period of T measurements. Denote this raw data by $(\mathbf{y}_i^\top, \mathbf{x}_i^{1\top}, ..., \mathbf{x}_i^{P\top})^\top$, where $\mathbf{y}_i(\mathbf{x}_i^p) \in \mathbb{R}^T$ is the outcome (p-th explanatory variable) $(p = 1, ..., P)$. The dimension of the observed (3-D) raw data is $n \times P \times T$, either P or T can be huge, provoking enormous challenges to demystify scientific findings. We now illustrate a feasible solution to first reduce the time dimension T.

For each pair $\mathbf{i}$, we integrate the information from each individual into pairwise measures to achieve the first-layer dimension reduction. Particularly, we can construct the pairwise responses $d_{\mathbf{i}}^y = d(\mathbf{y}_{i_1}, \mathbf{y}_{i_2})$ and respective explanatory distances $d_{\mathbf{i}}^{x_p} = d(\mathbf{x}_{i_1}^p, \mathbf{x}_{i_2}^p)$ for each $\mathbf{x}_i^p$ $(p = 1, ..., P)$. It works out favorably to reduced the data to a 2-D dimension of $n \times P$. If P is still large, further reductions may be needed. Fortunately, this unified GLM-type regression framework facilitates selecting predictors: either forward, backward, or stepwise selection is straightforward. A modern penalization-based approach is also suitable with special constructions during the inference (Kowalski et al., 2018).

Now the FRM depicting the relationship between $d_{\mathbf{i}}^y$ and $d_{\mathbf{i}}^{x_p}$ $(p = 1, ..., P)$ is

$$E\left(d_{\mathbf{i}}^y \mid \mathbf{d}_{\mathbf{i}}^x\right) \quad = \quad h\left(\beta^\top \mathbf{d}_{\mathbf{i}}^x\right) = \exp(\beta^\top \mathbf{d}_{\mathbf{i}}^x), \quad \mathbf{i} = (i_1, i_2) \in C_2^n, \tag{3.7}$$

where $\mathbf{d}_{\mathbf{i}}^x = \left(d_{\mathbf{i}}^{x_1}, ..., d_{\mathbf{i}}^{x_P}\right)^\top$, $\beta = (\beta_1, ..., \beta_P)^\top$, $\exp(\cdot)$ is adopted since $d_{\mathbf{i}}^y$ is non-negative. In (3.7), β_p is the difference in $d_{\mathbf{i}}^y$ per unit difference in $d_{\mathbf{i}}^{x_p}$. Interactions or non-linear effects such as quadratic terms can also be added to $h(\cdot)$.

For selecting the distance metric, we suggest scrutinizing features of raw data and involving content experts, some distances accounting for the phylogenetic structure (Lozupone et al., 2011) may be deemed appropriate. As a starting point, the squared Euclidean distance can naturally capture the variability since $E\left\{(y_{i_1} - y_{i_2})^2\right\} = 2\sigma^2$ if $y_i \sim N(\mu, \sigma^2)$. Other distances such as Wasserstein (Lin et al., 2021) also receive increasing attentions.

In a nutshell, by modeling several pairwise distances simultaneously, our proposed framework can potentially shed light on a variety of fields. For instance, in microbiome studies,

to explore the complex interplay of diet, microbiome, and metabolome in disease phenotypes; in mHealth and epidemiology studies, to connect disease status with sleep quality, physical activity, and even social networks; in genetic studies, to investigate interactions between genetic and environmental factors (G×E interaction) in disease development, etc.

3.3.3 Constructing Pairwise Explanatory Variable $d_{\mathbf{i}}^{x}$

Univariate x_i

Continuous or Count Type

Continuous or count x_i can be approached similarly as in Section 3.3.2, for example, by creating $d_{\mathbf{i}}^{x} = x_{i_1} - x_{i_2}$. Moreover, if we also adopt $d_{\mathbf{i}}^{y} = y_{i_1} - y_{i_2}$ as the pairwise response, the coefficient β in the FRM represents the directional difference between y_{i_1} and y_{i_2} per unit difference between x_{i_1} and x_{i_2}.

Now consider again the model in (3.5) with a non-directional $d_{\mathbf{i}}^{x_1}$ but appended with a sign function $sign(\mathbf{i})$, then adding more continuous pairwise covariates $d_{\mathbf{i}}^{x_p} = x_{i_1}^{p} - x_{i_2}^{p}$ $(p = 2, ..., P)$ will appreciably preserve the sign interpretations for β_p.

Categorical Type

For many studies, the major interest is to compare characteristics of $d_{\mathbf{i}}^{y}$ among subgroups, where FRM is desirable to discern sources of variation. We start with a categorical variable w_i with K levels. To transform w_i to a between-subject attribute for the $\mathbf{i}$-th pair, we define a set of pairwise indicators (or dummy variables) for $\mathbf{w_i} = \{w_{i_1}, w_{i_2}\}$ through the one-hot encoding function $\delta(\cdot) : \{1, ..., K\} \times \{1, ..., K\} \mapsto \{0, 1\}^{K + C_2^K}$:

$$\delta_{k_1 k_2}(\mathbf{w_i}) = \begin{cases} 1, & \text{if } \mathbf{w_i} = \{w_{i_1}, w_{i_2}\} = \{k_1, k_2\} \ , \\ 0, & \text{otherwise.} \end{cases} \tag{3.8}$$

$$\delta(\mathbf{w_i}) = \left(\delta_{11}(\mathbf{w_i}), \ldots, \delta_{(K-1)K}(\mathbf{w_i}), \delta_{KK}(\mathbf{w_i})\right)^{\top}, \quad 1 \le k_1 \le k_2 \le K.$$

where the vector $\delta(\mathbf{w_i}) \in \mathbb{R}^{K+C_2^K}$ denotes all combinations. Thus, $\delta_{k_1k_2}(\mathbf{w_i})$ indicates the pair with the same kth concordant ($k_1 = k_2 = k$) or discordant ($k_1 < k_2$) levels for $\mathbf{w_i}$.

For example, if w_i is gender, we form $\delta(\mathbf{w_i}) = (\delta_{FF}(\mathbf{w_i}), \delta_{MM}(\mathbf{w_i}), \delta_{MF}(\mathbf{w_i}))^\top$, where $\delta_{FF}(\mathbf{w_i})$ and $\delta_{MM}(\mathbf{w_i})$ index female-female and male-male pairs, and $\delta_{MF}(\mathbf{w_i})$ represents the mixed male-female pairs. By selecting one as the reference, we can add other levels to the linear predictors of the FRM in (3.7). The sign function can again resolve the situation when we choose a response $d_{\mathbf{i}}^y$ that is directional.

The coefficients of the dummy variables now reveal the heterogeneity in $d_{\mathbf{i}}^y$ among different groups defined by $\delta(\mathbf{w_i})$. Such a pairwise one-hot encode also facilitates disentagling different types of heterogeneity (such as "location" or "scale" difference) (Liu et al., 2021), which is laborious or not even feasible using existing approaches such as PERMANOVA.

Multivariate $\mathbf{x}_i$ or Inherent Between-subject $d_{\mathbf{i}}^x$

In addition to adjusting for covariates constructed from univariate x_i, FRM surpasses other comparable methods by its ease to incorporate multivariate $\mathbf{x}_i$ or inherent $d_{\mathbf{i}}^x$. Consider a distance $d_{\mathbf{i}}^x$ for the $\mathbf{i}$-th pair, either directly observed (such as the adjacency matrix of a social network) or constructed. We can readily add it to (3.7) to evaluate its effect on $d_{\mathbf{i}}^y$, or add more complex non-linear terms. For example, in microbiome studies, it is key to control for covariates such as dissimilarities comparing subjects' metabolites abundance profile. In studies investigating their interplay with the microbiome on a disease phenotype, FRM is even more suitable (e.g., $E\left(d_{\mathbf{i}}^y \mid d_{\mathbf{i}}^{x_1}, d_{\mathbf{i}}^{x_2}\right) = \exp(\beta_1 d_{\mathbf{i}}^{x_1} + \beta_2 d_{\mathbf{i}}^{x_2} + \beta_{12} d_{\mathbf{i}}^{x_1} \cdot d_{\mathbf{i}}^{x_2})$).

With blooming implementations of the between-subject attributes as effective dimension-reduction tools, this unified FRM framework will propel growing data-driven understandings. Nevertheless, the next challenge is to properly address the dependencies among pairs.

3.3.4 Inference

As the response function $d_{\mathbf{i}}^{y}$ in (3.7) involves pairs, we must address their interlocking dependent relationships for the inference of β. Classical asymptotic properties such as the central limit theorem (CLT) rely heavily on the assumption of independence, which precludes direct implementations to our correlated functional responses. Instead, we adopt a class of U-statistics-based generalized estimating equations (UGEE) (Kowalski and Tu, 2008a) to address such correlations. For a distance-based FRM in (3.7), let

$$\begin{aligned} S_{\mathbf{i}}(\beta) &= d_{\mathbf{i}}^{y} - h_{\mathbf{i}}(\beta), \quad \mathbf{D}_{\mathbf{i}} = \frac{\partial}{\partial \beta} h_{\mathbf{i}}(\beta), \quad V_{\mathbf{i}} = Var\left(d_{\mathbf{i}}^{y} \mid d_{\mathbf{i}}^{x}\right), \\ \mathbf{U}_{n,\mathbf{i}}(\beta) &= \mathbf{D}_{\mathbf{i}} V_{\mathbf{i}}^{-1} S_{\mathbf{i}}(\beta), \quad \mathbf{i} = (i_1, i_2) \in C_2^n. \end{aligned} \tag{3.9}$$

In practice, the unknown $V_{\mathbf{i}}$ is substituted by a working variance.

Despite (3.9) a comparable form as a previous discussion on inference for extending ANOVA to semiparametric regression for microbiome Beta-diversity in Chapter 2, special adjustments are needed here to accommodate the directional issue. Particularly, by symmetrizing $\mathbf{U}_{n,\mathbf{i}}(\beta)$ for the pair $\mathbf{i} = (i_1, i_2)$ using its mirror pair $\widetilde{\mathbf{i}} = (i_2, i_1)$, we obtain

$$\widetilde{\mathbf{U}}_{n,\mathbf{i}}(\beta) = \frac{1}{2}\left\{\mathbf{U}_{n,\mathbf{i}}(i_1, i_2) + \mathbf{U}_{n,\widetilde{\mathbf{i}}}(i_2, i_1)\right\}, \tag{3.10}$$

and the estimating equations summing over all C_2^n pairs are

$$\mathbf{U}_n(\beta) = \sum_{\mathbf{i} \in C_2^n} \widetilde{\mathbf{U}}_{n,\mathbf{i}}(\beta) = \mathbf{0}. \tag{3.11}$$

Given β, $\mathbf{U}_n(\beta)$ is a multivariate U-statistic that enjoys asymptotic normality under mild regularity conditions, (3.11) is hence a set of U-statistics-based generalized estimating equations (UGEE) that are uniquely-identified. Although similar in appearance, they generalize beyond the conventional GEE, as it is no longer a sum of independent random vectors. For example, for

pairs $\mathbf{i} = (i_1, i_2)$ and $\mathbf{i}' = (i_1, i_3)$, $i_2 \neq i_3$, $\widetilde{\mathbf{U}}_{n,\mathbf{i}}$ and $\widetilde{\mathbf{U}}_{n,\mathbf{i}'}$ are not independent, since both involve the same observations from subject i_1.

Now the inference follows with the projection of U-statistics. Let $\widehat{\beta}$ denote the estimator from solving (3.11), its asymptotic properties are summarized below.

Theorem 3.1 Let

$$\mathbf{v}_{i_1} = E\left\{\widetilde{\mathbf{U}}_{n,\mathbf{i}}(\beta)\,|y_{i_1}, x_{i_1}\right\}, \quad \mathbf{B} = E\left(\mathbf{D}_{\mathbf{i}} V_{\mathbf{i}}^{-1} \mathbf{D}_{\mathbf{i}}^{\top}\right), \tag{3.12}$$
$$\Sigma_U = 4var(\mathbf{v}_{i_1}), \quad \Sigma_\beta = \mathbf{B}^{-1}\Sigma_U\mathbf{B}^{-1}, \quad \mathbf{i} = (i_1, i_2) \in C_2^n.$$

Under mild regularity conditions,

(a) $\widehat{\beta}$ is consistent and asymptotically normal: $\sqrt{n}\left(\widehat{\beta} - \beta\right) \to_d N\left(0, \Sigma_\beta\right)$, where $\to_d$ denotes convergence in distribution.

(b) A consistent estimator of Σ_β is obtained by substituting consistent estimators of β and moments of the respective quantities in Σ_β.

Paralleling the robustness of GEE, UGEE also yields correct inference without explicitly specifying the sparse correlations among pairs. The added complexity of UGEE for pairs is readily addressed by leveraging the nice large-sample behaviors of U-statistics. Although $\widetilde{\mathbf{U}}_{n,\mathbf{i}}(\beta)$ are correlated, by projecting each onto a single subject i_1, we obtain the Hajek projection (Van der Vaart, 2000) $\mathbf{v}_{i_1}(\beta)$ that are now independent. And $\sum_{i_1=1}^{n} \mathbf{v}_{i_1}(\beta)$ yields the same asymptotic distribution as $\mathbf{U}_n(\beta)$ (Kowalski and Tu, 2008a), enabling CLT to be applied.

Theorem 1 readily permits testing the Neyman-Pearson-type hypotheses (Perezgonzalez, 2015) concerning β with the linear contrasts: $H_0 : \mathbf{C}\beta = 0$ vs. $H_a :\ \mathbf{C}\beta \neq 0$, where $\mathbf{C}$ is a matrix of known constants. Under the null, the Wald statistic has an asymptotic χ^2 distribution:

$$W_n = n\left(\mathbf{C}\widehat{\beta}\right)^{\top}\left(\mathbf{C}\Sigma_\beta\mathbf{C}^{\top}\right)^{-1}\left(\mathbf{C}\widehat{\beta}\right) \to_d \chi_s^2(0), \tag{3.13}$$

where s is the rank of $\mathbf{C}$, $\chi_s^2(0)$ denotes a central χ^2 distribution with s degrees of freedom.

3.4 Simulation Study

Our extensive simulation studies aim to demonstrate the asymptotic normality of the proposed semiparametric UGEE estimators, by comparing the discrepancy between asymptotic and empirical standard errors, along with correct type I error (or coverage) under the null.

It is quite often that between-subject attributes are sparsely correlated, where those pairs sharing at least one subject are correlated but other pairs are independent. Therefore, directly generating such pairwise outcomes is difficult, since it will not only break the interlocking correlation structure but also fail to resemble the real data.

Hence, we adopted an approach based on the empirical cumulative distribution function (eCDF) and copula (Liu et al., 2021), first generated $\mathbf{x}_i$ resembling the real microbiome taxonomic counts in Lang et al. (2020) and then computed Beta-diversity $d_{\mathbf{i}}^x$.

Without loss of generality, we set up the FRM regression under the null by including $d_{\mathbf{i}}^x$ as the main explanatory variable, but additionally control for a continuous $z_i \sim N(\mu_z, \sigma_z^2)$ and a binary $w_i \sim \text{Bernoulli}(p_w)$ $\left(i \in C_1^n\right)$. We then created their between-subject counterparts $d_{\mathbf{i}}^z$ and $\delta_{12}(w_{\mathbf{i}})$. We set $\mu_z = 5$, $\sigma_z^2 = 5$, $p_w = 0.75$ below. For Monte Carlo (MC) simulations, we set iterations $M = 1{,}000$, two-sided type I error rate $\alpha = 0.05$, and sample sizes $n = 150$, 200, 500. All analyses are performed with the R software platform, with code optimized using Rcpp (Eddelbuettel et al., 2011) for run-time improvement.

3.4.1 Continuous Univariate Outcome y_i

We start with a continuous univariate y_i. The proposed FRM becomes:

$$E\left(d_{\mathbf{i}}^y \mid d_{\mathbf{i}}^x, d_{\mathbf{i}}^z, w_{\mathbf{i}}; \beta\right) = h_{\mathbf{i}} = \beta_x d_{\mathbf{i}}^x + \beta_z d_{\mathbf{i}}^z + \beta_w \delta_{12}(w_{\mathbf{i}}). \tag{3.14}$$

To generate $d_{\mathbf{i}}^y$, we first simulated the within-subject error term $\varepsilon_i \sim N(\mu_\varepsilon, \sigma_\varepsilon^2)$ and created their difference $d_{\mathbf{i}}^\varepsilon = \varepsilon_{i_1} - \varepsilon_{i_2}$. The regression is hence created by setting $d_{\mathbf{i}}^y = h_{\mathbf{i}} + d_{\mathbf{i}}^\varepsilon$. To further

evaluate its robustness to misspecification of data distributions, we also simulated the error terms from a centered Chi-square with $\varepsilon_i \sim \left(\chi^2_d - d\right)\left(\sigma^2_\varepsilon/2\right)^{1/2}$, where χ^2_d denotes a Chi-square with d degrees of freedom.

We specified the parameters under the null β_0 as $\mu_\varepsilon = 0, \quad \sigma^2_\varepsilon = 1, \quad d = 1, \quad \beta_0 = \left(\beta^0_x, \beta^0_z, \beta^0_w\right)^\top = (1, 0.5, 0.5)^\top$. For the mth MC iteration, let $\widehat{\beta}^{(m)}$ and $\widehat{\Sigma}^{(m)}_\beta$ denote the estimator and its asymptotic variance, $\widehat{\beta}$ and $\widehat{\Sigma}^{(asymp)}_\beta$ denote their respective sample means. The sample variance of $\widehat{\beta}^{(m)}$ is denoted by $\widehat{\Sigma}^{(emp)}_\beta$. Let $W^{(m)}_n$ denote the Wald statistic in (3.13) for a hypothesis at the mth iteration. The Wald type I error is $\widehat{\alpha}^W = (1/M)\sum_{m=1}^M I\left(W^{(m)}_n \geq q_{s,0.95}\right)$, where $q_{s,0.95}$ denotes the 95^{th} percentile of a central χ^2_s distribution with s degrees of freedom.

We then assess the asymptotic performance by comparing asymptotic ($\widehat{\Sigma}^{(asymp)}_\beta$) and empirical variances ($\widehat{\Sigma}^{(emp)}_\beta$), also by comparing $\widehat{\alpha}^W$ ($\widehat{\alpha}^s$) with the nominal level $\alpha = 0.05$.

Shown in the top of Table 3.1 are results with normal errors: estimates were close to the truth, the asymptotic and empirical standard errors were close, and Type I error rates were close to the nominal values. The second panel of Table 3.1 are results with Chi-square error terms. Proper asymptotic performance was also observed, indicating that the UGEE approach worked quite well for both normal and non-normal data, thanks to its semiparametric nature.

3.4.2 Binary Univariate Outcome y_i

The FRM for a binary y_i can be specified by setting

$$d^y_{\mathbf{i}} = I(y_{i_1} = 1,\ y_{i_2} = 0), \quad h(\beta) = \text{logit}^{-1}\left\{\beta_x d^x_{\mathbf{i}} + \beta_z d^z_{\mathbf{i}} + \beta_w \delta_{12}(w_{\mathbf{i}})\right\}.$$

We simulated $d^y_{\mathbf{i}}$ for the pair $\mathbf{i} = (i_1, i_2) \in C^n_2$ by $d^y_{\mathbf{i}} \sim Bernoulli(h(\beta_0))$ under the null, where $\beta_0 = \left(\beta^0_x, \beta^0_z, \beta^0_w\right)^\top = (1, 0.5, 0.5)^\top$.

Shown in Table 3.2 are results for binary response that also demonstrate decent performances: all estimates were close to the truth, the asymptotic and empirical standard errors were close. Type I errors were also close to the nominal value.

Table 3.1. FRM with continuous responses (Normal and Chi-square residuals) under the null hypotheses, averaged over MC $M = 1,000$ iterations.

Normal residuals: Under Null Hypotheses				
Para.	Est.	Std. err		Type I Error
		Asymp.	Emp.	Wald
		$n = 150$		
β_x	0.9994	0.2835	0.2867	0.0500
β_z	0.5022	0.1606	0.1631	0.0540
β_w	0.5010	0.2528	0.2496	0.0510
		$n = 200$		
β_x	0.9973	0.2642	0.2648	0.0490
β_z	0.4986	0.1493	0.1500	0.0560
β_w	0.4996	0.2339	0.2280	0.0440
		$n = 500$		
β_x	1.0000	0.2102	0.2126	0.0520
β_z	0.5004	0.1187	0.1166	0.0430
β_w	0.4999	0.1849	0.1849	0.0550
Chi-square residuals: Under Null Hypotheses				
Para.	Est.	Std. err		Type I Error
		Asymp.	Emp.	Wald
		$n = 150$		
β	0.9974	0.3362	0.3336	0.0470
ξ_1	0.4973	0.1905	0.1934	0.0600
η_{12}	0.5004	0.3002	0.2982	0.0420
		$n = 200$		
β	1.0024	0.3142	0.3192	0.0470
ξ_1	0.5035	0.1780	0.1783	0.0440
η_{12}	0.4981	0.2780	0.2778	0.0440
		$n = 500$		
β	1.0031	0.2496	0.2500	0.0450
ξ_1	0.5003	0.1414	0.1404	0.0530
η_{12}	0.5005	0.2205	0.2216	0.0400

Table 3.2. MC estimates, standard errors (asymptotic and empirical) for FRM with binary responses under the null hypotheses, averaged over MC $M = 1,000$ iterations.

Binary outcomes: Under Null Hypotheses				
Para.	Est.	Std. err		Type I Error
		Asymp.	Emp.	Wald
		$n = 150$		
β_x	1.0015	0.1952	0.1987	0.0670
β_z	0.4999	0.0714	0.0707	0.0410
β_w	0.5011	0.1386	0.1364	0.0480
		$n = 200$		
β_x	0.9996	0.1673	0.1676	0.0440
β_z	0.4999	0.0616	0.0616	0.0410
β_w	0.4897	0.1196	0.1200	0.0580
		$n = 500$		
β_x	1.0005	0.1044	0.1049	0.0530
β_z	0.5000	0.0387	0.0400	0.0550
β_w	0.5001	0.0755	0.0762	0.0620

3.4.3 Multivariate Outcome $\mathbf{y}_i$

Now consider multivariate $\mathbf{y}_i \in \mathbb{R}^m$. Specifically, we chose $m = 10$ and generated $\mathbf{y}_i$ using the multivariate normal, then constructed $d^y_{\mathbf{i}} = d(\mathbf{y}_{i_1}, \mathbf{y}_{i_2})$ with the Euclidean distance. The proposed distance-based FRM under the null is

$$E\left(d^y_{\mathbf{i}} \mid d^x_{\mathbf{i}}, d^z_{\mathbf{i}}, w_{\mathbf{i}}\right) = h\left(\beta_0\right) = \exp\left\{\beta^0_0 + \beta^0_{w1}\delta_{22}\left(w_{\mathbf{i}}\right) + \beta^0_{w2}\delta_{12}\left(w_{\mathbf{i}}\right) + \beta^0_x d^x_{\mathbf{i}} + \beta^0_z d^z_{\mathbf{i}}\right\}, \quad (3.15)$$

where $\delta_{11}\left(w_{\mathbf{i}}\right)$ is designated as the reference level. We include the additional level $\delta_{22}\left(w_{\mathbf{i}}\right)$ to accommodate that the response is a Euclidean distance $d^y_{\mathbf{i}}$ concatenated from m dimensions. The coefficient β^0_{w1} can further discern concordant pairs in terms of $d^y_{\mathbf{i}}$.

********** Table 3 goes about here **********

The asymptotic behaviors are demonstrated in Table **??**. Again, all estimates were close to the truth $\beta_0 = (2.2618, 0.5, 0.5, 0.5, 0.5)^\top$, the asymptotic and empirical standard errors were close. Type I errors were also close to the nominal level.

3.5 Real Data Analyses

3.5.1 Microbiome Diversity

Recent studies suggest that the gut microbiome plays a major role in the development and functioning of the central nervous system via the "microbiome-gut-brain-axis" (Carabotti et al., 2015). In a recent study (Nguyen et al., 2021a), 184 participants across the lifespan (ages 20-100 years) provided fecal samples. DNA extraction and 16S rRNA amplicon sequencing were completed using the Earth Microbiome Project standard protocols, the dimension of the microbiome taxonomic units was quite high (m = 12,131). Instead of searching for individual signals accountable for the diseases of interest, investigators of the study were interested in the relationships between gut microbiome diversity and various psychological traits and social factors. We aim to investigate the role of Beta-diversity on clinical outcomes: (1) physical health, (2) mental health, and (3) positive states (traits) with the proposed FRM.

Continuous Univariate Outcome y_i

(1) and (2) were assessed by the self-administered standardized instruments based on the component scores from the Medical Outcomes Survey - Short Form 36 (SF-36) (Ware Jr and Sherbourne, 1992). Both are continuous with higher values indexing a better physical (mental) health condition. The FRM with Beta-diversity $d_{\mathbf{i}}^{x}$, Alpha-diversity difference $d_{\mathbf{i}}^{z1}$, age difference $d_{\mathbf{i}}^{z2}$, and gender $w_{\mathbf{i}}$ is $E\left(d_{\mathbf{i}}^{y} \mid d_{\mathbf{i}}^{x}, d_{\mathbf{i}}^{z1}, d_{\mathbf{i}}^{z2}, w_{\mathbf{i}}; \beta\right) = h_{\mathbf{i}} = \beta_x d_{\mathbf{i}}^{x} + \beta_{z1} d_{\mathbf{i}}^{z1} + \beta_{z2} d_{\mathbf{i}}^{z2} + \beta_w \delta_{12}\left(w_{\mathbf{i}}\right)$, where $d_{\mathbf{i}}^{y} = y_{i_1} - y_{i_2}$ denotes the difference of physical (mental) health for the $\mathbf{i}$-th pair.

Shown at the top of Table 3.4 are results of (1). The mean difference between physical health was $\widehat{\beta}_x = 0.214$ ($p = 0.018$) per unit difference in the Beta-diversity. The mean difference comparing physical health of male-female and homogeneous gender (male-male and female-female) pairs was $\left|\widehat{\beta}_w\right| = 0.005$ ($p = 0.835$). The mean (directional) difference between physical health was $\widehat{\beta}_{z1} = -0.057$ per unit difference in their Alpha-diversity, but not significant

Table 3.3. Estimates, asymptotic standard errors (Std. Error), Wald statistics, p-valules for the microbiome data using FRM, including continuous and categorical covariates.

Continuous outcome (Physical Health)				
Parameter	Est.	Std. Error	Statistic Wald	p-value Wald
β_x	0.2135	0.0905	5.5636	0.0183
β_{z1}	0.0575	0.0696	0.6823	0.4088
β_{z2}	-0.2300	0.0708	10.5571	0.0012
β_w	-0.0052	0.0254	0.0431	0.8355

Composite outcome (Positive traits/states)				
Parameter	Est.	Std. Error	Statistic Wald	p-value Wald
β_x	2.1210	0.1575	181.3488	$<$0.0001
β_{w1}	-0.1912	0.1762	1.1775	0.2779
β_{w2}	-0.1095	0.0815	1.8043	0.1792
β_{z1}	0.0292	0.0384	0.5804	0.4461
β_{z2}	0.1185	0.0618	3.6775	0.0552

($p = 0.409$). Per unit age difference was significantly associated with $\widehat{\beta}_{z2} = -0.230$ unit change in the mean physical health difference ($p = 0.001$).

Results for (2) mental health are in the **Supplement**, which also revealed an interesting finding that unlike physical health, mental health difference presented a significantly positive association with age difference ($p < 0.0001$).

Multivariate Outcome $\mathbf{y}_i$

In mental health studies, some traits are evaluated as a composite. For example, resilience, optimism, mental well-being, and wisdom are all (4) positive states (traits) measured by respective instruments (CD-RISC, LOTR, SF-36 and SD-WISE) in this dataset. The content experts intended to link such a multivariate outcome with microbiome, where the distanced-based FRM shines as it ideally handles the multivariate $\mathbf{y}_i$. For the ease of illustration, we constructed the composite outcome using the Euclidean distance $d^y_{\mathbf{i}} = \left\{ \sum_{k=1}^{4} \left(y^k_{i_1} - y^k_{i_2} \right)^2 \right\}^{1/2}$, and modified

$d_{\mathbf{i}}^{z1}$, $d_{\mathbf{i}}^{z2}$ using Euclidean distances to reflect variability, then adopted the FRM similar as (3.15) to elucidate their relationships.

The bottom of Table 3.4 are results of (4). The mean distance (variability) in positive states was significantly associated with microbiome Beta-diversity ($\widehat{\beta}_x = 2.12$, $p < 0.00001$); but not so with the variability in Alpha-diversity ($\widehat{\beta}_{z1} = 0.029$, $p = 0.446$). Age distance (variability) effect was only marginally significant on positive states ($\widehat{\beta}_{z2} = 0.118$, $p = 0.055$). The mean distance (variability) in positive states comparing male-male was 0.191 unit lower than female-female pairs ($p = 0.278$); male-female pairs was also 0.110 unit lower than female-female pairs ($p = 0.179$), but neither effect was significant.

These scientific insights uniquely driven by the proposed FRM may further lead to novel microbiota-related intervention strategies to improve mental health.

3.5.2 Sleep and Physical Activity in mHealth

In an mHealth study on population aging comparing schizophrenia (SZ) and health controls (HC), the investigators assessed objective sleep measures with a wrist-worn actigraph device (Actisleep-BT; Actigraph, Pensacola, FL).

For illustration purposes, we took the total sleep time (TST) as the outcome of interest. Our initial analysis revealed that the two groups have similar mean TST (HC: 366.12 vs. SZ: 367.81, $p = 0.96$), while the SZ group has significantly higher variability (HC: 78.52 vs. SZ: 136.59, $p = 0.0084$, see Figure 3.1). In view of this, we aim to to characterize such a heterogeneity across groups. The raw sequence for each person of each variable has $m = 19,751$ measurements. To reduce the dimension, we created respective pairwise explanatory variables for four variables: disease status (SZ vs. HC), gender, sleep efficiency, and total steps counts. We then fit the distance-based FRM to derive insights.

Our distance-based regression identified that the mean TST distance for SZ-SZ pairs was $e^{0.4062} = 1.5$ times of that for the reference HC-HC pairs, revealing the larger "dispersion" (variability) in the SZ group. The significant comparison of HC-SZ vs. HC-HC pairs ($p = 0.02$)

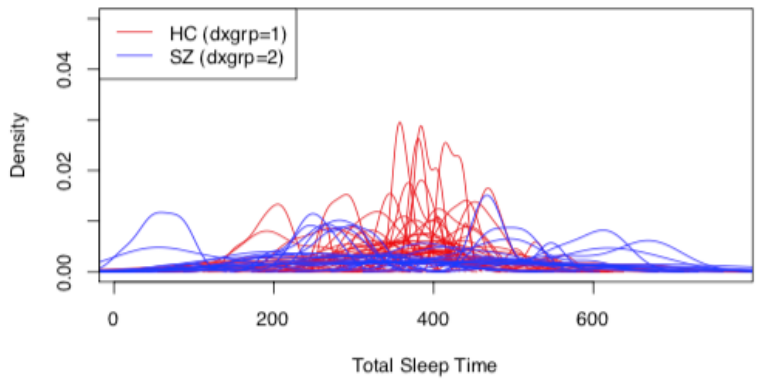

Figure 3.1. Comparison of total sleep time (TST) for raw data between 2 groups.

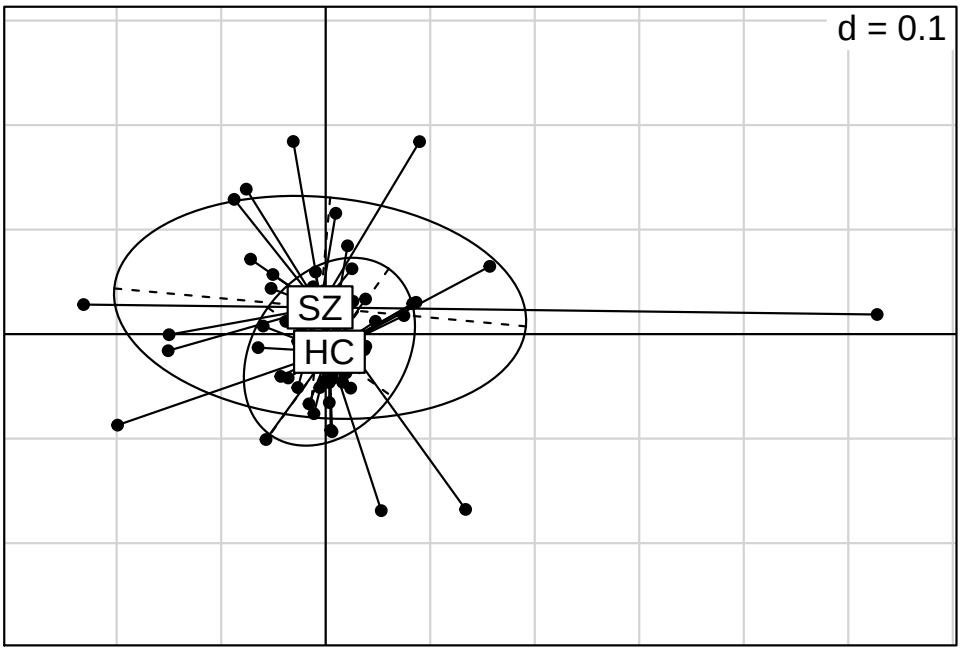

Figure 3.2. PCoA plot of total sleep time (TST) for between-subject distances of 2 groups.

Table 3.4. Results for the wearables data using FRM, MRM and PERMANOVA with composite outcome (Total Sleep Time).

FRM				
Parameter	Est.	Std. Error	Statistic Wald	p-value Wald
β_0	-2.5021	0.1989	158.2489	<0.0001
β_{w1}^{22}	0.4062	0.1758	5.3398	0.0208
β_{w1}^{12}	0.2822	0.1244	5.1466	0.0233
β_{w2}^{22}	0.0076	0.1591	0.0023	0.9621
β_{w2}^{12}	-0.0088	0.0938	0.0088	0.9254
β_{z1}	4.6944	1.6170	8.4285	0.0037
β_{z2}	0.1827	0.2793	0.4280	0.5130

MRM		
Parameter	Est.	p-value Permutation #: 999
ξ_0	271.9212	0.8639
ξ_{z1}	4.1268	0.0010
ξ_{z2}	0.0639	0.0040

PERMANOVA						
Parameter	Df.	Sums of Sqs.	Mean Sqs.	Statistic Pseudo-F	R^2	p-value Permutation #: 999
τ_{w1}	1	0.0374	0.0374	2.1985	0.0353	0.0750
τ_{w2}	1	0.0019	0.0019	0.1112	0.0018	0.9540
Residuals	60	1.0199	0.0170		0.9629	
Total	62	1.0592			1.0000	

also uncovers the "location" difference of the two groups. Since if HC and SZ are close in location, the mean distance between HC-SZ and HC-HC pairs should be small. In addition, FRM indicated significant associations between the variabilities in total sleep time and sleep efficiency ($p = 0.004$), but not so with total step counts ($p = 0.513$).

Comparing with existing approaches, MRM with only continuous distance of sleep efficiency ($p = 0.001$) and steps counts ($p = 0.004$) revealed that both are significantly associated with the distance of TST, but fails to include the most important disease indicator or control for gender. In comparison, PERMANOVA fails to include continuous distances as predictors but only uncovered the relationships between TST and factors of disease indicator ($p = 0.075$) and gender ($p = 0.954$). But the proposed FRM accommodates both types to derive more in-depth scientific findings. It successfully delineated the little-known relationships between the heterogeneity (beyond the mean) of sleep and activity with schizophrenia, which could further lead to personalized disease interventions.

3.6 Discussion

This chapter extended a unified semiparametric GLM-type regression paradigm for distances. In this era confronting high-dimensional data burgeoning from various disciplines, our proposed framework fills a critical gap in statistical modeling by leveraging the pairwise distances to achieve effective dimension reduction. By tackling the high dimensionality from an alternative angle, it complements regularization-based approaches (such as lasso). It may help reveal new scientific insights with the preserved dendrogram/tree structure of some specific distances.

Usually, after detecting a significant overall effect of the raw sequence on a clinical phenotype, further interest is to pinpoint the segment that drives such an effect. To do so, one can further create segmentation-based distances to locate such segment sequences.

In addition, as most clustering algorithms involve the distance of raw data, one can

repurpose the distance-based FRM to determine the optimal number of clusters. For instance, by specifying a certain number of clusters to a dataset such as k-mean, we can assign subjects to specific clusters. We then one-hot encode this clustering index as explanatory variables to fit the FRM. By evaluating test statistics such as the overall significance of the clustering index for each regression (with the specified cluster number), we can choose an optimal one.

Last but not least, by virtue of the semiparametric inference, UGEE estimators enjoy not only computational scalability but also robust asymptotic properties. Moreover, we will demonstrate in the next Chapter that akin to GEE for within-subject attributes (Tsiatis, 2007), the proposed estimators are highly efficient as well, reaching the semiparametric efficiency bound when the variance is specified correctly.

One major limitation is that it only focuses on cross-sectional data. Future extensions include distance-based regression for general clustered or longitudinal data with missingness. With this unified framework as a building block, those further developments will shed light on abundant research disciplines by disentangling the intricate interplays from data with astronomical dimensions. Chapter 3, in part is currently being prepared for submission for publication of the material. The dissertation author was the primary investigator and author of this material. The co-authors include Zhang, X., Zhong, Y., Lin, T., Chen, T., Wu, T., Nguyen, T.T, Jeste, D. V. and Tu, XM.

Chapter 4

On Semiparametric Efficiency of an Emerging Class of Distance-based Regression Models for Between-subject Attributes

4.1 Introduction

As a popular class of regression models, generalized linear models (GLM) encompass a wide variety of nonnormal and noncontinuous responses (or dependent variables) as a unified theory in modeling different types of responses (Agresti, 2003b). The maximum likelihood estimators (MLE) for GLM enjoy consistency and asymptotic normality (CAN) if both the random and systematic components, as well as the link function, are correctly specified. Estimators whose (asymptotic) variances achieve the Cramér-Rao bound are (asymptotically) efficient; MLEs are examples of efficient estimators for parametric models.

By relaxing the distributional assumption in the random component, the semiparametric GLM, also referred to as the restricted moment models (RMM) (Tsiatis, 2006), enable statistical inference for a broader class. Under suitable regularity conditions, estimators from the generalized estimating equations (GEE) are not only consistent and asymptotically normal (Liang and Zeger, 1986) but also achieve the semiparametric efficiency bound (Tsiatis, 2006).

Despite their broad applicability, the semiparametric GLM, or even the predominant regression paradigm, characterizes the relationship within the same subject from the raw data,

termed the "within-subject attributes" (Liu et al., 2021). But in the growing applications, of major interest are outcomes defined by a pair of subjects, or the "between-subject attributes" (Liu et al., 2021). The probability index $Pr(Y_{i_1} < Y_{i_2})$, $(i_1, i_2) \in C_2^n$ in the Mann-Whitney-Wilcoxon (MWW) rank-sum test is a classical example (Chen et al., 2016). Modern examples include pairwise dissimilarity/similarity summarizing data with astronomical dimensions fueled by gene-sequencing, wearable technology, etc (Nguyen et al., 2021b; Martinato et al., 2021).

Modeling between-subject attributes is challenging due to their complex correlation structures among pairwise observations. To address this, a class of semiparametric functional response models (FRM) have been proposed, as natural extensions of the semiparametric GLM. The FRM have been applied to model pairwise between-subject attributes in various settings, such as the microbiome Beta-diversity (Liu et al., 2021), reliability coefficients (Lu et al., 2014), causal inference for the MWW rank-sum test (Wu et al., 2014b; Lin et al., 2021), and rank-based robust regression for longitudinal data (Chen et al., 2014, 2016). Estimators of FRM regression based on a class of U-statistics-based generalized estimating equations (UGEE) also enjoy nice asymptotic properties just like their GEE counterparts of the semiparametric GLM.

Nevertheless, the efficiency of UGEE-based estimators for semiparametric FRM has not yet been thoroughly studied to the best of our knowledge. As in the study of estimators for within-subject attributes, the goal is to find the estimator(s) with the smallest asymptotic variance, termed the semiparametric efficient estimator(s). To this end, one first needs to extend key concepts such as influence functions and asymptotic linearity from the classical within-subject settings and then develop a coherent theory tailored for between-subject attributes in the context of FRM regression. In this paper, we leverage the Hilbert-space-based semiparametric efficiency theory to demonstrate that UGEE estimators achieve the semiparametric efficiency bound, just like GEE estimators for the semiparametric GLM. Harmonizing the efficiency and semiparametric robustness, the FRM provide an effective approach for modeling between-subject attributes, facilitate understandings of fundamental scientific questions that call for such

models, and inform new knowledge discovery and decision making.

4.2 Between-subject Regression

4.2.1 Semiparametric GLM and Functional Response Model

Here we consider a class of semiparametric functional response models (FRM):

$$E\left[f\left(Y_{i_1},\ldots,Y_{i_s}\right) \mid \mathbf{X}_{i_1},\ldots,\mathbf{X}_{i_s}\right] = h\left(\mathbf{X}_{i_1},\ldots,\mathbf{X}_{i_s};\beta\right),\ \left(i_1,\ldots,i_s\right) \in C_s^n,\ s \geq 1, \tag{4.1}$$

where $f(\cdot)$ is some scalar-valued function, $h(\cdot)$ is some smooth function (e.g., with continuous derivatives up to the second order), β is a vector of parameters, s is a positive integer. Akin to (3.2), (4.1) is also semiparametric with no distributional assumption imposed for the response. In practice, this is particularly appealing due to the increased difficulty to specify such an assumption for multi-subject based response function $f\left(Y_{i_1},\ldots,Y_{i_s}\right)$ to resemble real study data. As a special case when $s=1$ and $f\left(Y_i\right)=Y_i$, (4.1) reduces to the SPGLM in (3.2). Like SPGLM, the FRM is also readily extended to model a vector-valued response function (see **Supplement and Example 3 in** Section 4.2.2).

For many studies involving endogenous or exogenous pairwise outcomes, $s=2$, $f_{\mathbf{i}} = f\left(Y_{i_1},Y_{i_2}\right)$ and $\mathbf{X}_{\mathbf{i}} = \left(\mathbf{X}_{i_1}^\top,\mathbf{X}_{i_2}^\top\right)^\top$. The FRM below models the between-subject attributes $f_{\mathbf{i}}$ as a function of $\mathbf{X}_{\mathbf{i}}$:

$$E\left(f_{\mathbf{i}} \mid \mathbf{X}_{\mathbf{i}}\right) = h\left(\mathbf{X}_{\mathbf{i}};\beta\right),\ \mathbf{i} = \left(i_1,i_2\right) \in C_2^n, \tag{4.2}$$

where C_2^n denotes the set of combinations (i_1,i_2) from the integer set $\{1,\ldots,n\}$. When research interests involve exogenous or endogenous between-subject attributes, semiparametric FRM uniquely positions itself to facilitate data-driven knowledge discoveries, which could otherwise be hindered by the predominant paradigm of merely modeling within-subject attributes. We highlight the versatility of semiparametric FRM with some additional examples. More applications can be found in the references of Introduction Section.

4.2.2 Examples of Functional Response Models

Example 1: The Beta-diversity for High-throughput Data in Microbiome

In microbiome studies, our major interest lies in the association between Beta-diversity and certain clinical outcomes, such as disease status. Hence, if we could construct appropriate between-subject clinical variables such as a pairwise group indicator, (4.2) can be readily applied to reveal whether the mean Beta-diversity is higher, say, in the healthy than the diseased population.

Now the challenge is to convert the conventional within-subject group indicator to a between-subject attribute. Specifically, we consider K total disease groups with n_k denoting the sample size of the k-th group $(1 \leq k \leq K)$, and $n = \sum_{k=1}^{K} n_k$ is the total sample size. Let X_i denote the group membership for the i-th subject $(1 \leq X_i \leq K,\ 1 \leq i \leq n)$. For each pair $\mathbf{i}$ $(= (i_1, i_2) \in C_2^n)$, we observe a pairwise group indicator $\mathbf{X_i} = \{X_{i_1}, X_{i_2}\}$ $(1 \leq X_{i_1}, X_{i_2} \leq K)$. We can include all different combinations of $\mathbf{X_i}$ in a vector $\delta(\mathbf{X_i}) \in \{0,1\}^{K+C_2^K}$ through a one-hot encoding function $\delta : \{1,...,K\} \times \{1,...,K\} \mapsto \{0,1\}$ such that for $\mathbf{k} = \{k_1, k_2\}$ $(1 \leq k_1 \leq k_2 \leq K)$:

$$\delta_{\mathbf{k}}(\mathbf{X_i}) = \begin{cases} 1, & \text{if } \mathbf{X_i} = \{X_{i_1}, X_{i_2}\} = \{k_1, k_2\} = \mathbf{k}, \\ 0, & \text{otherwise.} \end{cases} \tag{4.3}$$

Let $f_{\mathbf{i}}$ denote the Beta-diversity for the $\mathbf{i}$-th pair such as the Aichison distance in (2.2), we can model its mean among different groups with the FRM:

$$E\left(f_{\mathbf{i}} \mid \mathbf{X_i}\right) = \exp\left[\beta^{\top} \delta(\mathbf{X_i})\right],\ \beta = \left(\tau_{11}, \ldots, \tau_{KK}\right)^{\top},\ \delta\left(\mathbf{X_i}\right) = \left(\delta_{11}\left(\mathbf{X_i}\right), \ldots, \delta_{KK}\left(\mathbf{X_i}\right)\right)^{\top}, \tag{4.4}$$

where $\exp(\cdot)$ is adopted to ensure that the response is non-negative. We can also include either between- or within-subject attributes as covariates in (4.4). For between-subject covariates, it is straightforward. For within-subject attributes, we can readily create their between-subject counterparts as shown above and in (Liu et al., 2021).

Example 2: Mann-Whitney-Wilcoxon Rank-sum Test and Rank Regression

Let Y_i $(1 \leq i \leq n)$ denote a univariate continuous within-subject response. In the presence of outliers, the Mann-Whitney-Wilcoxon rank-sum test provides an alternative to the two-sample t-test when comparing the centers of two distributions (**?**). To extend the rank-sum test into a regression, let $\mathbf{X}_i$ denote a vector of explanatory variables for subject i and consider the FRM:

$$E\left(f_{\mathbf{i}} \mid \mathbf{X}_{\mathbf{i}}\right) = h\left(\mathbf{X}_{\mathbf{i}}; \beta\right) = \Phi\left[-\beta^{\top}\left(\mathbf{X}_{i_1} - \mathbf{X}_{i_2}\right)\right], \; f_{\mathbf{i}} = f\left(Y_{i_1}, Y_{i_2}\right) = I\left(Y_{i_1} \leq Y_{i_2}\right), \tag{4.5}$$

where $\mathbf{i} = (i_1, i_2) \in C_2^n$, $\mathbf{X}_{\mathbf{i}} = \left(\mathbf{X}_{i_1}^{\top}, \mathbf{X}_{i_2}^{\top}\right)^{\top}$, and $\Phi(\cdot)$ denotes the cumulative distribution function (CDF) of the standard normal distribution. The model above has a parameter β that preserves its interpretation in a linear model by regressing the within-subject Y_i on $\mathbf{X}_i$ but addresses outliers in Y_i. Unlike Example 1, the response and explanatory variables here are both exogenous, and research interest lies in the relationship between the within-subject Y_i and $\mathbf{X}_i$. An extension of (4.5) to longitudinal settings with missing values is discussed by Chen et al. (2016). More examples for such probability index models can be found in Thas et al. (2012).

Example 3: Intraclass Correlations for Rater Agreement

Consider a study of n subjects in which each subject is rated by K judges. Let Y_{ik} denote the rating for the i-th subject by the k-th judge $(1 \leq i \leq n,\ 1 \leq k \leq K)$, which is commonly characterized by a two-way mixed-effects model:

$$\begin{aligned} &Y_{ik} = \mu + \beta_i + \gamma_k + (\beta\gamma)_{ik} + \varepsilon_{ik},\ \varepsilon_{ik} \sim N\left(0, \sigma_{\varepsilon}^2\right), \\ &\beta_i \sim N\left(0, \sigma_{\beta}^2\right),\ \sum_{k=1}^{K} \gamma_k = 0,\ (\beta\gamma)_{ik} \sim N\left(0, \sigma_{\beta\gamma}^2\right),\ \sum_{k=1}^{K} (\beta\gamma)_{ik} = 0, \end{aligned} \tag{4.6}$$

where $N\left(0, \sigma^2\right)$ denotes a normal distribution with mean 0 and variance σ^2. The intraclass correlation (ICC), $\rho = \left[\sigma_{\beta}^2 - \sigma_{\beta\gamma}^2/(K-1)\right] / \left(\sigma_{\beta}^2 + \sigma_{\beta\gamma}^2 + \sigma^2\right)$, is a widely applied index of agreement among K judges (Shrout and Fleiss, 1979; McGraw and Wong, 1996). However, the major issue of the above model is the difficulty to validate multiple normal assumptions in (4.6)

(especially for random effects due to their latent nature), rendering likelihood-based approaches prone to invalid inference under non-normal rating data.

Now consider a semiparametric alternative. For $\mathbf{i} = (i_1, i_2) \in C_2^n$, let

$$\overline{Y}_{i\cdot} = \frac{1}{K}\sum_{k=1}^{K} Y_{ik},\ g_{\mathbf{i}} = \frac{1}{2}\left(\overline{Y}_{i_1\cdot} - \overline{Y}_{i_2\cdot}\right)^2,\ g_{\mathbf{i}k} = \frac{1}{2}\left(Y_{i_1k} - Y_{i_2k}\right)^2,$$
$$f_{\mathbf{i}1} = g_{\mathbf{i}},\ f_{\mathbf{i}2} = \frac{1}{K}\sum_{k=1}^{K} g_{\mathbf{i}k},\ h_{\mathbf{i}1} = \frac{\left[1+(K-1)\,\rho\right]\tau^2}{K},\ h_{\mathbf{i}2} = \tau^2,$$

we construct the following (multivariate) FRM:

$$E\left(\mathbf{f}_{\mathbf{i}}\right) = \mathbf{h}_{\mathbf{i}}\left(\theta\right),\ \mathbf{f}_{\mathbf{i}} = \left(f_{\mathbf{i}1}, f_{\mathbf{i}2}\right)^{\top},\ \mathbf{h}_{\mathbf{i}} = \left(h_{\mathbf{i}1}, h_{\mathbf{i}2}\right)^{\top},\ \theta = \left(\tau^2, \rho\right)^{\top}. \tag{4.7}$$

The ρ in (4.7) is exactly the ICC (Lu et al., 2014). In addition to its robustness, this model also allows for an immediate extension to longitudinal settings.

4.2.3 Inference for the U-statistics and UGEE

The FRM reinvigorates regression by extending within- to between-subject attributes. However, popular asymptotic methods, such as the law of large numbers and central limit theorem (CLT), rely on the pivotal assumption of independence and as such are not directly applicable to FRM due to complex correlation structures of functional responses. We address such challenges by leveraging theory of U-statistics (Hoeffding and Robbins, 1948).

Asymptotic Properties of U-statistics

Most statistics from classical models of within-subject attributes are in the form of a summation of $i.i.d.$ elements, such as the score and estimating equations. However, statistics formed by between-subject attributes from FRM are correlated hence prohibiting direct applications of the CLT for inference. To resolve this issue, a class of U-statistics-based generalized estimating equations (UGEE) have been developed to facilitate inference. We first give a brief review of

the U-statistics, which play an instrumental role in studying multi-subject-based statistics. More details and examples can be found in Kowalski and Tu (2008b) and the references in Section **??**.

Definition Consider a sample of *i.i.d.* random vectors $\mathbf{Y}_i \in \mathbb{R}^m$ $(1 \leq i \leq n)$. Let $\mathbf{f}^{d\times 1}(\mathbf{Y}_1,\ldots,\mathbf{Y}_s)$ be a d-dimensional symmetric function with s arguments, or input vectors, i.e., $\mathbf{f}(\mathbf{Y}_1,\ldots,\mathbf{Y}_s) = \mathbf{f}(\mathbf{Y}_{i_1},\ldots,\mathbf{Y}_{i_s})$ for any permutation $(i_1,\ldots,i_s)$ of $(1,\ldots,s)$. A d-variate, one-sample, s-argument *U-statistic* is

$$\mathbf{U}_n = \binom{n}{s}^{-1} \sum_{(i_1,\ldots,i_s)\in C_s^n} \mathbf{f}(\mathbf{Y}_{i_1},\ldots,\mathbf{Y}_{i_s}),\ s \geq 1, \tag{4.8}$$

where $C_s^n = \{(i_1,\ldots,i_s); 1 \leq i_1 < \ldots < i_s \leq n\}$ denotes the set of all distinct s-combinations from the integer set $\{1,\ldots,n\}$. Let $\theta = E[\mathbf{f}(\mathbf{Y}_{i_1},\ldots,\mathbf{Y}_{i_s})]$, then it can be checked that $E(\mathbf{U}_n) = \theta$, i.e., $\mathbf{U}_n$ is an unbiased estimator of θ.

Since $\mathbf{f}(\mathbf{Y}_{i_1},\ldots,\mathbf{Y}_{i_s})$ (also termed the *kernel function*) involves multiple rather than a single subject, dependencies between any two kernel functions arise when they share at least one common subject (e.g., $f(\mathbf{Y}_{i_1},\mathbf{Y}_{i_2})$ and $f(\mathbf{Y}_{i_1},\mathbf{Y}_{i_3})$ are correlated as they share $\mathbf{Y}_{i_1}$). This dependency is tackled through the *Hájek projection* $\widetilde{\mathbf{U}}_n$ (Hájek, 1968):

$$\widetilde{\mathbf{U}}_n = \frac{s}{n}\sum_{i_1=1}^{n} E[\mathbf{f}(\mathbf{Y}_1,\ldots,\mathbf{Y}_s) \mid \mathbf{Y}_i]. \tag{4.9}$$

The conditional expectations of $\mathbf{f}(\mathbf{Y}_1,\ldots,\mathbf{Y}_s)$ given each $\mathbf{Y}_i$of the *i.i.d.* sample are *i.i.d.*, permitting applications of conventional asymptotic techniques (Kowalski and Tu, 2008b). As shown below, the U-statistic and its projection have the same asymptotic distribution.

Theorem 4.1. Let

$$\widetilde{\mathbf{v}}_1(\mathbf{Y}_1) = E[\mathbf{f}(\mathbf{Y}_1,\ldots,\mathbf{Y}_s) \mid \mathbf{Y}_1] - \theta,\ \mathbf{e}_n = \sqrt{n}\left(\mathbf{U}_n - \widetilde{\mathbf{U}}_n\right),\ \Sigma_v = Var[\widetilde{\mathbf{v}}_1(\mathbf{Y}_1)]. \tag{4.10}$$

Under mild regularity conditions (see **Supplement** for details), $\mathbf{e}_n \to_p \mathbf{0}$ and thus,

(i) $\mathbf{U}_n$ is consistent, i.e., $\mathbf{U}_n \to_p \theta$.

(ii) $\mathbf{U}_n$ is asymptotically (multivariate) normal:

$$\sqrt{n}\,(\mathbf{U}_n - \theta) \to_d N\left(\mathbf{0}, \Sigma_U = s^2\Sigma_v\right), \tag{4.11}$$

where $\to_{p\,(d)}$ denotes convergence in probability (distribution).

U-statistics-based Generalized Estimating Equations

As between-subject outcomes in most applications, including all in this paper, involve pairs of subjects, we will focus on pairwise outcomes for notational brevity. Extensions to between-subject outcomes involving more than two subjects are straightforward.

For pairwise outcomes, inference for β must address their interlocking dependent relationships. We tackle this based on a class of U-statistics-based Generalized Estimating Equations (UGEE) (Kowalski and Tu, 2008b) defined by

$$\mathbf{U}_n(\beta) = \sum_{\mathbf{i}\in C_2^n} \mathbf{U}_{n,\mathbf{i}}(\beta) = \sum_{\mathbf{i}\in C_2^n} \mathbf{D}_\mathbf{i}^\top V_\mathbf{i}^{-1} S_\mathbf{i}(\beta) = \mathbf{0},\ \mathbf{i} = (i_1, i_2) \in C_2^n, \tag{4.12}$$

$$S_\mathbf{i}(\beta) = f_\mathbf{i} - h_\mathbf{i}(\mathbf{X_i};\beta),\ \mathbf{D_i} = \frac{\partial}{\partial \beta^\top} h_\mathbf{i}(\mathbf{X_i};\beta),\ V_\mathbf{i} = Var(f_\mathbf{i} \mid \mathbf{X_i}).$$

In practice, $V_\mathbf{i}$ is unknown and substituted by a working variance. Estimators $\widehat{\beta}_{\text{ugee}}$ of β are readily obtained by solving (4.12) numerically such as with the Newton-Raphson method. Although similar in appearance to GEE (Tang et al., 2012b), UGEE is not a sum of independent variables. But $\widehat{\beta}_{\text{ugee}}$ is asymptotically normal as shown below, which is readily proved using Theorem 1 (see the **Appendix**).

Theorem 4.2. Let

$$\widetilde{\mathbf{v}}_{i_1} = 2E\left(\mathbf{U}_{n,\mathbf{i}} \mid \mathbf{Y}_{i_1}, \mathbf{X}_{i_1}\right),\ \mathbf{B} = E\left(\mathbf{D}_\mathbf{i}^\top V_\mathbf{i}^{-1} \mathbf{D}_\mathbf{i}\right),\ \Sigma_U = Var(\widetilde{\mathbf{v}}_{i_1}),\ \Sigma_\beta^{\text{ugee}} = \mathbf{B}^{-1}\Sigma_U\mathbf{B}^{-1}. \tag{4.13}$$

Under mild regularity conditions (see **Supplements** for details), $\widehat{\beta}_{\text{ugee}}$ is a consistent and asymptotically normal (CAN) estimator of β in (4.2):

$$\sqrt{n}\left(\widehat{\beta}_{\text{ugee}}-\beta\right)\rightarrow_{d} N\left(\mathbf{0},\Sigma_{\beta}^{\text{ugee}}\right).$$

A consistent estimator of $\Sigma_{\beta}^{\text{ugee}}$ can be obtained by substituting consistent estimators of β and moment estimators of the respective quantities in (4.13). Conforming to the appealing features of its within-subject counterpart GEE, UGEE also yields valid inference without explicitly delineating the potentially more complex correlation structures.

4.3 Asymptotic Linearity and Influence Function

To study semiparametric efficiency for between-subject attributes, we first need to extend concepts such as asymptotic linearity and influence functions (Bickel et al., 1998; Hampel, 1974), of those classical within-subject attributes $\mathbf{Z}_1,...,\mathbf{Z}_n \sim^{i.i.d.} \{p(\mathbf{Z}_i;\theta);\ \theta\in\Omega\}$, where $p(\mathbf{Z}_i;\theta)$ is a probability density or distribution function characterized by parameter θ. In what follows, we assume $\theta=\left(\beta^{\top},\eta\right)^{\top}$ (i.e., β and η are variationally independent with no overlapping components), where β is a $q\times 1$ vector of *parameters of interest* and η is the *nuisance parameter*. Here the only component that differentiates parametric from semiparametric models is the dimension of η; a finite-dimensional vector η yields the parametric while an infinite-dimensional nuisance parameter, denoted by $\eta(\cdot)$, encompasses the semiparametric representation (Tsiatis, 2006).

For classical within-subject attributes, an estimator of $\widehat{\beta}$ is *asymptotically linear (AL)* if there exists an expansion $n^{1/2}\left(\widehat{\beta}-\beta_0\right)=n^{-1/2}\sum_{i=1}^{n}\phi(\mathbf{Z}_i;\theta_0)+\mathbf{o}_p(1)$, where $\phi(\mathbf{Z}_i;\theta_0)$ is termed the *influence function (I.F.)* for the i-th observation with θ_0 denoting the truth. The I.F. has mean zero and finite and nonsingular $E\left(\phi\phi^{\top}\right)$, its name reflects the *influence* of an observation unit on the estimator (Ichimura and Newey, 2015). The asymptotic normality can

be readily derived from this expansion by CLT, $n^{1/2}\left(\widehat{\beta}-\beta_0\right)\to_d N\left(\mathbf{0},E\left(\phi\phi^\top\right)\right)$, i.e., the asymptotic variance of $\widehat{\beta}$ is determined by its I.F., hence, $\phi\left(\mathbf{Z}_i\right)$ defines the efficiency of $\mathbf{Z}_i$ (Tsiatis, 2006).

Those $\mathbf{Z}_i$'s induce a sequence of $i.d.$ (identically distributed but not necessarily independent) random vectors, $\mathbf{Z}_\mathbf{i}=\left(\mathbf{Z}_{i_1}^\top,\mathbf{Z}_{i_2}^\top\right)^\top\sim\left\{p\left(\mathbf{Z}_\mathbf{i};\theta\right);\ \theta\in\Omega\right\}$ for $\mathbf{i}=(i_1,i_2)\in C_2^n$. For $\mathbf{Z}_\mathbf{i}$, we consider two classes of models, either parametric or semiparametric, and associated estimators and I.F.s.

4.3.1 Non-overlap Model Class 1

We first discuss constructing likelihood or estimating equations based on $i.i.d.$ pairs $\mathbf{Z}_{\mathbf{i}_j}$, $1\le j\le\lfloor n/2\rfloor=m$, where $\lfloor\cdot\rfloor$ denotes the floor function. Namely, we reorganize the data into independent non-overlapping pairs. This reorganization is not unique, and without loss of generality, we choose one of them. For example, when $n=4$, we can choose $\{\mathbf{Z}_{i_1},\mathbf{Z}_{i_2}\}$ and $\{\mathbf{Z}_{i_3},\mathbf{Z}_{i_4}\}$ to form two independent pairs $\mathbf{i}_1=(i_1,i_2)$ and $\mathbf{i}_2=(i_3,i_4)$. It removes the hurdle of dependencies originated from overlapping pairs, providing results in parallel with the classical within-subject case.

Definition. $\widetilde{\beta}$ is an *asymptotically linear (AL)* estimator of between-subject attributes for the *non-overlap model class 1* if it belongs to

$$\Omega_1^\beta=\left\{\widetilde{\beta}(\mathbf{Z}_{\mathbf{i}_j}):\sqrt{m}\left(\widetilde{\beta}-\beta_0\right)=\sqrt{m}\frac{1}{m}\sum_{j=1}^{m}\psi\left(\mathbf{Z}_{\mathbf{i}_j};\theta_0\right)+\mathbf{o}_p(1)\right\},\tag{4.14}$$

where $\psi^{q\times1}\left(\mathbf{Z}_\mathbf{i};\theta_0\right)$ is a measurable function with mean zero, finite and nonsingular $E\left(\psi\psi^\top\right)$, defined as the *non-overlap influence function 1* for the $\mathbf{i}$-th pair at the truth. The set of all such I.F.s is denoted by $\Gamma_1^{I.F.}$.

Under mild regularity conditions, CLT yields $\sqrt{m}\left(\widetilde{\beta}-\beta_0\right)\to_d N\left(\mathbf{0},\Sigma_1\right)$, $\Sigma_1=E\left[\psi\left(\mathbf{Z}_{\mathbf{i}_j}\right)\psi^\top\left(\mathbf{Z}_{\mathbf{i}_j}\right)\right]$. Hence, the asymptotic variance for $\widetilde{\beta}\in\Omega_1^\beta$ is determined by its I.F., and the *efficient* estimator in Ω_1^β should have minimum variance.

In practice, we do not fit data with between-subject attributes using model class 1, since they only deploy part of the data. But this class of conceptual models will help determine the efficient estimator for the FRM, termed model class 2, which we now introduce.

4.3.2 Enumerated Model Class 2

If making the inference based on all possible pairs of $\mathbf{Z_i}$, $\mathbf{i} = (i_1, i_2) \in C_2^n$, including those with overlapping subjects, we reimpose the dependencies and form a class of enumerated models. The FRM in (4.2) that engage all possible pairs is an example of this class.

Definition. $\widehat{\beta}$ is an *AL* estimator of between-subject attributes for the *enumerated model 2* if it belongs to

$$\Omega_2^{\beta} = \left\{ \widehat{\beta}(\mathbf{Z_i}) : n^{1/2}\left(\widehat{\beta} - \beta_0\right) = \sqrt{n}\binom{n}{2}^{-1} \sum_{\mathbf{i} \in C_2^n} \varphi\left(\mathbf{Z_i}; \theta_0\right) + \mathbf{o}_p(1) \right\}, \tag{4.15}$$

where the measurable function $\varphi^{q\times 1}(\mathbf{Z_i}; \theta_0)$ with mean zero and finite and nonsingular $E\left(\varphi\varphi^{\top}\right)$ is defined as the *enumerated influence function 2* for the **i**-th pair at the truth. Denote the set of all such $\varphi(\mathbf{Z_i}; \theta_0)$ by $\Gamma_2^{I.F.}$.

As (4.15) involves the summation of dependent $\varphi(\mathbf{Z_i})$, we apply (4.10) to obtain:

$$\begin{aligned} \sqrt{n}\left(\widehat{\beta} - \beta_0\right) &= \sqrt{n}\frac{1}{n}\sum_{i_1=1}^{n} 2E\left[\varphi\left(\mathbf{Z_i}; \theta_0\right) \mid \mathbf{Z}_{i_1}\right] + \mathbf{o}_p(1) \to_d N(\mathbf{0}, \Sigma_2), \\ \Sigma_2 &= Var\left\{2E\left[\varphi\left(\mathbf{Z_i}; \theta_0\right) \mid \mathbf{Z}_{i_1}\right]\right\} = E\left\{2E\left[\varphi\left(\mathbf{Z_i}; \theta_0\right) \mid \mathbf{Z}_{i_1}\right] \cdot 2E\left[\varphi^{\top}\left(\mathbf{Z_i}; \theta_0\right) \mid \mathbf{Z}_{i_1}\right]\right\}. \end{aligned} \tag{4.16}$$

Hence, for enumerated model class 2, the asymptotic variance Σ_2 of $\widehat{\beta} \in \Omega_2^{\beta}$ is also determined by its I.F., which apparently differs from Σ_1 for the model class 1, since Σ_2 involves an additional step of mapping from a function of between-subject attribute $\mathbf{Z_i}$ to a function of within-subject attribute $\mathbf{Z}_{i_1}$.

4.3.3 Relationships between I.F.s for the Two Model Classes

As mentioned, the influence function is key to studying efficiency. We now discuss the relationships between influence functions for the two classes of models.

Equivalence of Two Classes of AL estimators

For AL estimators of between-subject attributes, the I.F.s for model class 1 are equivalent to I.F.s for class 2. Namely, for any $\psi(\mathbf{Z_i};\theta_0) \in \Gamma_1^{I.F.}$ and associated AL estimator $\widetilde{\beta} \in \Omega_1^{\beta}$, we can construct another estimator $\widehat{\beta} = \beta_0 + \binom{n}{2}^{-1} \sum_{\mathbf{i} \in C_2^n} \psi(\mathbf{Z_i};\theta_0)$. It is readily shown that this estimator belongs to the class of estimators defined by (4.15), indicating that $\widehat{\beta}$ is AL for the enumerated class 2.

Conversely, for any $\varphi(\mathbf{Z_i};\theta_0) \in \Gamma_2^{I.F.}$ and corresponding AL estimator $\widehat{\beta} \in \Omega_2^{\beta}$ satisfying (4.15), we define an estimator $\widetilde{\beta} = \beta_0 + m^{-1} \sum_{j=1}^{m} \varphi\left(\mathbf{Z}_{\mathbf{i}_j};\theta_0\right)$. It is again readily shown that this estimator satisfies (4.14), indicating that $\widetilde{\beta} \in \Omega_1^{\beta}$ and $\varphi(\mathbf{Z_i};\theta_0) \in \Gamma_1^{I.F.}$, i.e., $\varphi(\mathbf{Z_i};\theta_0)$ is also an I.F. for model class 1.

Equivalence of Two Classes of Regular and AL estimators

As in the literature, to avoid estimators with undesirable local properties such as *super-efficiency* (LeCam, 1953), we restrict considerations to *regular* estimators by considering a *local data generating process* (LDGP). Suppose the underlying within-subject attributes are generated from $\mathbf{Z}_{in} \sim^{i.i.d.} \{p(\mathbf{Z}_{in};\theta_n)\}$ for each θ_n, and $n^{1/2}(\theta_n - \theta_*)$ converges to a constant where θ_* denote some fixed parameter. Let $\widehat{\theta}(\mathbf{Z}_{\mathbf{i}n})$ denote an estimator of θ_n based on the between-subject attributes, where $\mathbf{Z}_{\mathbf{i}n} = (\mathbf{Z}_{i_1 n}^{\top}, \mathbf{Z}_{i_2 n}^{\top})^{\top}$, $\mathbf{i} = (i_1, i_2) \in C_2^n$. Then $\widehat{\theta}(\mathbf{Z}_{\mathbf{i}n})$ is *regular* if, for some fixed θ_*, the limiting distribution of $n^{1/2}\left(\widehat{\theta}(\mathbf{Z}_{\mathbf{i}n}) - \theta_n\right)$ does not depend on the LDGP (or θ_n). More details with an example violating the LDGP (i.e., a super-efficient $\widehat{\theta}(\mathbf{Z}_{\mathbf{i}n})$) are discussed in the **Supplements**. In what follows, we focus on *regular* and *asymptotically linear* (RAL) estimators unless stated otherwise. The theorem below further declares the equivalence between the two classes of I.F.s for RAL estimators. A proof is deferred to the **Appendix.**

Theorem 4.3 For RAL estimators of between-subject attributes, the I.F.s in the set $\Gamma_1^{I.F.}$

for model class 1 are equivalent to I.F.s in $\Gamma_2^{I.F.}$ for model class 2, i.e., $\Gamma_1^{I.F.} = \Gamma_2^{I.F.}$.

Although our goal is to find the efficient I.F. for the FRM in model class 2, it is much more difficult to directly work with this class of models due to the added complexity in the asymptotic variance of their estimators. Accordingly, Theorem 3 is critical by allowing us to achieve our goal by virtue of the simplicity of model class 1.

4.4 Hilbert Space and Projection

In this section, we start with a brief review of Hilbert space (Walter, 1987) and its application to the classical within-subject attributes and then extend such considerations to their between-subject counterparts. More details can be found in the **Supplement**.

4.4.1 Within-subject Attributes

Let $(\mathscr{L}, \mathscr{A}, P)$ be a probability space (where $\mathscr{L}$ is the sample space, $\mathscr{A}$ is the σ-algebra, and P is the probability measure). Consider a q-dimensional measurable function $\mathbf{Z} : \mathscr{L} \to \mathbb{R}^q$. Suppose we observe $i.i.d.$ within-subject attributes $\mathbf{Z}_1, ..., \mathbf{Z}_n$, where $\mathbf{Z}_i$ is the random vector for subject i. We denote by $\mathscr{H}_w$ the Hilbert space consisting of all q-dimensional functions of $\mathbf{Z}_i$, $\mathbf{h} : \mathscr{L} \to \mathbb{R}^q$, that are measurable with mean zero and finite second-order moments, associated with an inner product and a norm induced by this inner product (we emphasize quantities of within-subject attributes with a subscript w) :

$$\langle \mathbf{h}_1(\mathbf{Z}_i), \mathbf{h}_2(\mathbf{Z}_i) \rangle_w = E\left[\mathbf{h}_1^\top(\mathbf{Z}_i)\mathbf{h}_2(\mathbf{Z}_i)\right], \ \|\mathbf{h}(\mathbf{Z}_i)\|_w = \langle \mathbf{h}, \mathbf{h} \rangle_w^{1/2} = E^{1/2}\left[\mathbf{h}^\top(\mathbf{Z}_i)\mathbf{h}(\mathbf{Z}_i)\right]. \tag{4.17}$$

Let $\mathbf{v}(\mathbf{Z}_i) = (v_1(\mathbf{Z}_i), ..., v_r(\mathbf{Z}_i))^\top$ be an r-dimensional random function with $E[\mathbf{v}(\mathbf{Z}_i)] = \mathbf{0}$ and $\langle \mathbf{v}, \mathbf{v} \rangle_w < \infty$. For the linear subspace spanned by $\mathbf{v}(\mathbf{Z}_i)$:

$$\mathscr{U}_w = \{\mathbf{B}\mathbf{v}(\mathbf{Z}_i); \text{ for an arbitrary matrix } \mathbf{B}^{q\times r} \text{ of real numbers}\},$$

the projection of $\mathbf{h}^{q\times 1}(\mathbf{Z}_i) \in \mathscr{H}_w$ onto $\mathscr{U}_w$ is unique by the closest point theorem (Sehgal et al., 1987):

$$\Pi_w\{\mathbf{h}(\mathbf{Z}_i) \mid \mathscr{U}_w\} = E\left[\mathbf{h}(\mathbf{Z}_i)\mathbf{v}(\mathbf{Z}_i)^\top\right] E^{-1}\left[\mathbf{v}(\mathbf{Z}_i)\mathbf{v}(\mathbf{Z}_i)^\top\right]\mathbf{v}(\mathbf{Z}_i). \tag{4.18}$$

4.4.2 Between-subject Attributes

For the induced pairwise observations $\mathbf{Z}_\mathbf{i} = (\mathbf{Z}_{i_1}^\top, \mathbf{Z}_{i_2}^\top)^\top$, $\mathbf{i} = (i_1, i_2) \in C_2^n$, we consider the Hilbert space $\mathscr{H}_b$ (with a subscript b reflecting between-subject attributes) of all q-dimensional symmetric measurable functions $\mathbf{h}(\mathbf{Z}_\mathbf{i}) = \mathbf{h}(\mathbf{Z}_{i_1}, \mathbf{Z}_{i_2})$ with mean zero and finite $E\left[\mathbf{h}(\mathbf{Z}_\mathbf{i})\mathbf{h}^\top(\mathbf{Z}_\mathbf{i})\right]$. We consider two inner products and associated norms for $\mathscr{H}_b$.

Definition. The *non-overlap inner product 1* and associated *norm b1* are defined as

$$\begin{aligned}\langle \mathbf{h}_1(\mathbf{Z}_\mathbf{i}), \mathbf{h}_2(\mathbf{Z}_\mathbf{i})\rangle_{b1} &= E\left[\mathbf{h}_1^\top(\mathbf{Z}_\mathbf{i})\mathbf{h}_2(\mathbf{Z}_\mathbf{i})\right], \\ \|\mathbf{h}(\mathbf{Z}_\mathbf{i})\|_{b1} &= \langle \mathbf{h}(\mathbf{Z}_\mathbf{i}), \mathbf{h}(\mathbf{Z}_\mathbf{i})\rangle_{b1}^{1/2} = E^{1/2}\left[\mathbf{h}^\top(\mathbf{Z}_\mathbf{i})\mathbf{h}(\mathbf{Z}_\mathbf{i})\right].\end{aligned} \tag{4.19}$$

For the linear span of $\mathbf{v}(\mathbf{Z}_\mathbf{i}) = (v_1(\mathbf{Z}_\mathbf{i}), ..., v_r(\mathbf{Z}_\mathbf{i}))^\top$ (as a function of $\mathbf{Z}_\mathbf{i}$ for the $\mathbf{i}$-th pair):

$$\mathscr{U}_{b1} = \{\mathbf{B}\mathbf{v}(\mathbf{Z}_\mathbf{i});\ \text{for an arbitrary matrix } \mathbf{B}^{q\times r} \text{ of real numbers}\},$$

the projection of $\mathbf{h}^{q\times 1}(\mathbf{Z}_\mathbf{i}) \in \mathscr{H}_b$ onto $\mathscr{U}_{b1}$ is:

$$\Pi_{b1}\{\mathbf{h}(\mathbf{Z}_\mathbf{i}) \mid \mathscr{U}_{b1}\} = E\left[\mathbf{h}(\mathbf{Z}_\mathbf{i})\mathbf{v}(\mathbf{Z}_\mathbf{i})^\top\right] E^{-1}\left[\mathbf{v}(\mathbf{Z}_\mathbf{i})\mathbf{v}(\mathbf{Z}_\mathbf{i})^\top\right]\mathbf{v}(\mathbf{Z}_\mathbf{i}). \tag{4.20}$$

It follows from the Pythagorean triangle inequality that

$$\|\mathbf{h}(\mathbf{Z}_\mathbf{i})\|_{b1}^2 = \|\Pi_{b1}\{\mathbf{h} \mid \mathscr{U}_{b1}\}\|_{b1}^2 + \|\mathbf{h} - \Pi_{b1}\{\mathbf{h} \mid \mathscr{U}_{b1}\}\|_{b1}^2 \geq \|\Pi_{b1}\{\mathbf{h}(\mathbf{Z}_\mathbf{i}) \mid \mathscr{U}_{b1}\}\|_{b1}^2, \tag{4.21}$$

i.e., the norm b1 of any element $\mathbf{h}(\mathbf{Z}_\mathbf{i})$ is always larger than or equal to that of its projection onto the subspace $\mathscr{U}_{b1}$.

Under the q-replicating linear spaces that we consider (See **Supplement** for details), the multivariate Pythagoras holds: the orthogonality between $\mathbf{h}(\mathbf{Z_i})$ and $\mathscr{U}_{b1}$ is equivalent to the uncorrelatedness between $\mathbf{h}(\mathbf{Z_i})$ and $\mathbf{v}(\mathbf{Z_i})$ (i.e., $E\left[\mathbf{h}^{\top}(\mathbf{Z_i})\,\mathbf{v}(\mathbf{Z_i})\right]=0$ implies $E\left[\mathbf{h}(\mathbf{Z_i})\,\mathbf{v}^{\top}(\mathbf{Z_i})\right]=\mathbf{0}$). Thus (6.17) shows that the variance (matrix) of the element $\mathbf{h}(\mathbf{Z_i})\in\mathscr{H}_b$ also satisfies

$$Var\left[\mathbf{h}(\mathbf{Z_i})\right]=Var\left[\Pi_{b1}\left\{\mathbf{h}\mid\mathscr{U}_{b1}\right\}\right]+Var\left[\mathbf{h}-\Pi_{b1}\left\{\mathbf{h}\mid\mathscr{U}_{b1}\right\}\right]\geq Var\left[\Pi_{b1}\left\{\mathbf{h}(\mathbf{Z_i})\mid\mathscr{U}_{b1}\right\}\right]. \tag{4.22}$$

Hence, the variance of any element $\mathbf{h}(\mathbf{Z_i})$ is larger than or equal to its projection $\Pi_{b1}\left\{\mathbf{h}\mid\mathscr{U}_{b1}\right\}$ onto a subspace, i.e., their difference is non-negative definite. This will inspire the construction of the efficient estimator using the projection later.

As the norm b1 for $\mathscr{H}_b$ does not yield the asymptotic variance for the UGEE estimator, we now introduce another inner product motivated by the form of asymptotic variance based on the I.F.s for the enumerated model class 2.

Definition. The *enumerated inner product 2* and *norm b2* are defined as

$$\begin{aligned}\left\langle\mathbf{h}_1(\mathbf{Z_i}),\mathbf{h}_2(\mathbf{Z_i})\right\rangle_{b2}&=E\left\{2E\left[\mathbf{h}_1^{\top}(\mathbf{Z_i})\mid\mathbf{Z}_{i_1}\right]\cdot 2E\left[\mathbf{h}_2(\mathbf{Z_i})\mid\mathbf{Z}_{i_1}\right]\right\},\\ \left\|\mathbf{h}(\mathbf{Z_i})\right\|_{b2}&=\left\langle\mathbf{h}(\mathbf{Z_i}),\mathbf{h}(\mathbf{Z_i})\right\rangle_{b2}^{1/2}=E^{1/2}\left\{2E\left[\mathbf{h}^{\top}(\mathbf{Z_i})\mid\mathbf{Z}_{i_1}\right]\cdot 2E\left[\mathbf{h}(\mathbf{Z_i})\mid\mathbf{Z}_{i_1}\right]\right\}.\end{aligned} \tag{4.23}$$

Definition. Define a projection *mapping* (Luenberger, 1997), $\mathscr{M}\colon\mathscr{H}_b\to\mathscr{H}_w$, referred to as the *U-statistics, or Hajek, projection*, such that for $\mathbf{h}(\mathbf{Z_i})\in\mathscr{H}_b$,

$$\mathscr{M}\left[\mathbf{h}(\mathbf{Z_i})\right]=2E\left[\mathbf{h}(\mathbf{Z_i})\mid\mathbf{Z}_{i_1}\right]\in\mathscr{H}_w. \tag{4.24}$$

Now consider a linear subspace of $\mathscr{H}_w$ spanned by $\mathscr{M}\left[\mathbf{B}^{q\times r}\,\mathbf{v}(\mathbf{Z_i})\right]=\mathbf{B}E\left[\mathbf{v}(\mathbf{Z_i})\mid\mathbf{Z}_{i_1}\right]$,

$$\mathscr{U}_{b2}=\mathscr{M}\left(\mathscr{U}_{b1}\right)=\left\{\mathbf{B}E\left[\mathbf{v}(\mathbf{Z_i})\mid\mathbf{Z}_{i_1}\right];\ \text{for an arbitrary matrix }\mathbf{B}^{q\times r}\ \text{of real numbers}\right\}.$$

Projecting any $\mathbf{h}(\mathbf{Z_i}) \in \mathscr{H}_b$ onto $\mathscr{U}_{b2}$ involves two steps: we first map $\mathbf{h}(\mathbf{Z_i})$ to $\mathscr{M}[\mathbf{h}(\mathbf{Z_i})] \in \mathscr{H}_w$ and then project it onto $\mathscr{U}_{b2}$ with the projection theorem for within-subject attributes in (4.18), i.e.:

$$\Pi_{b2}\{\mathbf{h}(\mathbf{Z_i}) \mid \mathscr{U}_{b2}\} = \Pi_w\{\mathscr{M}[\mathbf{h}(\mathbf{Z_i})] \mid \mathscr{U}_{b2}\}. \tag{4.25}$$

Similarly, the norm b2 of any element $\mathbf{h}(\mathbf{Z_i})$ is always larger than or equal to that of its projection onto $\mathscr{U}_{b2}$, which by (4.25), equals the squared norm of the mapped element $\mathscr{M}[\mathbf{h}(\mathbf{Z_i})]$, i.e.,

$$\|\mathbf{h}(\mathbf{Z_i})\|_{b2}^2 \geq \|\Pi_{b2}\{\mathbf{h}(\mathbf{Z_i}) \mid \mathscr{U}_{b2}\}\|_{b2}^2 = \|\Pi_w\{\mathscr{M}[\mathbf{h}(\mathbf{Z_i})] \mid \mathscr{U}_{b2}\}\|_w^2. \tag{4.26}$$

Accordingly, the variance of any $\mathscr{M}[\mathbf{h}(\mathbf{Z_i})]$ is larger than or equal to its projection $\Pi_w\{\mathscr{M}[\mathbf{h}(\mathbf{Z_i})] \mid \mathscr{U}_{b2}\}$ by the multivariate Pythagoras:

$$Var\{\mathscr{M}[\mathbf{h}(\mathbf{Z_i})]\} \geq Var[\Pi_w\{\mathscr{M}[\mathbf{h}(\mathbf{Z_i})] \mid \mathscr{U}_{b2}\}] = Var[\Pi_{b2}\{\mathbf{h}(\mathbf{Z_i}) \mid \mathscr{U}_{b2}\}]. \tag{4.27}$$

The above links norm b2 with the asymptotic variance of the UGEE estimator. We now define equivalence classes within each norm and discuss the relationship between the two.

Definition. For a given norm, any two functions $\mathbf{h}_1(\mathbf{Z_i})$ and $\mathbf{h}_2(\mathbf{Z_i})$ are considered *equivalent*, if the norm of their difference is zero. The *equivalence class* of $\mathbf{h}(\mathbf{Z_i})$ under norm b1 includes all q-dimensional measurable functions $\mathbf{g}(\mathbf{Z_i}) \in \mathscr{H}_b$ that equal $\mathbf{h}(\mathbf{Z_i})$ almost surely (a.s.), denoted by:

$$\Gamma_{b1}^{\mathbf{h}} = \{\mathbf{g}(\mathbf{Z_i}) \in \mathscr{H}_b : \mathbf{g}(\mathbf{Z_i}) = \mathbf{h}(\mathbf{Z_i}) \text{ a.s.}\}.$$

The equivalence class under norm b2 contains all functions $\mathbf{g}(\mathbf{Z_i}) \in \mathscr{H}_b$ whose U-statistics projection mapping are equal to that of $\mathbf{h}(\mathbf{Z_i})$ a.s.:

$$\Gamma_{b2}^{\mathbf{h}} = \{\mathbf{g}(\mathbf{Z_i}) \in \mathscr{H}_b : \mathscr{M}[\mathbf{g}(\mathbf{Z_i})] = \mathscr{M}[\mathbf{h}(\mathbf{Z_i})] \text{ a.s.}\}.$$

The projections onto subspaces, $\Pi_{b1}\{\mathbf{h}(\mathbf{Z_i}) \mid \mathscr{U}_{b1}\}$ and $\Pi_{b2}\{\mathbf{h}(\mathbf{Z_i}) \mid \mathscr{U}_{b2}\}$, are unique up to their respective equivalence classes $\Gamma^{\mathbf{h}}_{b1}$ and $\Gamma^{\mathbf{h}}_{b2}$. Since all estimators in the same equivalence class deliver the same asymptotic variance (or efficiency) under the respective norm, it suffices to find one of them. Since the projection mapping $\mathscr{M}$ is many-to-one, i.e., different elements in $\mathscr{H}_b$ can be mapped to the same element in $\mathscr{H}_w$, the origin of $\mathscr{H}_b$ under inner product 2 is not the equivalence class of $\mathbf{h}(\mathbf{Z_i})$ with $\mathbf{h}(\mathbf{Z_i}) = \mathbf{0}$ a.s., but a larger one consisting of functions $\mathbf{h}(\mathbf{Z_i})$ such that $\mathscr{M}[\mathbf{h}(\mathbf{Z_i})] = \mathbf{0}$ a.s. (**see Supplement** for an example of $\mathbf{h}(\mathbf{Z_i}) \neq \mathbf{0}$, but $\mathscr{M}[\mathbf{h}(\mathbf{Z_i})] = \mathbf{0}$ a.s.).

Akin to the classical theory for within-subject attributes (Tsiatis, 2007), for model class 1, the I.F. $\psi(\mathbf{Z_i};\theta_0)$ is an element in $\mathscr{H}_b$, whose norm b1 is always larger than or equal to its projection onto a subspace $\mathscr{U}_{b1}$, hence, this projection $\Pi_{b1}\{\psi(\mathbf{Z_i};\theta_0) \mid \mathscr{U}_{b1}\}$ yields an RAL estimator with the minimum variance within class 1. Likewise, for model class 2, the projection of an I.F. onto $\mathscr{U}_{b2}$, $\Pi_{b2}\{\varphi(\mathbf{Z_i};\theta_0) \mid \mathscr{U}_{b2}\}$, has the smallest norm b2 thus also yields the efficient RAL estimator for class 2. And both $\Pi_{b1}\{\psi(\mathbf{Z_i};\theta_0) \mid \mathscr{U}_{b1}\}$ and $\Pi_{b2}\{\varphi(\mathbf{Z_i};\theta_0) \mid \mathscr{U}_{b2}\}$ are unique up to their respective equivalence classes.

4.5 Tangent Space and Dual Geometric Interpretation

The Hilbert space repositions the search for the efficient estimator to the efficient influence function with the smallest norm. Another important tool we implement is a "bridge" between parametric and semiparametric models, termed "parametric submodels" (Newey, 1990). We extend this idea to help find the semiparametric efficient estimator for between-subject attributes.

4.5.1 Parametric Submodels

The distribution of pairwise observations $\mathbf{Z_i} = \left(\mathbf{Z}_{i_1}^\top, \mathbf{Z}_{i_2}^\top\right)^\top$ can be characterized by $p_{\mathbf{Z}}(\mathbf{Z_i})$ that belongs to

$$\mathscr{P} = \{p_{\mathbf{Z}}(\mathbf{Z_i}; \beta, \eta(\cdot));\ \beta \in \mathbb{R}^q \text{ and } \eta(\cdot) \text{ is infinite-dimensional.}\} \tag{4.28}$$

Let $p_0(\mathbf{Z_i}; \theta_0) = p_{\mathbf{Z}}(\mathbf{Z_i}; \beta_0, \eta_0(\cdot))$ denote the truth, where β and $\eta(\cdot)$ are variationally independent as indicated previously. The infinite-dimensional nuisance parameter $\eta(\cdot)$ makes $\mathscr{P}$ a class of semiparametric models.

Now consider as if the data were generated from a conceptual class of parametric models, referred to as the *parametric submodels* (Newey, 1990):

$$\mathscr{P}_\gamma^{sub} = \{p_{\mathbf{Z}}(\mathbf{Z_i}; \beta, \gamma);\ \beta \in \mathbb{R}^q,\ \gamma \in \mathbb{R}^r\} \subset \mathscr{P}, \tag{4.29}$$

where $p_{\mathbf{Z}}(\mathbf{Z_i}; \beta_0, \gamma_0) = p_0(\mathbf{Z_i}; \theta_0) = p_{\mathbf{Z}}(\mathbf{Z_i}; \beta_0, \eta_0(\cdot))$ for some $\gamma_0 \in \mathbb{R}^r$. Thus the parametric submodels $\mathscr{P}_\gamma^{sub}$ of $\mathscr{P}$ are described by a finite-dimensional nuisance parameter vector $\gamma \in \mathbb{R}^r$ that contains the truth generating the data. Let $\theta^{sub} = \left(\beta^\top, \gamma^\top\right)^\top \in \mathbb{R}^p$ denote the parameter vector for the submodel with $p = q + r$.

Distinct from usual parametric models that can be used to describe real study data (by estimating model parameters), a parametric submodel cannot be applied to fit data since it involves the true but unknown parameter $\eta_0(\cdot)$.

An RAL estimator of β for a *semiparametric* model is an RAL estimator for every parametric submodel in $\mathscr{P}_\gamma^{sub}$ (Tsiatis, 2006). Unlike semiparametric models involving the infinite-dimensional $\eta(\cdot)$, parametric submodels are granted the well-defined score vectors at the truth θ_0:

$$\mathbf{S}_{\theta^{sub}}^{p\times 1}(\mathbf{Z_i}; \theta_0) = \left(\mathbf{S}_\beta^\top(\mathbf{Z_i}; \theta_0), \mathbf{S}_\gamma^\top(\mathbf{Z_i}; \theta_0)\right)^\top, \tag{4.30}$$

where $\mathbf{S}_{\theta^{sub}}(\mathbf{Z_i};\theta_0) = \partial \log p_0(\mathbf{Z_i};\theta_0)/\partial \theta^{sub\top}$, for θ^{sub} (or β,γ).

4.5.2 Tangent Spaces

Consider again $\mathscr{H}_b$ that consists of all measurable functions of $\mathbf{h}^{q\times 1}(\mathbf{Z_i})$ with mean zero and finite variances, equipped with both inner product 1: $\langle \mathbf{h}_1(\mathbf{Z_i}), \mathbf{h}_2(\mathbf{Z_i})\rangle_{b1} = E\left(\mathbf{h}_1^\top \mathbf{h}_2\right)$ and inner product 2: $\langle \mathbf{h}_1(\mathbf{Z_i}), \mathbf{h}_2(\mathbf{Z_i})\rangle_{b2} = E\left[2E\left(\mathbf{h}_1^\top \mid \mathbf{Z}_{i_1}\right)\cdot 2E\left(\mathbf{h}_2 \mid \mathbf{Z}_{i_1}\right)\right]$. We can then use the well-defined score vectors in (4.30) to span linear subspaces, termed *parametric submodel tangent spaces*.

Parametric Submodel Tangent Spaces

Non-overlap Model Class 1

The *parametric submodel tangent space* for model class 1 spanned by $\mathbf{S}_{\theta^{sub}}(\mathbf{Z_i};\theta_0)$ is a linear subspace of $\mathscr{H}_b$, where

$$\begin{aligned}
\pounds_{\beta\gamma}^{sub} &= \{\mathbf{B}\mathbf{S}_{\theta^{sub}}^{p\times 1}(\mathbf{Z_i};\theta_0);\ \forall \mathbf{B}^{q\times p}\} = \pounds_\beta \oplus \Lambda_\gamma,\\
\pounds_\beta &= \left\{\mathbf{B}\mathbf{S}_\beta^{q\times 1}(\mathbf{Z_i};\theta_0);\ \forall \mathbf{B}^{q\times q}\right\},\ \Lambda_\gamma = \left\{\mathbf{B}\mathbf{S}_\gamma^{r\times 1}(\mathbf{Z_i};\theta_0),\ \forall \mathbf{B}^{q\times r}\right\},
\end{aligned} \tag{4.31}$$

with $\oplus$ denoting the direct sum. Since $\theta^{sub} = \left(\beta^\top,\gamma^\top\right)^\top \in \mathbb{R}^p$, $\pounds_{\beta\gamma}^{sub}$ is the direct sum of two linear subspaces: $\pounds_\beta$, the *tangent space* for β; and Λ_γ, the *tangent space* for γ, also termed the *parametric submodel nuisance tangent space* (submodel n.t.s.).

Enumerated Model Class 2

The *parametric submodel tangent space* for model class 2 spanned by $\mathscr{M}\left[\mathbf{S}_{\theta^{sub}}(\mathbf{Z_i};\theta_0)\right]$ is

$$\begin{aligned}
\widetilde{\pounds}_{\beta\gamma}^{sub} &= \{\mathbf{B}\mathscr{M}\left[\mathbf{S}_{\theta^{sub}}(\mathbf{Z_i};\theta_0)\right];\ \forall \mathbf{B}^{q\times p}\} = \widetilde{\pounds}_\beta \oplus \widetilde{\Lambda}_\gamma,\\
\widetilde{\pounds}_\beta &= \left\{\mathbf{B}\mathscr{M}\left[\mathbf{S}_\beta^{q\times 1}(\mathbf{Z_i};\theta_0)\right];\ \forall \mathbf{B}^{q\times q}\right\},\ \widetilde{\Lambda}_\gamma = \left\{\mathbf{B}\mathscr{M}\left[\mathbf{S}_\gamma^{r\times 1}(\mathbf{Z_i};\theta_0)\right],\ \forall \mathbf{B}^{q\times r}\right\},
\end{aligned}$$

where $\widetilde{\pounds}_{\beta\gamma}^{sub}$, $\widetilde{\pounds}_{\beta}$ and $\widetilde{\Lambda}_{\gamma}$ are the respectively mapped subspaces from $\mathscr{H}_b$ to $\mathscr{H}_w$ with the U-statistics projection mapping $\mathscr{M}$ in (4.24).

Semiparametric Tangent Spaces

We are now in a position to define semiparametric tangent spaces. Since β is unchanged for a semiparametric model, $\pounds_{\beta}$ $\left(\widetilde{\pounds}_{\beta}\right)$ remains the same, but the nuisance tangent spaces need to be expanded to accommodate the infinite-dimensional nuisance parameter for semiparametric models.

Non-overlap Model Class 1

Let Υ be the collection of nuisance parameter γ for all possible parametric submodels in $\mathscr{P}_{\gamma}^{sub}$ defined in (4.29).

Definition. The *semiparametric nuisance tangent space* (semiparametric n.t.s.) Λ_{η} for model class 1 is the *mean-square closure* (in terms of the norm b1) of the unions of points $(\mathbf{h}(\mathbf{Z_i}))$ in all the parametric submodel nuisance tangent spaces, which, with a slight abuse of notation, we denote by $\Lambda^{\cup} = \cup_{\{\gamma\in\Upsilon\}}\Lambda_{\gamma}$. Specifically, Λ_{η} consists of all $\mathbf{h}(\mathbf{Z_i})$ in $\mathscr{H}_b$ for which there exists a sequence of $\mathbf{B}_j\mathbf{S}_{\gamma_j}(\mathbf{Z_i}) \in \Lambda^{\cup}$ $(j = 1,2,...)$ such that

$$\lim_{j\to\infty} \|\mathbf{h}^{q\times 1}(\mathbf{Z_i}) - \mathbf{B}_j^{q\times r_j}\mathbf{S}_{\gamma_j}^{r_j\times 1}(\mathbf{Z_i})\|_{b1}^2 = 0, \tag{4.32}$$

where $\mathbf{S}_{\gamma_j}^{r_j\times 1}(\mathbf{Z_i})$ corresponds to a sequence of submodels $\mathscr{P}_{\gamma_j}^{sub}$ characterized by $\gamma_j \in \mathbb{R}^{r_j}$, where each submodel and its associated dimension (r_j) are allowed to vary with j. Denote the whole *semiparametric tangent space for class 1* by $\pounds = \pounds_{\beta} \oplus \Lambda_{\eta}$.

Enumerated Model Class 2

The *semiparametric n.t.s.* for model class 2, denoted by $\widetilde{\Lambda}_{\eta}$, is the *mean-square closure* of $\widetilde{\Lambda}^{\cup} = \cup_{\{\gamma\in\Upsilon\}}\widetilde{\Lambda}_{\gamma}$. It consists of all $\mathbf{h}(\mathbf{Z}_i)$ in $\mathscr{H}_w$ which is either in $\widetilde{\Lambda}^{\cup}$ or the limit of a

convergent sequence $\mathbf{h}_j(\mathbf{Z}_i) \in \widetilde{\Lambda}^{\cup}$ ($j = 1, 2, ...$), i.e.,

$$\lim_{j\to\infty} \|\mathbf{h}(\mathbf{Z}_i) - \mathbf{h}_j(\mathbf{Z}_i)\|_w^2 = 0. \tag{4.33}$$

The *semiparametric tangent space for model class 2* is hence $\widetilde{\mathcal{L}} = \widetilde{\mathcal{L}}_\beta \oplus \widetilde{\Lambda}_\eta$.

Both Λ_η and $\widetilde{\Lambda}_\eta$ are closed by definition. The theorem below shows that linearity and closedness are preserved under the projection mapping $\mathscr{M}$.

Theorem 4.4 The *semiparametric n.t.s.* Λ_η and $\widetilde{\Lambda}_\eta$ are both linear subspaces and $\widetilde{\Lambda}_\eta = \mathscr{M}(\Lambda_\eta)$.

Therefore, all $\mathbf{h}(\mathbf{Z_i}) \in \mathscr{H}_b$ has a unique projection (up to its equivalence class) onto each subspace of Λ_η and $\widetilde{\Lambda}_\eta$.

4.5.3 Dual Geometric Interpretations for Semiparametric Models

We now show a fundamental connection, termed *dual orthogonality*, for the two classes of models, which is a direct generalization of the results for parametric models by leveraging the bridge of submodels (See **Appendix** for details). It geometrically characterizes the semiparametric RAL estimators through properties of their I.F.s with respect to their respective semiparametric nuisance tangent spaces.

Theorem 4.5 A semiparametric RAL estimator of β for either class of models must have an influence function (I.F.) $\varphi(\mathbf{Z_i})$ satisfying

$$\begin{aligned}
&\text{(i)} : \left\langle \varphi(\mathbf{Z_i}), \mathbf{S}_\beta(\mathbf{Z_i}; \theta_0) \right\rangle_{b1} = E\left[\varphi(\mathbf{Z_i})\mathbf{S}_\beta^\top(\mathbf{Z_i}; \theta_0)\right] = \mathbf{I}_q, \\
&\text{(ii)} : \Pi_{b1}\left\{\varphi(\mathbf{Z_i}) \mid \Lambda_\eta\right\} = \mathbf{0}, \\
&\text{(iii)} : \Pi_{b2}\left\{\varphi(\mathbf{Z_i}) \mid \widetilde{\Lambda}_\eta\right\} = \Pi_w\left\{2E\left[\varphi(\mathbf{Z_i}) \mid \mathbf{Z}_{i_1}\right] \mid \widetilde{\Lambda}_\eta\right\} = \mathbf{0},
\end{aligned} \tag{4.34}$$

where $\mathbf{I}_q$ is the $q \times q$ identity matrix, $\Pi_{b1}\left\{\varphi(\mathbf{Z_i}) \mid \Lambda_\eta\right\}$ is the unique projection (w.r.t. inner product 1) of $\varphi(\mathbf{Z_i})$ onto Λ_η, and $\Pi_{b2}\left\{\varphi(\mathbf{Z_i}) \mid \widetilde{\Lambda}_\eta\right\}$ is the unique projection (w.r.t. inner prod-

uct 2) of $\varphi(\mathbf{Z_i})$ onto $\widetilde{\Lambda}_\eta$, therefore, $\varphi(\mathbf{Z_i})$ is deemed *dual orthogonal* to both the semiparametric n.t.s. Λ_η (corresponding to model class 1) and its mapping$\widetilde{\Lambda}_\eta$ (for model class 2).

While Theorem 4.3 asserts that the two classes of models share the same RAL estimators and I.F.s., Theorem 4.5 further identifies such estimators through the dual orthogonality property for the I.F.s with respect to their respective semiparametric n.t.s. Λ_η and $\widetilde{\Lambda}_\eta$. Recall that the variance of any element is always larger than or equal to its projection onto a linear subspace in (4.22) and (4.27). This intrinsic connection between the two model classes allows us to locate the efficient estimator for model class 2 through that for model class 1, which serves as a "conjugate" model class as we now discuss.

4.6 Semiparametric Efficiency Bound

In this section, our goal is to identify the efficient semiparametric RAL estimator for the FRM in (4.2), or enumerated model class 2. Directly tackling the efficiency for class 2 is much more difficult, but the dual orthogonality motivates a strategy to find this estimator via the more straightforward model class 1. Although the efficient I.F. for class 2 corresponds to multiple I.F.s in class 1, our goal is fulfilled if we can identify one in the equivalence class of the efficient I.F. for model 1. In essence, we first establish the efficient I.F. for model class 1 and then show that its mapping is in the intended equivalence class for model 2.

Definition. The *efficient I.F.* is the unique influence function (up to its equivalence class) belonging to the tangent space that has the smallest asymptotic variance.

Recall that a semiparametric RAL estimator of β in $\mathscr{P}$ is an RAL estimator for *every* parametric submodel. In terms of influence functions, the class of I.F.s for a semiparametric model will be a subset of the class of I.F.s for *all* parametric submodels. Hence, the asymptotic variance of a semiparametric model must be greater than or equal to the parametric efficiency bound for any submodel, or the *supremum* of such bounds for all submodels. We define the semiparametric efficiency bound via the bridge of parametric submodels for each model class.

4.6.1 Parametric Submodels

Non-overlap Model Class 1

The *efficient I.F.* for a *parametric* submodel in class 1, denoted by $\varphi^{sub}_{\gamma,\text{eff1}}(\mathbf{Z_i};\theta_0)$, is the unique I.F. in the tangent space $\pounds^{sub}_{\beta\gamma} = \pounds_\beta \oplus \Lambda_\gamma$ with the smallest norm b1, i.e., for any I.F. $\varphi^{sub}(\mathbf{Z_i};\theta_0)$ of a submodel in $\mathscr{P}^{sub}_\gamma$,

$$\left\|\varphi^{sub}_{\gamma,\text{eff1}}(\mathbf{Z_i};\theta_0)\right\|^2_{b1} \leq \left\|\varphi^{sub}(\mathbf{Z_i};\theta_0)\right\|^2_{b1},\ \varphi^{sub}_{\gamma,\text{eff1}}(\mathbf{Z_i};\theta_0) = \Pi_{b1}\left\{\varphi^{sub}(\mathbf{Z_i};\theta_0) \mid \pounds^{sub}_{\beta\gamma}\right\},$$

then the *efficiency bound* for *parametric* submodels of class 1 is its variance

$$\upsilon^{sub}_{1,\gamma} = Var\left[\varphi^{sub}_{\gamma,\text{eff1}}(\mathbf{Z_i};\theta_0)\right].$$

The *semiparametric efficiency bound* for non-overlap model class 1 is defined as the supremum of $\upsilon^{sub}_{1,\gamma}$over all submodels:

$$\upsilon_1 = \sup_{\left\{\mathscr{P}^{sub}_\gamma\right\}} \upsilon^{sub}_{1,\gamma} = \sup_{\left\{\mathscr{P}^{sub}_\gamma\right\}} Var\left[\varphi^{sub}_{\gamma,\text{eff1}}(\mathbf{Z_i};\theta_0)\right], \tag{4.35}$$

where sup is defined based on the non-negative definite criterion for comparing matrices using their differences.

Enumerated Model Class 2

Likewise, the *efficient I.F.* for *parametric* submodels in class 2, $\psi^{sub}_{\gamma,\text{eff2}}(\mathbf{Z_i};\theta_0)$, is the I.F. lying in $\pounds^{sub}_{\beta\gamma} = \pounds_\beta \oplus \Lambda_\gamma$ with the smallest norm b2. Hence, any I.F. $\psi^{sub}(\mathbf{Z_i};\theta_0)$ of a submodel satisfies

$$\left\|\psi^{sub}_{\gamma,\text{eff2}}(\mathbf{Z_i};\theta_0)\right\|^2_{b2} = \left\|\mathscr{M}\left[\psi^{sub}_{\gamma,\text{eff2}}(\mathbf{Z_i};\theta_0)\right]\right\|^2_w \leq \left\|\mathscr{M}\left[\psi^{sub}(\mathbf{Z_i};\theta_0)\right]\right\|^2_w = \left\|\psi^{sub}(\mathbf{Z_i};\theta_0)\right\|^2_{b2}.$$

By the multivariate Pythagoras, any two I.F.s with zero difference in norm b2 are

equivalent as they determine the same efficiency (asymptotic variance). The *equivalence class* for $\psi^{sub}_{\gamma,\text{eff2}}(\mathbf{Z_i};\theta_0)$ is hence defined to be

$$\Gamma^{sub}_{\text{eff2}} = \left\{ \psi^{sub}(\mathbf{Z_i};\theta_0) \in \pounds^{sub}_{\beta\gamma} : \mathscr{M}\left[\psi^{sub}(\mathbf{Z_i};\theta_0)\right] = \mathscr{M}\left[\psi^{sub}_{\gamma,\text{eff2}}(\mathbf{Z_i};\theta_0)\right] \text{ a.s.} \right\}, \tag{4.36}$$

$\psi^{sub}_{\gamma,\text{eff2}}(\mathbf{Z_i};\theta_0)$ is unique up to this equivalence class $\Gamma^{sub}_{\text{eff2}}$, and the *efficiency bound* for *parametric* submodels of class 2 is defined by

$$\upsilon^{sub}_{2,\gamma} = Var\left\{ \mathscr{M}\left[\psi^{sub}_{\gamma,\text{eff2}}(\mathbf{Z_i};\theta_0)\right] \right\}.$$

The associated *semiparametric efficiency bound* for class 2 is defined by the supremum $\upsilon_2 = \sup_{\left\{\mathscr{P}^{sub}_{\gamma}\right\}} \upsilon^{sub}_{2,\gamma}$. The theorem below connects the two classes of submodels regarding the efficient I.F., with a proof in the **Appendix**.

Theorem 4.6 The norm b2 of the efficient I.F. for class 1, $\varphi^{sub}_{\gamma,\text{eff1}}(\mathbf{Z_i};\theta_0)$, equals the norm b2 of the efficient I.F. for class 2, $\psi^{sub}_{\gamma,\text{eff2}}(\mathbf{Z_i};\theta_0)$, hence $\varphi^{sub}_{\gamma,\text{eff1}}(\mathbf{Z_i};\theta_0)$ is in the equivalence class $\Gamma^{sub}_{\text{eff2}}$ defined in (4.36), i.e.,

$$\left\| \varphi^{sub}_{\gamma,\text{eff1}}(\mathbf{Z_i};\theta_0) \right\|^2_{b2} = \left\| \psi^{sub}_{\gamma,\text{eff2}}(\mathbf{Z_i};\theta_0) \right\|^2_{b2}, \ \varphi^{sub}_{\gamma,\text{eff1}}(\mathbf{Z_i};\theta_0) \in \Gamma^{sub}_{\text{eff2}}.$$

$\varphi^{sub}_{\gamma,\text{eff1}}(\mathbf{Z_i};\theta_0)$ is already shown to be a valid I.F. for model 2, now with the same norm b2 as $\psi^{sub}_{\gamma,\text{eff2}}(\mathbf{Z_i};\theta_0)$, it is indeed in the equivalence class $\Gamma^{sub}_{\text{eff2}}$. It now follows from Theorem 6 that after the mapping, $\mathscr{M}\left[\varphi^{sub}_{\gamma,\text{eff1}}(\mathbf{Z_i};\theta_0)\right] = \mathscr{M}\left[\psi^{sub}_{\gamma,\text{eff2}}(\mathbf{Z_i};\theta_0)\right]$ a.s., hence they determine the same asymptotic variance. Essentially, $\varphi^{sub}_{\gamma,\text{eff1}}(\mathbf{Z_i};\theta_0)$ delivers exactly what we aim to find: one element lying in the submodel tangent space $\pounds^{sub}_{\gamma}$ that yields the efficiency for the model class 2.

4.6.2 Semiparametric Models

Now we switch from parametric submodels to semiparametric models using the semi-

parametric n.t.s. Λ_η defined in (4.32). For notational brevity, we drop the superscripts (of "sub") and subscripts γ for quantities of semiparametric models to differentiate from those of submodels. By definition, the *efficient I.F.* for the *semiparametric* non-overlap model 1 is the I.F. in $\pounds$ whose variance achieves the semiparametric efficiency bound υ_1. With variationally independent parameters $\theta = \{\beta, \eta(\cdot)\}$, the *efficient score* is shown to be the residual of the score vector for β after projecting it onto the nuisance tangent space (Tsiatis, 2006). For model class 1, the *semiparametric efficient score* is hence

$$\mathbf{S}_{\text{eff1}}(\mathbf{Z_i}; \theta_0) = \mathbf{S}_\beta(\mathbf{Z_i}; \theta_0) - \Pi_{b1}\left\{\mathbf{S}_\beta(\mathbf{Z_i}; \theta_0) \mid \Lambda_\eta\right\}. \tag{4.37}$$

The theorem below shows how to find the efficient I.F. for the semiparametric model class 1, with a proof in the **Appendix**.

Theorem 4.7 Let

$$\varphi_{\text{eff1}}(\mathbf{Z_i}; \theta_0) = E^{-1}\left(\mathbf{S}_{\text{eff1}}\mathbf{S}_{\text{eff1}}^\top\right)\mathbf{S}_{\text{eff1}}(\mathbf{Z_i}; \theta_0). \tag{4.38}$$

Then $\varphi_{\text{eff1}}(\mathbf{Z_i}; \theta_0)$ is the unique element in $\pounds = \pounds_\beta \oplus \Lambda_\eta$ whose variance achieves υ_1.

Akin to submodels, this semiparametric efficient I.F. $\varphi_{\text{eff1}}(\mathbf{Z_i}; \theta_0)$ for model class 1 is also mapped to an element in $\widetilde{\pounds}$ that achieves the efficiency for the model class 2, as summarized in the following theorem.

Theorem 4.8 Let $\psi_{\text{eff2}}(\mathbf{Z_i}; \theta_0)$ denote the efficient I.F. for the *semiparametric* model class 2. Then $\varphi_{\text{eff1}}(\mathbf{Z_i}; \theta_0)$ has the same norm b2 as $\psi_{\text{eff2}}(\mathbf{Z_i}; \theta_0)$ and hence is in its equivalence class denoted by Γ_{eff2}, i.e.,

$$\|\varphi_{\text{eff1}}(\mathbf{Z_i}; \theta_0)\|_{b2}^2 = \|\psi_{\text{eff2}}(\mathbf{Z_i}; \theta_0)\|_{b2}^2, \text{ or } \|\mathscr{M}[\varphi_{\text{eff1}}(\mathbf{Z_i}; \theta_0)]\|_w^2 = \|\mathscr{M}[\psi_{\text{eff2}}(\mathbf{Z_i}; \theta_0)]\|_w^2,$$

where $\Gamma_{\text{eff2}} = \{\psi(\mathbf{Z_i}; \theta_0) \in \pounds : \mathscr{M}[\psi(\mathbf{Z_i}; \theta_0)] = \mathscr{M}[\psi_{\text{eff2}}(\mathbf{Z_i}; \theta_0)] \text{ a.s.}\}$.

Again, the multivariate Pythagoras implies that the variance of $\mathscr{M}\left[\varphi_{\text{eff1}}\left(\mathbf{Z_i};\theta_0\right)\right]$ equals the *semiparametric efficiency bound* υ_2 for enumerated model class 2, i.e.,

$$Var\left\{\mathscr{M}\left[\varphi_{\text{eff1}}\left(\mathbf{Z_i};\theta_0\right)\right]\right\} = Var\left\{\mathscr{M}\left[\psi_{\text{eff2}}\left(\mathbf{Z_i};\theta_0\right)\right]\right\} = \upsilon_2.$$

Based on Theorem 4.8, we can identify the efficient estimator for model class 2 via that for model class 1. In model class 1, it is more straightforward to construct the efficient I.F. $\varphi_{\text{eff1}}\left(\mathbf{Z_i};\theta_0\right)$.

4.7 The Efficiency for the FRM

By the conditions in Theorem 4.5 that any I.F. satisfies, to derive the semiparametric efficient estimator $\varphi_{\text{eff1}}\left(\mathbf{Z_i};\theta_0\right)$ for model class 1, we first identify the specific form of the semiparametric n.t.s. Λ_η and then find elements that are orthogonal to it, which form a pool of candidates for the optimal one.

Let $\mathbf{Z_i} = (\mathbf{Y_i}^\top, \mathbf{X_i}^\top)^\top$, $\mathbf{Y_i} = (\mathbf{Y}_{i_1}^\top, \mathbf{Y}_{i_2}^\top)^\top$, $\mathbf{X_i} = (\mathbf{X}_{i_1}^\top, \mathbf{X}_{i_2}^\top)^\top$, $\mathbf{i} = (i_1, i_2) \in C_2^n$, where $\mathbf{X}_i$ $(\mathbf{Y}_i)$ is a $q \times 1$ $(m \times 1)$ vector of explanatory variables (outcomes) for the i-th subject. Let $f_{\mathbf{i}}(\mathbf{Y}_{i_1}, \mathbf{Y}_{i_2})$ be a univariate continuous response for the $\mathbf{i}$-th pair such as the microbiome Beta-diversity in (2.2) (same considerations apply to more general types, see Tsiatis (2006)). The semiparametric FRM in this case is

$$f_{\mathbf{i}} = h\left(\mathbf{X_i};\beta\right) + \varepsilon_{\mathbf{i}},\ E(\varepsilon_{\mathbf{i}} \mid \mathbf{X_i}) = 0,\ \mathbf{i} = (i_1, i_2) \in C_2^n. \tag{4.39}$$

The goal is to identify the semiparametric RAL estimator of β with the smallest variance for the FRM through $\varphi_{\text{eff1}}\left(\varepsilon_{\mathbf{i}}, \mathbf{X_i};\theta_0\right)$.

4.7.1 Identifying Λ_η with the Joint Likelihood and Score

The joint density of the observed outcomes for pairs, $(\varepsilon_{\mathbf{i}}, \mathbf{X}_{\mathbf{i}})$, where $\varepsilon_{\mathbf{i}} = f_{\mathbf{i}} - h(\mathbf{X}_{\mathbf{i}}, \beta)$, belongs to a class of semiparametric models

$$\mathscr{P} = \left\{ p_{\varepsilon,\mathbf{X}}(\varepsilon_{\mathbf{i}}, \mathbf{X}_{\mathbf{i}}; \beta, \eta(\cdot));\ \beta \in \mathbb{R}^q \text{ and } \eta(\cdot) \text{ is infinite-dimensional} \right\}. \tag{4.40}$$

We assume that the underlying true data are generated from $p(\mathbf{Y}_i, \mathbf{X}_i; \theta_0)$, which induces $p(\mathbf{Y}_{\mathbf{i}}, \mathbf{X}_{\mathbf{i}}; \theta_0) = p(\mathbf{Y}_{i_1}, \mathbf{X}_{i_1}) p(\mathbf{Y}_{i_2}, \mathbf{X}_{i_2})$. By independence and the change of variables, θ_0 remains the same for describing the individual-level $p(\mathbf{Y}_i, \mathbf{X}_i)$ and pairwise-level $p(\mathbf{Y}_{\mathbf{i}}, \mathbf{X}_{\mathbf{i}})$ or $p(\varepsilon_{\mathbf{i}}, \mathbf{X}_{\mathbf{i}})$ (see **Appendix** for details). So we denote the truth by $p_{\varepsilon,\mathbf{X}}(\varepsilon_{\mathbf{i}}, \mathbf{X}_{\mathbf{i}}; \theta_0)$. Its parametric submodels are given by

$$\mathscr{P}_\gamma^{sub} = \{ p_{\varepsilon,\mathbf{X}}(\varepsilon_{\mathbf{i}}, \mathbf{X}_{\mathbf{i}}; \beta, \gamma);\ \beta \in \mathbb{R}^q,\ \gamma \in \mathbb{R}^r \} \subset \mathscr{P},$$

which contain the truth $\theta_0 = \{\beta_0, \eta_0(\cdot)\}$ for all γ. Let Λ_η denote the semiparametric n.t.s. for model class 1 resulting from the mean-square closure of the parametric submodels n.t.s. $\Lambda^{\cup} = \cup_{\{\gamma \in \Upsilon\}} \Lambda_\gamma$. We can readily determine the form of Λ_η by applying arguments similar to those for the classical within-subject semiparametric models (see Chapter 4 of Tsiatis (2006)), summarized by the theorem below.

Theorem 4.9 The space Λ_η contains all mean-zero functions $\lambda(\varepsilon_{\mathbf{i}}, \mathbf{X}_{\mathbf{i}})$ satisfying the constraint on the conditional mean in (4.39), namely,

$$\Lambda_\eta = \left\{ \lambda^{q\times 1}(\varepsilon_{\mathbf{i}}, \mathbf{X}_{\mathbf{i}}) : E\left[\lambda(\varepsilon_{\mathbf{i}}, \mathbf{X}_{\mathbf{i}}) \varepsilon_{\mathbf{i}} \mid \mathbf{X}_{\mathbf{i}} \right] = \mathbf{0}^{q\times 1} \right\}. \tag{4.41}$$

Its orthogonal complement (w.r.t. inner product 1) is defined by

$$\Lambda_\eta^{\perp} = \left\{ \chi^{q\times 1}(\varepsilon_{\mathbf{i}}, \mathbf{X}_{\mathbf{i}}) \in \mathscr{H}_b : \langle \chi(\varepsilon_{\mathbf{i}}, \mathbf{X}_{\mathbf{i}}), \lambda(\varepsilon_{\mathbf{i}}, \mathbf{X}_{\mathbf{i}}) \rangle_{b1} = 0 \right\}.$$

The form of Λ_η for model class 1 above is conformable with that for the semiparametric GLM in Tsiatis (2006), as both models have restrictions only on the conditional mean.

4.7.2 The Efficient Influence Function of the FRM

Recall that the efficient score is the residual after projecting $\mathbf{S}_\beta$ onto Λ_η by (4.37). In $\mathscr{H}_b$, the projection (w.r.t. inner product 1) of an arbitrary element $g(\varepsilon_\mathbf{i}, \mathbf{X_i}) \in \mathscr{H}_b$ onto Λ_η is readily shown to satisfy:

$$\Pi_{b1}\left\{g(\varepsilon_\mathbf{i}, \mathbf{X_i}) \mid \Lambda_\eta^\perp\right\} = g - \Pi_{b1}\left\{g(\varepsilon_\mathbf{i}, \mathbf{X_i}) \mid \Lambda_\eta\right\} = E\left[g(\varepsilon_\mathbf{i}, \mathbf{X_i})\,\varepsilon_\mathbf{i} \mid \mathbf{X_i}\right] E^{-1}(\varepsilon_\mathbf{i}^2 \mid \mathbf{X_i})\varepsilon_\mathbf{i}, \quad (4.42)$$

which is verified by the fact that $\left\langle \Pi_{b1}\left\{g \mid \Lambda_\eta^\perp\right\}, \lambda^*(\varepsilon_\mathbf{i}, \mathbf{X_i})\right\rangle_{b1} = 0$ for any $\lambda^*(\varepsilon, \mathbf{X}) \in \Lambda_\eta$. Substituting $\mathbf{S}_\beta(\varepsilon_\mathbf{i}, \mathbf{X_i})$ in place of $g(\varepsilon_\mathbf{i}, \mathbf{X_i})$ in (4.42) yields the efficient score for model 1:

$$\mathbf{S}_{\text{eff1}}(\varepsilon_\mathbf{i}, \mathbf{X_i}; \theta_0) = \mathbf{S}_\beta - \Pi_{b1}\left\{\mathbf{S}_\beta(\varepsilon_\mathbf{i}, \mathbf{X_i}) \mid \Lambda_\eta\right\} = E\left[\mathbf{S}_\beta(\varepsilon_\mathbf{i}, \mathbf{X_i})\,\varepsilon_\mathbf{i} \mid \mathbf{X_i}\right] V^{-1}(\mathbf{X_i})\,\varepsilon_\mathbf{i}, \quad (4.43)$$

where $V(\mathbf{X_i}) = E\left(\varepsilon_\mathbf{i}^2 \mid \mathbf{X_i}\right)$. By fixing $\eta(\cdot)$ at the truth $\eta_0(\cdot)$ and taking partial derivatives w.r.t. β of the conditional mean restriction $E\left[f_\mathbf{i} - h(\mathbf{X_i}; \beta) \mid \mathbf{X_i}\right] = 0$, we obtain

$$E\left[\varepsilon_\mathbf{i}\mathbf{S}_\beta^\top(\varepsilon_\mathbf{i}, \mathbf{X_i}) \mid \mathbf{X_i}\right] = \frac{\partial}{\partial \beta^\top} h(\mathbf{X_i}; \beta_0) \stackrel{\text{def}}{=} \mathbf{D}(\mathbf{X_i}), \quad (4.44)$$

which is the partial derivatives of β for the mean function $h(\mathbf{X_i}; \beta_0)$ in (4.39). Then the efficient score in (4.43) simplifies to

$$\mathbf{S}_{\text{eff1}}(\varepsilon_\mathbf{i}, \mathbf{X_i}; \theta_0) = E\left[\mathbf{S}_\beta(\varepsilon_\mathbf{i}, \mathbf{X_i})\,\varepsilon_\mathbf{i} \mid \mathbf{X_i}\right] V^{-1}(\mathbf{X_i})\,\varepsilon_\mathbf{i} = \mathbf{D}^\top(\mathbf{X_i}) V^{-1}(\mathbf{X_i})\,\varepsilon_\mathbf{i}. \quad (4.45)$$

By (4.38) in Theorem 7, the unique efficient I.F. for model class 1 is obtained by scaling $\mathbf{S}_{\text{eff1}}(\varepsilon_{\mathbf{i}}, \mathbf{X}_{\mathbf{i}}; \theta_0)$:

$$\varphi_{\text{eff1}}(\varepsilon_{\mathbf{i}}, \mathbf{X}_{\mathbf{i}}; \theta_0) = E^{-1}\left(\mathbf{S}_{\text{eff1}}\mathbf{S}_{\text{eff1}}^{\top}\right)\mathbf{S}_{\text{eff1}} = E^{-1}\left(\mathbf{D}_{\mathbf{i}}^{\top}V_{\mathbf{i}}^{-1}\mathbf{D}_{\mathbf{i}}\right)\mathbf{D}_{\mathbf{i}}^{\top}V_{\mathbf{i}}^{-1}\left[f_{\mathbf{i}} - h(\mathbf{X}_{\mathbf{i}}; \beta_0)\right], \quad (4.46)$$

which is easily verified to satisfy (i) - (iii) in (4.34).

By Theorem 4.8, this semiparametric efficient I.F. $\varphi_{\text{eff1}}(\varepsilon_{\mathbf{i}}, \mathbf{X}_{\mathbf{i}}; \theta_0)$ is in the equivalence class of the efficient I.F. for model class 2, thus achieving the *semiparametric efficiency bound* υ_2 :

$$\upsilon_2 = Var\{\mathscr{M}[\varphi_{\text{eff1}}(\varepsilon_{\mathbf{i}}, \mathbf{X}_{\mathbf{i}}; \theta_0)]\} = Var[2E(\varphi_{\text{eff1}}(\varepsilon_{\mathbf{i}}, \mathbf{X}_{\mathbf{i}}; \theta_0) \mid \mathbf{Z}_{i_1})] = \mathbf{B}^{-1}\Sigma_U\mathbf{B}^{-1}, \quad (4.47)$$

where

$$\mathbf{B} = E\left[\mathbf{D}^{\top}(\mathbf{X}_{\mathbf{i}})V^{-1}(\mathbf{X}_{\mathbf{i}})\mathbf{D}(\mathbf{X}_{\mathbf{i}})\right],\ \widetilde{\mathbf{v}}_{i_1} = 2E\left\{\mathbf{D}^{\top}(\mathbf{X}_{\mathbf{i}})V(\mathbf{X}_{\mathbf{i}})^{-1}\left[f_{\mathbf{i}} - \mu(\mathbf{X}_{\mathbf{i}}, \beta_0)\right] \mid \mathbf{Z}_{i_1}\right\},$$
$$\Sigma_U = Var(\widetilde{\mathbf{v}}_{i_1}) = \mathbf{v}_{i_1}\mathbf{v}_{i_1}^{\top},\ \mathbf{i} = (i_1, i_2) \in C_2^n,\ \mathbf{Z}_{i_1} = (\mathbf{Y}_{i_1}^{\top}, \mathbf{X}_{i_1}^{\top})^{\top}. \quad (4.48)$$

Consequently, the efficient score equations

$$\sum_{\mathbf{i}\in C_2^n}\mathbf{S}_{\text{eff1}}(\varepsilon_{\mathbf{i}}, \mathbf{X}_{\mathbf{i}}) = \sum_{\mathbf{i}\in C_2^n}\mathbf{D}^{\top}(\mathbf{X}_{\mathbf{i}})V^{-1}(\mathbf{X}_{\mathbf{i}})\left[f_{\mathbf{i}} - h(\mathbf{X}_{\mathbf{i}}, \beta)\right] = \mathbf{0}, \quad (4.49)$$

yield an estimator $\widehat{\beta}_{\text{eff}}$ whose variance (after mapping) is the smallest among all semiparametric RAL estimators of the FRM.

This υ_2 coincides with $\Sigma_{\beta}^{\text{ugee}}$ in (4.13), which is the asymptotic variance of the UGEE estimator in Theorem 4.2. Hence, the UGEE in (4.12) for between-subject FRM is the efficient estimating equation (4.49), and the resulting UGEE estimator does achieve the semiparametric efficiency bound υ_2.

4.8 Examples of Efficient UGEE

In this section, we demonstrate that semiparametric UGEE estimators for continuous responses do achieve the efficiency bound. For space consideration, examples of binary or count responses are included in the **Supplements.**

4.8.1 Exogenous Between-subject Responses

Consider a classical linear regression

$$Y_i = X_i\beta + \varepsilon_i,\ \varepsilon_i \sim^{i.i.d} N\left(0, \sigma_Y^2\right),\ 1 \le i \le n.$$

For simplicity, we assume $X_i \sim^{i.i.d} N\left(0, \sigma_X^2\right)$. The maximum likelihood estimator (MLE) of β reaches the Cramér-Rao (CR) bound $\sigma_Y^2\sigma_X^{-2}$.

Let $f_{\mathbf{i}} = Y_{i_1} - Y_{i_2}$ and $X_{\mathbf{i}} = X_{i_1} - X_{i_2}$ for $\mathbf{i} = (i_1, i_2) \in C_2^n$. Consider an FRM: $E\left(f_{\mathbf{i}} \mid X_{\mathbf{i}}\right) = X_{\mathbf{i}}\beta$. Let

$$S_{\mathbf{i}} = f_{\mathbf{i}} - X_{\mathbf{i}}\beta,\ D_{\mathbf{i}} = \frac{\partial}{\partial\beta}\left(X_{\mathbf{i}}\beta\right) = X_{\mathbf{i}},\ V_{\mathbf{i}} = Var\left(f_{\mathbf{i}}\right) = 2\sigma_Y^2.$$

The UGEE and associated I.F. for this exogenous FRM are given by

$$U_n(\beta) = \sum_{\mathbf{i}\in C_2^n} D_{\mathbf{i}}V_{\mathbf{i}}^{-1}S_{\mathbf{i}} = \sum_{\mathbf{i}\in C_2^n} X_{\mathbf{i}}\left(2\sigma_Y^2\right)^{-1}\left(f_{\mathbf{i}} - X_{\mathbf{i}}\beta\right) = 0,$$

$$\varphi_{\text{ugee}}\left(\varepsilon_{\mathbf{i}}, X_{\mathbf{i}}; \beta_0\right) = E(X_{\mathbf{i}}V_{\mathbf{i}}^{-1}X_{\mathbf{i}})X_{\mathbf{i}}V_{\mathbf{i}}^{-1}\left(f_{\mathbf{i}} - X_{\mathbf{i}}\beta_0\right) = \left(2\sigma_X^2\right)^{-1}\left(\varepsilon_{\mathbf{i}}X_{\mathbf{i}}\right).$$

The asymptotic variance calculated based on norm b2 is

$$\upsilon_2 = \left\|\varphi_{\text{ugee}}\left(\varepsilon_{\mathbf{i}}, X_{\mathbf{i}}; \beta_0\right)\right\|_{b2}^2 = Var\left[2E\left(\varphi_{\text{ugee}}\left(\varepsilon_{\mathbf{i}}, X_{\mathbf{i}}; \beta_0\right) \mid \varepsilon_{i_1}, X_{i_1}\right)\right] = \sigma_Y^2\sigma_X^{-2},$$

which is exactly the same as the CR bound for the MLE of β for the classic linear regression. The semiparamatric UGEE estimator is efficient.

4.8.2 Endogenous Between-subject Responses

Now consider (identically but not independently distributed) endogenous between-subject responses $f_{\mathbf{i}} = f_{i_1,i_2} \sim^{i.d.} (\mu, \sigma^2)$, where, unlike the exogenous example above, subject-level outcomes may be latent. Let $\beta = (\mu, \sigma^2)^\top$ be the parameters of interest and $\beta_0 = (\mu_0, \sigma_0^2)^\top$ denote the truth. To obtain the efficient (parametric) estimator for β as our benchmark for this case, assume $f_{\mathbf{i}} \sim^{i.d.} N(\mu, \sigma^2)$. The efficient (parametric) I.F. in model class 1 is

$$\varphi_{\text{eff1}}(f_{\mathbf{i}}) = \left(f_{\mathbf{i}} - \mu_0,\ -\sigma_0^2 + (f_{\mathbf{i}} - \mu_0)^2\right)^\top, \tag{4.50}$$

which is also in the equivalent class of the efficient I.F. $\psi_{\text{eff2}}(f_{\mathbf{i}})$ for the model class 2 with the variance (based on norm b2)

$$\Sigma_{\beta_0}^{\text{eff2}} = 4E\left[E\left(\varphi_{\text{eff1}}(f_{\mathbf{i}}) \mid f_{i_1}\right) E\left(\varphi_{\text{eff1}}^\top(f_{\mathbf{i}}) \mid f_{i_1}\right)\right].$$

For endogenous responses where the benchmark based on individuals is intractable, we use this $\Sigma_{\beta_0}^{\text{eff2}}$ (from a parametric model) as the efficiency bound.

Now consider a semiparametric FRM with $E(f_{\mathbf{i}}) = \mu$, $E\left[(f_{\mathbf{i}} - \mu)^2\right] = \sigma^2$, $\mathbf{i} = (i_1, i_2) \in C_2^n$, and let

$$\mathbf{S}_{\mathbf{i}} = \left(f_{\mathbf{i}} - \mu,\ (f_{\mathbf{i}} - \mu)^2 - \sigma^2\right)^\top,\ \mathbf{D}_{\mathbf{i}} = \frac{\partial}{\partial \beta^\top}\beta,\ \mathbf{V}_{\mathbf{i}} = diag\left(Var(f_{\mathbf{i}}),\ Var\left[(f_{\mathbf{i}} - \mu)^2\right]\right), \tag{4.51}$$

The UGEE, the resulting estimator, and the associated I.F. for the FRM are

$$\mathbf{U}_n(\beta) = \sum_{\mathbf{i} \in C_2^n} \mathbf{D}_{\mathbf{i}}^\top \mathbf{V}_{\mathbf{i}}^{-1} \mathbf{S}_{\mathbf{i}} = \mathbf{0},\ \widehat{\beta}_f^{\text{ugee}} = \binom{n}{2}^{-1} \sum_{\mathbf{i} \in C_2^n} \left(f_{\mathbf{i}},\ \left(f_{\mathbf{i}} - \overline{f_{\mathbf{i}}}\right)^2\right)^\top,$$
$$\varphi_{\text{ugee}}(f_{\mathbf{i}}) = \left(f_{\mathbf{i}} - \mu_0,\ -\sigma_0^2 + (f_{\mathbf{i}} - \mu_0)^2\right)^\top.$$

Since $\varphi_{\text{ugee}}(f_{\mathbf{i}}) = \varphi_{\text{eff1}}(f_{\mathbf{i}})$ in (4.50), this UGEE estimator does achieve the benchmark $\Sigma_{\beta_0}^{\text{eff2}}$

and hence is optimal. Therefore, in the endogenous case, UGEE also yields the most efficient semiparametric RAL estimator.

4.9 Adaptive Estimator for the FRM

We first clarify the concept of local and global efficiency. *Local efficiency* refers to the efficiency for particular assumptions of the nonparametric component of the model (Newey, 1990). Such estimators are optimal for a particular distribution, subject to the constraint implied by the semiparametric model (Tsiatis and Ma, 2004; Robinson, 1988), while the more ambitious *global efficiency* refers to the efficiency for all values of the nonparametric component (Bickel, 1982).

We define local and global efficiency for FRM in the same vein as for within-subject models. Namely, any semiparametric RAL estimator $\widehat{\beta}$ with the asymptotic variance achieving the bound v_2 in (4.47) for the true model $p_0(f_\mathbf{i}, \mathbf{X_i}) = p\left(f_\mathbf{i}, \mathbf{X_i}; \theta_0\right)$ is *locally efficient* at $p_0\left(f_\mathbf{i}, \mathbf{X_i}\right)$. If the same $\widehat{\beta}$ is semiparametric efficient regardless of $p_0(f_\mathbf{i}, \mathbf{X_i}) \in \mathscr{P}$, then it is *globally efficient*. For FRM, the nonparametric component refers to the unknown true conditional distribution $p_0\left(f_\mathbf{i} \mid \mathbf{X_i}\right)$ left unspecified, which yields an unknown conditional variance $V\left(\mathbf{X_i}\right) = Var\left(f_\mathbf{i} \mid \mathbf{X_i}\right)$. As in the case of models for within-subject attributes?, adaptive estimators can be used to find approximations to this variance by imposing additional working variance assumptions to improve efficiency as shown in our simulations (see Section 4.9.3 and Supplements). In the following, we demonstrate global and local efficiency for FRM.

4.9.1 Globally Efficient Estimators

Example 1. (Binary responses) Consider an FRM for binary responses $f_\mathbf{i}$ with a vector of explanatory variables $\mathbf{X_i}$, where $E\left(f_\mathbf{i} \mid \mathbf{X_i}\right) = \text{expit}\left(\beta^\top \mathbf{X_i}\right) = \exp\left(\beta^\top \mathbf{X_i}\right)\left[1 + \exp\left(\beta^\top \mathbf{X_i}\right)\right]^{-1}$. The variance of the binary $f_\mathbf{i}$ conditional on $\mathbf{X_i}$ takes the form

$$V\left(\mathbf{X_i}; \beta\right) = \exp\left(\beta^\top \mathbf{X_i}\right)\left[1 + \exp\left(\beta^\top \mathbf{X_i}\right)\right]^{-2}, \tag{4.52}$$

which does not involve any additional unknown parameter (aside from β). By (4.46), the optimal UGEE is:

$$\sum_{\mathbf{i}\in C_2^n} \mathbf{D}_\mathbf{i}^\top V_\mathbf{i}^{-1} S_\mathbf{i} = \sum_{\mathbf{i}\in C_2^n} \mathbf{X}_\mathbf{i}\left[f_\mathbf{i} - \text{expit}\left(\beta^\top \mathbf{X}_\mathbf{i}\right)\right] = \mathbf{0}.$$

Since the above only contains β with no other parameter, the resulting UGEE estimator $\widehat{\beta}$ has the efficient I.F. depending only on β_0 :

$$\varphi_{\text{eff1}}\left(f_\mathbf{i}, \mathbf{X}_\mathbf{i}; \beta_0\right) = E^{-1}\left[\mathbf{X}_\mathbf{i} V(\mathbf{X}_\mathbf{i}; \beta_0)\mathbf{X}_\mathbf{i}^\top\right]\mathbf{X}_\mathbf{i}\left[f_\mathbf{i} - \text{expit}\left(\beta_0^\top \mathbf{X}_\mathbf{i}\right)\right].$$

This $\widehat{\beta}$ is semiparametric efficient regardless of $p\left(f_\mathbf{i}, \mathbf{X}_\mathbf{i}; \theta_0\right) \in \mathscr{P}$ and thus is globally efficient.

4.9.2 Locally Efficient Estimators

Example 2. (**Count responses)** Consider an FRM for a count response $f_\mathbf{i}$ with $E\left(f_\mathbf{i} \mid \mathbf{X}_\mathbf{i}\right) = \exp\left(\beta^\top \mathbf{X}_\mathbf{i}\right)$, where $f_\mathbf{i}$ is over-dispersed. We specify a working variance that is proportional to the conditional mean, i.e., $V\left(\mathbf{X}_\mathbf{i}; \tau^2, \beta\right) = \tau^2 \exp\left(\beta^\top \mathbf{X}_\mathbf{i}\right)$, with $\tau^2 = 1$ for non-overdispersed and $\tau^2 > 1$ for overdispersed $f_\mathbf{i}$. We then estimate τ^2 and β by iterating between (1) minimizing the squared sum of residuals $\left\{\left[f_\mathbf{i} - \exp\left(\beta^\top \mathbf{X}_\mathbf{i}\right)\right]^2 - V\left(\mathbf{X}_\mathbf{i}; \tau^2, \beta\right)\right\}^2$ for τ^2 with a given $\widehat{\beta}$ and (2) solving the UGEE for β with a given $\widehat{\tau}^2$, until convergence.

Under mild regularity conditions, $\widehat{\tau}^2 \to_p \tau_*^2$ (a constant may or may not be the truth), leading to a UGEE estimator $\widehat{\beta}^P$ with the efficient I.F.

$$\varphi_{\text{eff1}}\left(f_\mathbf{i}, \mathbf{X}_\mathbf{i}; \tau_*^2, \beta_0\right) = E^{-1}\left[\mathbf{X}_\mathbf{i} \exp\left(\beta_0^\top \mathbf{X}_\mathbf{i}\right)\mathbf{X}_\mathbf{i}^\top\right]\mathbf{X}_\mathbf{i}\left[f_\mathbf{i} - \exp\left(\beta_0^\top \mathbf{X}_\mathbf{i}\right)\right]. \tag{4.53}$$

This estimator is locally efficient; if the conditional variance is indeed proportional to the conditional mean, i.e., $\tau_*^2 = \tau_0^2$, then it is semiparametric efficient.

Alternatively, we specify a working variance from the Negative Binomial (NB) distribu-

tion with a dispersion parameter ζ, and substitute $\exp\left(\beta^\top \mathbf{X_i}\right)\left[1+\zeta\exp\left(\beta^\top \mathbf{X_i}\right)\right]$ in place of $V(\mathbf{X_i};\zeta,\beta)$, leading to an UGEE estimator $\widehat{\beta}^{NB}$ with the I.F.
$E^{-1}\left\{\mathbf{X_i}\left[1+\zeta_*\exp\left(\beta_0^\top \mathbf{X_i}\right)\right]^{-1}\exp\left(\beta_0^\top \mathbf{X_i}\right)\mathbf{X_i}^\top\right\}$
$\mathbf{X_i}\left[1+\zeta_*\exp\left(\beta_0^\top \mathbf{X_i}\right)\right]^{-1}\left[f_\mathbf{i}-\exp\left(\beta_0^\top \mathbf{X_i}\right)\right]$. Again, it has the form of the efficient I.F., but with respect to the limiting point ζ_* that may or may not be the truth. If the assumed working variance is the same as the true variance, then the resulting $\widehat{\beta}^{NB}$ is semiparametric efficient.

The distinct forms of efficient I.F.s between (4.53) and the above result from different working variance assumptions made. For count responses, other forms of non-negative working variance can be assumed, each leads to a different variance of $\widehat{\beta}$. Adaptive estimators have been shown empirically to improve efficiency for classical semiparametric GLMs for within-subject attributes in Tsiatis (2006). Our simulation studies also demonstrate this feature, some of which are discussed below.

4.9.3 Simulation Studies

To illustrate the local efficiency of adaptive estimators, we consider again overdispersed count response. The data are generated from the Negative Binomial distribution and parameters are then estimated using both parametric and semiparametric models (with different working variances). For Monte Carlo (MC) simulations, we set total MC iterations $M = 1,000$ and sample sizes $n =$ 100, 300, 500. All analyses are performed with the R software platform (R Development Core Team, 2012), with code optimized using Rcpp (Eddelbuettel et al., 2011) for run-time improvement, which is available as **Supplement**. We demonstrate between-subject attributes here, similar performances of within-subject attributes can be found in the **Supplement**.

Without loss of generality, we include one continuous predictor. By first generating $X_i \sim^{i.i.d} U(a,b)$ with $U(a,b)$ denoting a uniform distribution over (a,b), we create between-subject $X_\mathbf{i}$ with $X_\mathbf{i} = X_{i_1} + X_{i_2}$ $\left(\mathbf{i} = (i_1,i_2) \in C_2^n\right)$. Given $X_\mathbf{i}$, we generate $f_\mathbf{i} \sim NB\left(\zeta, h\left(\beta_i^\top X_\mathbf{i}\right)\right)$,

Table 4.1. Simulation results comparing MLE with UGEE for between-subject attributes.

Method	Assump.	β_0			β_1		
		$n = 100$					
		Est.	Variance		Est.	Variance	
			Asy.			Asy.	
Work-MLE	NB	2.99	0.0002		3.00	0.0001	
			Asy.	Emp.		Asy.	Emp.
UGEE	NB	3.00	0.0002	0.0002	3.00	0.0001	0.0001
	Pois	3.00	0.0007	0.0007	3.00	0.0005	0.0005
	Const.	2.99	0.006	0.006	3.00	0.0028	0.0029
		$n = 300$					
		Est.	Variance		Est.	Variance	
			Asy.			Asy.	
Work-MLE	NB	2.9970	1.8e-05		3.00	1.5e-05	
			Asy.	Emp.		Asy.	Emp.
UGEE	NB	3.00	1.8e-05	1.8e-05	3.00	1.5e-05	1.5e-05
	Pois	3.00	7.5e-05	7.4e-05	3.00	5.1e-05	5.00e-05
	Const.	3.00	0.0007	0.0007	3.00	0.0003	0.0003
		$n = 500$					
		Est.	Variance		Est.	Variance	
			Asy.			Asy.	
Work-MLE	NB	2.99	6.3e-06		3.00	5.3e-06	
			Asy.	Emp.		Asy.	Emp.
UGEE	NB	3.00	6.3e-06	6.2e-06	3.00	5.2e-06	5.1e-06
	Pois	3.00	2.7e-05	2.7e-05	3.00	1.9e-05	1.8e-05
	Const.	2.99	0.0002	0.0002	3.00	0.0001	0.0001

where $h(\beta_i^\top X_\mathbf{i}) = \exp[\beta_0 + \beta_1(X_\mathbf{i})]$ and $NB(\zeta, \mu)$ denotes a Negative Binomial with mean μ and dispersion parameter ζ. We estimate $\beta = (\beta_0, \beta_1)^\top$ using (i) MLE from Negative Binomial (NB); and (ii) semiparametric UGEE with working variances from (1) NB, (2) Poisson and (3) as a constant (See the **Supplement** for details). We set $\zeta = 10$, $\beta_0 = 3$, $\beta_1 = 3$, $a = 0$, $b = 1$ and report the parameter estimators (Est.), asymptotic (Asy.) and empirical (Emp.) variances under different sample sizes.

The MLE from NB is the benchmark for efficiency in this setting. As expected, Table 1 shows that UGEE estimators with the working variance of NB reach the local efficiency

bound, while the other two yield larger variances. As expected, the constant working variance yields the largest variance, since the Poisson working variance has a better approximation to the true variance than a constant. Thus, akin to within-subject attributes, adaptive approaches demonstrate efficiency gains for semiparametric models of between-subject attributes as well, with improvement depending on how well the working variance resembles the true variance.

4.10 Discussion

By leveraging the Hilbert-space-based semiparametric efficiency theory, we demonstrated that UGEE estimators are semiparametric efficient for functional response models (FRM) modeling between-subject attributes. Such estimators deliver the smallest asymptotic variances among a class of regular and asymptotic linear (RAL) estimators for this emerging class of semiparametric models. Specifying mathematical distributions such as normality for between-subject attributes is much more difficult than for their within-subject counterparts, as between-subject attributes are not only correlated, but generally follow more complex distributions. Extending the semiparametric efficiency theories to between-subject attributes will not only enrich the body of research on this topic, but will also greatly facilitate the applications of FRM to provide valid and efficient inferences in practice.

To show the efficiency of UGEE estimators for FRM, or model class 2, we first generalized all relevant concepts and properties of estimators to between-subject attributes, such as asymptotic linear, regular estimators, and efficiency bounds. Since directly establishing the efficiency theory is difficult for UGEE estimators, we also introduced a class of models involving only a subset of independent pairs of between-subject responses, or model class 1. Although this "conjugate" class of models has no practical utility in practice given its lower efficiency (compared to FRM, see **Supplements** for details), it provides a powerful tool to help determine the efficiency of the UGEE estimator for FRM. By connecting estimators from the two classes of models with a dual orthogonality property with respect to their respective

nuisance tangent spaces, we determined the efficiency of the UGEE estimator for FRM through the efficiency of the estimator for the "conjugate" class of models, which are much easier to address by leveraging the existing Hilbert-space-based semiparametric efficiency theory.

Therefore, not only does UGEE enjoy the semiparametric robustness, but also the efficiency in inference, just like its counterpart GEE for the classical within-subject attributes. With blooming implementations of between-subject attributes as effective summary metrics of high-dimensional data in biomedical and other research disciplines, the developed efficiency will propel growing applications of FRM.

One of the limitations is that we only focus on the efficiency bound for semiparametric FRM when applied to the cross-sectional data. We are currently working on extending the results to clustered data such as repeated assessments in longitudinal studies. A major challenge is to address the missing data arising from study dropouts and elucidate its impact on estimators through different missing data mechanisms.

Chapter 4, in part is currently being prepared for submission for publication of the material. The dissertation author was the primary investigator and author of this material. The co-authors include Lin, T., Zhang, X., Chen, T. and Tu, XM.

Chapter 5

Future Directions

Rooted in the high-dimensional data from real applications, this dissertation achieved our main objectives to develop new statistical methodologies to 1) overcome the challenges of analyzing high-dimensional data through effective dimension reductions; 2) unify a framework to fill the vital gaps in quantifying their effects by addressing the inherent correlations properly; 3) ground the implementations of the new method from a rigorous theoretical perspective.

In summary, the unified paradigm for between-subject attributes we proposed here reduces the astronomical data dimensions effectively, harmonizes robustness and efficiency in statistical modeling to inform scientific insights, and is theoretically grounded in statistical inference to accelerate blooming applications in biomedical, psychosocial, and related research.

This building block also provides a premise for extensions to longitudinal data and causal effects in the future. In this Chapter, we discuss some future directions including causal inference and longitudinal data analysis.

5.1 Doubly Robust Causal Effects for High-dimensional Outcomes

Exciting study results have revealed associations between the high-dimensional data and many diseases or health problems. However, with the current scientific frontier shifting towards discovering causal effects of the underlying biological mechanism, simply evaluating

associations is no longer a sufficient research goal. Instead, finding causal relationships provides more insights for disease prevention and intervention. For example, in a recent work, we constructed a double robust (DR) causal estimator for Mann-Whitney-Wilcoxon MWW rank sum test (MWWRST) in observational studies. The original MWWRST test is widely used to compare two treatment groups in randomized control trials (RCT) when data distributions are highly skewed, especially in the presence of outliers. As it generally yields invalid inference when applied to observational study data due to confounders. Wu et al. (2014b) introduced an approach to address confounding effects by incorporating the inverse probability weighting (IPW) technique into this rank-based statistic. In this work, we further address an important limitation in Wu et al. (2014b) by extending their approach to a doubly robust setting to provide causal inference with functional response models (FRM) by integrating the two modules to model both the mean outcome and the missing probability, where "doubly robust (DR)" means that it's robust to misspecification of either module. Additionally, if both are specified correctly, this proposed new estimator is the most efficient.

For the future, we aim to further extend such an estimator to the high-dimensional outcome, where we can leverage the pairwise distance to modify the definition of causal effects. For example, most human microbiome studies are observational due to cost, logistics, and difficulties in experimental control. Existing frameworks leading to a causal effect of exposure to certain individual taxa proportions often suffer from weak signals (Zhang et al., 2017), given the high-dimensional nature of microbiome data. To overcome the challenge of finding the causal effect for high-dimensional data, we propose to extend the definition of average causal effect (ATE) to high-dimensional outcome by deploying their pairwise distances. This building block allows for an overall characterization of the causal effect for microbiome composition. Building upon this, we could apply various observational study methods (inverse probability weighting, mean score imputation, etc.) to address confounding effects. Additionally, we will extend "doubly robust (DR)" estimators to our current context of between-subject attributes and

develop a class of weighted-UGEEs to provide DR estimators of causal effects, and this will significantly improve robustness and efficiency.

Particularly, this appealing extension will lead to a framework to identify exposure-causing changes in microbiome composition and assess the efficacy of some microbiota-related therapies, and furthermore, the development of a new class of therapies for certain diseases. Similar metrics can also be implemented in mHealth. By a similar construction of between-subject attributes as the causal effect for high-dimensional wearable data, this framework can also unbury causal effects of personalized treatments on physical activities, circadian rhythm, etc.

5.2 Triply Robust Causal Mediation Effect of High-dimensional Outcomes: Applications to Microbiome Sequence Data

Recently, additional evidence has implicated the casual mediating effect of the human microbiome from exposure or treatment to clinical outcomes. In the future, we also aim to construct a powerful and robust semiparametric model for such mediation effects. For classical within-subject attributes, the counterfactual mediation framework to define causal direct and mediation effects allow for the exposure – mediator interactions have been developed (Van der Vaart, 2000). Facing the growing need to uncover possible mediating roles of the human microbiome in relationships between exposures/treatments and clinical outcomes, we aim to extend this counterfactual mediation framework to between-subject attributes by combining two FRMs (one for the outcome and the other for the mediation) and developing weighted-UGEEs for joint inference of model parameters. Additionally, we also want to extend the "triply robust (TR)" estimator for mediation effect in (Tchetgen and Shpitser, 2012) to between-subject attributes in our context. This will help us develop a triply robust estimator for the mediation effect of microbiome composition using the Beta-diversity, where we could afford model

misspecifications from treatment, mediator, and outcome (i.e., the name "triply-robust") without sacrificing the asymptotic consistency. By focusing on the between-subject Beta-diversity, this approach will not only achieve effective dimension reduction by pooling individual weak signals but also inform an overall mediation effect of the human microbiome by integrating information across all sequenced genomes.

5.3 Distance-based Between-subject Regression for Longitudinal Data

Besides this ongoing extension in causal effects, our collaborators also strongly recommend us to further extend our approach to longitudinal study data where the sample are sequenced every 6 months, say. Given the dynamic and highly personalized nature of the human microbiome, valuable information is likely to be obtained from studies following subjects over time. To overcome the challenges in analyzing pairwise outcomes for such studies with missing data, we plan to extend the distance-based semiparametric framework to longitudinal settings, which can comprehensively capture changes in microbiome composition over time. We also address missing observations using a class of weighted-UGEE that yields valid inference under the missing at random (MAR) mechanism. In addition, we will further work on integrating a semiparametric "doubly robust (DR)" framework into our approach, to provide valid inferences under less stringent assumptions on modeling missing data in longitudinal studies.

Chapter 6

Supplemental Material

6.1 S1: Supporting Information for Chapter 2: A Semi-parametric Model for Between-Subject Attributes: Applications to Beta-diversity of Microbiome Data

6.1.1 Proof of Theorem 2.1.

Without loss of generality, consider the normalized quantity $\binom{n}{2}^{-1}\mathbf{U}_n$. A Taylor's series expansion gives

$$\sqrt{n}\left(\widehat{\theta}-\theta\right)=\left(-\frac{\partial}{\partial\theta}\mathbf{U}_n(\theta)\right)^{-\top}\sqrt{n}\mathbf{U}_n(\theta)+\mathbf{o}_p(1). \tag{6.1}$$

From the theory of multivariate U-statistics that (Kowalski and Tu, 2008a),

$$\begin{aligned}\frac{\partial}{\partial\theta}\mathbf{U}_n(\theta)&=\binom{n}{2}^{-1}\sum_{\mathbf{i}\in C_2^n}\frac{\partial}{\partial\theta}(-D_{\mathbf{i}}V_{\mathbf{i}}^{-1}h_{\mathbf{i}}(\theta))\to^p E\left(\frac{\partial}{\partial\theta}h_{\mathbf{i}}(\theta)\left(-D_{\mathbf{i}}V_{\mathbf{i}}^{-1}\right)^{\top}\right)\\&=-E\left(D_{\mathbf{i}}V_{\mathbf{i}}^{-1}D_{\mathbf{i}}^{\top}\right)=-B,\end{aligned}$$

where $\mathbf{o}_p(1)$ denotes the stochastic version of $\mathbf{o}(1)$. Since $\mathbf{U}_{n,\mathbf{i}}$ is a U-statistic-like quantity, it again follows from the theory of multivariate U-statistics that:

$$\begin{aligned}\sqrt{n}\mathbf{U}_n &= \sqrt{n}\binom{n}{2}^{-1}\sum_{\mathbf{i}\in C_2^n}\mathbf{U}_{n,\mathbf{i}} = \sqrt{n}\frac{2}{n}\sum_{i_1=1}^{n}E\left(\mathbf{U}_{n,\mathbf{i}} \mid \mathbf{y}_{i_1},\mathbf{x}_{i_1},\mathbf{z}_{i_1}\right)+\mathbf{o}_p(1) \\ &= \sqrt{n}\frac{2}{n}\sum_{i_1=1}^{n}\mathbf{v}_{i_1}+\mathbf{o}_p(1)\rightarrow_d N(\mathbf{0},\Sigma_U),\end{aligned} \tag{6.2}$$

By combining (6.11) and (6.12), we have:

$$\sqrt{n}\left(\widehat{\theta}-\theta\right) = \left(-\frac{\partial}{\partial\theta}\mathbf{U}_n\right)^{-\top}\sqrt{n}\mathbf{U}_n+\mathbf{o}_p(1) = B^{-1}\sqrt{n}\frac{2}{n}\sum_{i_1=1}^{n}\mathbf{v}_{i_1}+\mathbf{o}_p(1)\rightarrow_d N(\mathbf{0},\Sigma_\theta).$$

6.1.2 Proof of Theorem 2.2.

Again consider the normalized quantity $\binom{n}{2}^{-1}\mathbf{U}_n$. By the theory of multivariate U-statistics that (Kowalski and Tu, 2008a):

$$\frac{\partial}{\partial\theta}\mathbf{U}_n(\theta) = \begin{pmatrix} \frac{\partial}{\partial\theta_{(1)}}\mathbf{U}_{n(1)}(\theta) & \frac{\partial}{\partial\theta_{(1)}}\mathbf{U}_{n(2)}(\theta) \\ \frac{\partial}{\partial\theta_{(2)}}\mathbf{U}_{n(1)}(\theta) & \frac{\partial}{\partial\theta_{(2)}}\mathbf{U}_{n(2)}(\theta)\end{pmatrix}\rightarrow_p B = E\left(D_{\mathbf{i}}V_{\mathbf{i}}^{-1}D_{\mathbf{i}}^{\top}\right) = \begin{pmatrix} B_{11} & B_{12} \\ B_{12}^{\top} & B_{22}\end{pmatrix}. \tag{6.3}$$

It follows from a Taylor's series expansion and (6.3) that

$$\begin{aligned}\mathbf{0} &= \mathbf{U}_{n(1)}\left(\widetilde{\theta}_{(1)},\theta_{(20)}\right) = \mathbf{U}_{n(1)}(\theta)+\left(\frac{\partial^{\top}}{\partial\theta_{(1)}}\mathbf{U}_{n(1)}(\theta)\right)\left(\widetilde{\theta}_{(1)}-\theta_{(1)}\right)+\mathbf{o}_p\left(n^{-\frac{1}{2}}\right) \\ &= \mathbf{U}_{n(1)}(\theta)+B_{11}\left(\widetilde{\theta}_{(1)}-\theta_{(1)}\right)+\mathbf{o}_p\left(n^{-\frac{1}{2}}\right).\end{aligned}$$

Thus,

$$\widetilde{\theta}_{(1)}-\theta_{(1)} = -B_{11}^{-1}\mathbf{U}_{n(1)}(\theta)+\mathbf{o}_p\left(n^{-\frac{1}{2}}\right). \tag{6.4}$$

Similarly, since $B_{12}^{\top} = B_{21}$, we have:

$$\begin{aligned}
\mathbf{U}_{n(2)}\left(\widetilde{\theta}_{(1)}, \theta_{(20)}\right) &= \mathbf{U}_{n(2)}(\theta) + \left(\frac{\partial^{\top}}{\partial \theta_{(1)}}\mathbf{U}_{n(2)}\right)\left(\widetilde{\theta}_{(1)} - \theta_{(1)}\right) + \mathbf{o}_p\left(n^{-\frac{1}{2}}\right) \\
&= \mathbf{U}_{n(2)}(\theta) + B_{12}^{\top}\left(\widetilde{\theta}_{(1)} - \theta_{(1)}\right) + \mathbf{o}_p\left(n^{-\frac{1}{2}}\right) \\
&= \mathbf{U}_{n(2)}(\theta) + B_{21}\left(\widetilde{\theta}_{(1)} - \theta_{(1)}\right) + \mathbf{o}_p\left(n^{-\frac{1}{2}}\right).
\end{aligned} \tag{6.5}$$

It follows from (6.4) and (6.5) that

$$\begin{aligned}
\mathbf{U}_{n(2)}\left(\widetilde{\theta}_{(1)}, \theta_{(20)}\right) &= \mathbf{U}_{n(2)}(\theta) + B_{21}\left[-B_{11}^{-\top}\mathbf{U}_{n(1)}(\theta) + \mathbf{o}_p\left(n^{-\frac{1}{2}}\right)\right] + \mathbf{o}_p\left(n^{-\frac{1}{2}}\right) \\
&= \mathbf{U}_{n(2)}(\theta) - \left[B_{21}B_{11}^{-\top}\mathbf{U}_{n(1)}(\theta) + \mathbf{o}_p\left(n^{-\frac{1}{2}}\right)\right] + \mathbf{o}_p\left(n^{-\frac{1}{2}}\right) \\
&= \begin{pmatrix} -B_{21}B_{11}^{-1} & \mathbf{I}_q \end{pmatrix}\mathbf{U}_n(\theta) + \mathbf{o}_p\left(n^{-\frac{1}{2}}\right) \\
&= G\mathbf{U}_n(\theta) + \mathbf{o}_p\left(n^{-\frac{1}{2}}\right).
\end{aligned}$$

By the central limit theorem,

$$\sqrt{n}\mathbf{U}_{n(2)}\left(\widetilde{\theta}_{(1)}, \theta_{(20)}\right) = \sqrt{n}G\mathbf{U}_n(\theta) + \mathbf{o}_p(1) \to_d N\left(\mathbf{0}, \Sigma_{(2)} = G\Sigma_U G^{\top}\right). \tag{6.6}$$

The asymptotic normality of $\mathbf{U}_{n(2)}\left(\widetilde{\theta}_{(1)}, \theta_{(20)}\right)$ implies that the score statistic $S_n\left(\widetilde{\theta}_{(1)}, \theta_{(20)}\right)$ has the asymptotic χ_q^2 distribution.

6.1.3 PERMANOVA

If $\mathbf{x}_i$ consists of only one categorical variable for groups, PERMANOVA can be used to compare Beta-diversity across different groups. Consider a total of K groups for this categorical variable, PERMANOVA uses the pseudo-F statistic for inference about overall group differences

in Beta-diversity:

$$\text{pseudo-}F = \frac{tr\left(HGH\right)/\left(p-1\right)}{tr[(\mathbf{I}_n - H)G(\mathbf{I}_n - H)]/(n-p)}, \tag{6.7}$$
$$G = \left(\mathbf{I}_n - \frac{\mathbf{1}_n\mathbf{1}_n^\top}{n}\right) A \left(\mathbf{I}_n - \frac{\mathbf{1}_n\mathbf{1}_n^\top}{n}\right), \quad A = \left(-\frac{1}{2}d_{\mathbf{i}}^2\right),$$

where $tr\left(\cdot\right)$ denotes the trace of a matrix, X is the design matrix that contains the group information, p is the length of $\mathbf{x}_i$, $H = X(X^\top X)^{-1}X^\top$ is the projection of the design matrix X, G is the Gower's centered matrix obtained from the distance matrix $D = (d_{\mathbf{i}})$, $\mathbf{1}_n$ denotes a $n \times 1$ column vector of 1's, and $\mathbf{I}_n$ denotes the $n \times n$ identity matrix. For example, if $K = 2$, and $x_i = 1$ if the ith subject is from diseased group and $x_i = 0$ otherwise, then $X = \left(\mathbf{1}_n, \mathbf{x}^\top\right)$, where $\mathbf{x}^\top = (x_1, x_2, \ldots, x_n)^\top$.

6.1.4 Details of Data Generating Procedure with eCDF and Copula

For notational clarity, we use upper-case to denote random variables and lower-case to denote their values. Consider a random variable X and let $F\left(x\right)$ denote the cumulative distribution function (CDF) of X. Then the probability integral transformation of X, $U = F\left(X\right)$, follows $U\left(0,1\right)$, where $U\left(0,1\right)$ is a uniform between 0 and 1 (Kowalski and Tu, 2008a). Thus, if $F\left(x\right)$ is known, we can simulate X from $X = F^{-1}\left(U\right)$, where $F^{-1}\left(u\right)$ is the inverse of $F\left(x\right)$ defined by $F^{-1}\left(u\right) = \inf\left\{x \mid F\left(x\right) \geq u\right\}$, $0 < u < 1$. If $F\left(x\right)$ is unknown, we can instead use the empirical CDF (eCDF) of the observed X, i.e., $F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I\left(X_i \leq x\right)$, where $I\left(A\right)$ is an indicator with value 1 if A is true and 0 otherwise.

For a $p \times 1$ random vector $\mathbf{X} = \left(X_1, X_2, ..., X_p\right)^\top$ such as OTU counts, let $F\left(\mathbf{x}\right) = F\left(x_1, x_2, ..., x_p\right)$ denote the CDF. It can be expressed in terms of uniformly distributed marginals $F_j\left(X_j\right)$ and a copula, defined as the joint CDF of a $p \times 1$ random vector $\mathbf{U} = \left(U_1, U_2, ..., U_p\right)^\top$ with uniform marginals $U_j = F_j\left(X_j\right)$ $\left(1 \leq j \leq p\right)$ (Sklar, 1959). Similar to the univariate case, we can simulate correlated multivariate random vectors $\mathbf{X} = \left(X_1, X_2, ..., X_p\right)^\top$ where

$X_j = {F_j}^{-1}\left(U_j\right)$, with specified marginals $F_j\left(X_j\right)$ through copula.

To simulate $\mathbf{X}$ with distributions similar to those OTUs from a real study, we first use the copula to create a correlated multivariate uniform $\mathbf{U}_n$ based on the eCDF $F_n(\mathbf{x})$ of the observed OTUs, with the uniform marginals $U_{nj} = F_{nj}\left(X_j\right) = \frac{1}{n}\sum_{i=1}^{n} I\left(X_{ij} \leq x_j\right)$. Then by smoothing U_{nj} (de repartition an dimensions et leurs marges, gen), we apply the copula again to create a multivariate normal $\mathbf{V}$ with correlations similar to those of the original OTUs. Afterward, by simulating from $\mathbf{V}$, we obtain correlated multivariate uniform $\mathbf{U}$ with correlations and marginals similar to those of $\mathbf{U}_n$. Finally, by smoothing $F_{nj}\left(X_j\right)$ and inverting the simulated U_j to X_j with $X_j = {F_j}^{-1}\left(U_j\right)$, where $F_j(\cdot)$ is a smoothed version of $F_{nj}(\cdot)$**,** we obtain the simulated OTUs $\mathbf{X} = (X_1, X_2, ..., X_p)^\top$ with a distribution similar to $F_n(\mathbf{x})$ of the real OTUs. Beta-diversity was then calculated from simulated OTU counts after appropriate normalization.

As this procedure does not involve analytical distributional models, population-level characteristics such as mean and standard deviation are estimated by Monte Carlo (MC) simulation with a large MC size of 5,000.

6.1.5 Details of Simulation for Group Comparison Accounting for Covariates

We simulate the two covariates from parametric distributions with $x_i^g \sim \text{Bern}(p)$ and $z_i^a \sim U(a,b)$ and then created their respective pairwise counterparts $x_{\mathbf{i}}^g$ and $z_{\mathbf{i}}^a$, where $\text{Bern}(p)$ denotes Bernoulli with mean p and $U(a,b)$ a uniform over (a,b). We set:

$$p = 0.45, \quad a = 0, \quad b = 1$$
$$\theta_0 = \left(\beta_0, \beta_{22}^d, \beta_{12}^d, \beta_{22}^g, \beta_{12}^g, \xi^a\right)^\top = (-0.4595, 0, 0, 0.5, 0.5, 0.5)^\top.$$

To simulate $f(\mathbf{y_i})$ for the regression with covariates, we first simulate Beta-diversity distance $d_{\mathbf{i}}(\mathbf{y_i})$ and then use the two covariates $x_{\mathbf{i}}^g$ and $z_{\mathbf{i}}^a$ to create the mean $h(\mathbf{x_i}, z_{\mathbf{i}}; \theta_0) = \exp\left(\mathbf{u_i}^\top \theta_0\right)$. We next center $d_{\mathbf{i}}(\mathbf{y_i})$ with the true value of $\beta_0\,(= -0.4595)$ to create a "residual"

$\varepsilon_{\mathbf{i}} = d_{\mathbf{i}}(\mathbf{y_i}) - \beta_0$, which is then added to $u_{\mathbf{i}}^{\top}\theta_0$ and expenentiated to create:

$$\widetilde{d}_{\mathbf{i}}(\mathbf{y_i}) = \exp\left(\mathbf{u}_{\mathbf{i}}^{\top}\theta_0 + \varepsilon_{\mathbf{i}}\right) = \exp\left(\mathbf{u}_{\mathbf{i}}^{\top}\theta_0\right)\exp(\varepsilon_{\mathbf{i}}).$$

By setting $C_0 = E(\exp(\varepsilon_{\mathbf{i}}))$, we obtain simulated $f(\mathbf{y_i}) = C_0^{-1}\widetilde{d}_{\mathbf{i}}(\mathbf{y_i})$. This ensures that $E[f(\mathbf{y_i}) \mid \mathbf{x_i}, z_{\mathbf{i}}] = h(\mathbf{x_i}, z_{\mathbf{i}}; \theta) = \exp\left(\mathbf{u}_{\mathbf{i}}^{\top}\theta\right)$.

We estimate C_0 by the sample mean $C_0 = \binom{n}{2}^{-1}\sum_{\mathbf{i}\in C_2^n}\exp(\varepsilon_{\mathbf{i}})$ using a large $n = 5{,}000$, where $C_0 = 1.000796$ in our setting.

6.1.6 Details to Obtain Parameter Estimates from UGEE

The method to find $\widehat{\theta}$ is through Newton-Raphson using the pseudo-score $\mathbf{U}_n(\theta)$. For example, in a model with

$$E[f_{\mathbf{i}} \mid x_{\mathbf{i}}] = h_{\mathbf{i}}(\mathbf{x_i}; \theta) = \exp\left\{\theta^{\top} g(x_{\mathbf{i}})\right\},\ \mathbf{i} = (i_1, i_2) \in C_2^n, \tag{6.8}$$

where $x_{\mathbf{i}} = \{x_{i_1}, x_{i_2}\}$, $g(\cdot)$ is some symmetric smooth function such as the Euclidean distance. Let

$$S_{\mathbf{i}} = f_{\mathbf{i}} - h_{\mathbf{i}}(\mathbf{x_i}; \theta),\ D_{\mathbf{i}} = \frac{\partial}{\partial\theta}h_{\mathbf{i}}(\mathbf{x_i}; \theta),\ V_{\mathbf{i}} = Var(f_{\mathbf{i}} \mid \mathbf{x_i}, \mathbf{z_i}) = \exp\left\{\theta^{\top} g(x_{\mathbf{i}})\right\},$$

with

$$\mathbf{U}_n(\theta) = \sum_{\mathbf{i}\in C_2^n}\mathbf{U}_{n,\mathbf{i}} = \sum_{\mathbf{i}\in C_2^n} D_{\mathbf{i}} V_{\mathbf{i}}^{-1} S_{\mathbf{i}} = \mathbf{0}, \tag{6.9}$$

we can obtain $\widehat{\theta}$ by iterating through

$$\begin{aligned}\theta^{(t+1)} - \theta^{(t)} &= \sum_{\mathbf{i}\in C_2^n}\left(D_{\mathbf{i}} V_{\mathbf{i}}^{-1} D_{\mathbf{i}}\right)^{-1} D_{\mathbf{i}} V_{\mathbf{i}}^{-1} S_{\mathbf{i}} \\ &= \sum_{\mathbf{i}\in C_2^n}\left(D_{\mathbf{i}} V_{\mathbf{i}}^{-1} D_{\mathbf{i}}\right)^{-1} D_{\mathbf{i}} V_{\mathbf{i}}^{-1}\left\{f_{\mathbf{i}} - h_{\mathbf{i}}(\mathbf{x_i}; \theta^{(t)})\right\}\end{aligned} \tag{6.10}$$

until convergence, where all relevant quantities of $(D_{\mathbf{i}}, V_{\mathbf{i}})$ are evaluated at the t^{th} step with $\theta^{(t)}$.

6.1.7 FDR-corrected Test Results for the Real Data Analyses

We applied the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995) to control the family-wise FDR at 5%, and provided comparisons of p-values before and after FDR-correction for the real data analyses.

Shown in top panel of the table Table 6.1 are estimates (Est.) of θ, Wald and score test p-values (Wald under "W.p", score under "S.p", Bootstrap Wald under "B.W.p" and Bootstrap score under "B.S.p") for testing the nulls of no difference for the diagnostic groups and no effect for the two covariates**.** The bottom panel includes Wald and score test p-values for the three major types of hypotheses and covariate effects**.**

The comparisons indicate that major conclusions in the real data application remain unchanged after FDR-corrections, except for comparing the between-group variability of AUD-HC pairs vs. the within-group variability of AH-AH pairs with $\beta_{23}^{d} = 0$, where the score test p-value (S.p) was .020 before and .060 after correction.

6.1.8 Simulation Details of Power Comparison with the Existing Approach.

To control for the effect size that allows for appropriate power comparisons in the simulation, the data were generated from the alternative using the Dirichlet-Multinomial distribution (DM) with parameters calibrated from the real data using R package 'dirmult' (Tvedebrink, 2010), with effect size estimated with $\frac{\widehat{\theta}-0}{\sqrt{n}se(\widehat{\theta})}$ as a rough quantification. This allows us to vary effect sizes more easily for the power comparison and continues to generate Beta-diversity outcomes with their distributions resembling the real data as shown in the Figure S2 below.

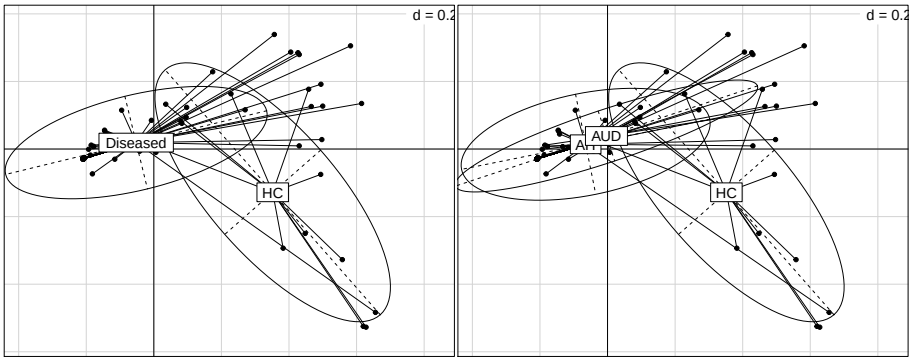

Figure 6.1. Principal Coordinates Analysis (PCoA) plots of Beta-diverisity distance for (1) combined diseased (AH and AUD patients) group and non-alcoholic controls (HC) (left) and (2)alcoholic hepatitis (AH) patients, alcohol user disorder (AUD) patients and non-alcoholic controls (HC) (right)

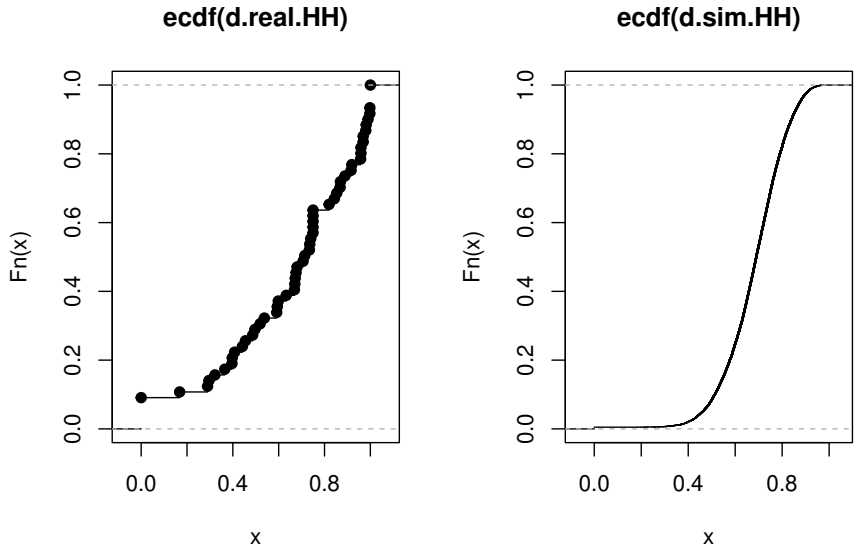

Figure 6.2. Empirical CDFs of Real vs. Simulated Beta-diversity.

6.2 S2: Supporting Information for Chapter 3: A Distance-based Semiparametric Regression Framework for Between-subject attributes

6.2.1 Details for Considering Pairs with Discordant Binary Responses

For each ith subject, $y_i = 1$ if diseased, $y_i = 0$ otherwise. Then there will be 4 pairwise possibilities: $y_i = y_j = 0$, $y_i = y_j = 1$, $y_i = 0, y_j = 1$ and $y_i = 1, y_j = 0$.

In logistic regression, we have:

$$\Pr(y_i = 1 \mid x_i) = \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)}, \Pr(y_i = 0 \mid x_i) = \frac{1}{1 + \exp(\beta_0 + \beta_1 x_i)}$$
$$\Pr(y_i = 1 \mid x_i) = \frac{\exp(\beta_0 + \beta_1 x_j)}{1 + \exp(\beta_0 + \beta_1 x_j)}, \Pr(y_j = 0 \mid x_j) = \frac{1}{1 + \exp(\beta_0 + \beta_1 x_j)}$$

Thus,

$$\begin{aligned}\Pr(y_i = 1, y_j = 1 \mid x_i, x_j) &= \Pr(y_i = 1 \mid x_i) \cdot \Pr(y_j = 1 \mid x_j) \text{ by independence} \\ &= \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)} \cdot \frac{\exp(\beta_0 + \beta_1 x_j)}{1 + \exp(\beta_0 + \beta_1 x_j)} \\ &= \frac{\exp[2\beta_0 + \beta_1 (x_i + x_j)]}{[1 + \exp(\beta_0 + \beta_1 x_i)] \cdot [1 + \exp(\beta_0 + \beta_1 x_j)]}\end{aligned}$$

Similarly,

$$\Pr(y_i = 0, y_j = 0 \mid x_i, x_j) = \frac{1}{[1 + \exp(\beta_0 + \beta_1 x_i)] \cdot [1 + \exp(\beta_0 + \beta_1 x_j)]},$$
$$\Pr(y_i = 1, y_j = 0 \mid x_i, x_j) = \frac{\exp(\beta_0 + \beta_1 x_i)}{[1 + \exp(\beta_0 + \beta_1 x_i)] \cdot [1 + \exp(\beta_0 + \beta_1 x_j)]},$$
$$\Pr(y_i = 0, y_j = 1 \mid x_i, x_j) = \frac{\exp(\beta_0 + \beta_1 x_j)}{[1 + \exp(\beta_0 + \beta_1 x_i)] \cdot [1 + \exp(\beta_0 + \beta_1 x_j)]}.$$

If we choose $y_i = 0, y_j = 0$ as reference, then,

$$\frac{\Pr(y_i = 1, y_j = 1 \mid x_i, x_j)}{\Pr(y_i = 0, y_j = 0 \mid x_i, x_j)} = \exp[2\beta_0 + \beta_1(x_i + x_j)],$$
$$\frac{\Pr(y_i = 1, y_j = 0 \mid x_i, x_j)}{\Pr(y_i = 0, y_j = 0 \mid x_i, x_j)} = \exp[\beta_0 + \beta_1 x_i],$$
$$\frac{\Pr(y_i = 0, y_j = 1 \mid x_i, x_j)}{\Pr(y_i = 0, y_j = 0 \mid x_i, x_j)} = \exp[\beta_0 + \beta_1 x_j], (i, j) \in C_2^n,$$

All of those relative probabilities are not related to the differences of x_i, x_j.

However, if we choose $y_i = 0, y_j = 1$ as reference, then

$$\frac{\Pr(y_i = 1, y_j = 0 \mid x_i, x_j)}{\Pr(y_i = 0, y_j = 1 \mid x_i, x_j)} = \exp[\beta_1(x_i - x_j)], (i, j) \in C_2^n.$$

i.e. if we define an indicator $\psi(i, j) = 1$ if $y_i = 1, y_j = 0$, $\psi(i, j) = 0$ if $y_i = 0, y_j = 1$, then this amounts to modeling with logistic regression

$$\frac{\Pr\left\{\psi(i, j) = 1 \mid x_i, x_j\right\}}{1 - \Pr\left\{\psi(i, j) = 1 \mid x_i, x_j\right\}} = \exp\left\{\beta_1(x_i - x_j)\right\},$$
$$\text{i.e. } \log it[\Pr\left\{\psi(i, j) = 1 \mid x_i, x_j\right\}] = \beta_1(x_i - x_j), \quad (i, j) \in C_2^n.$$

Therefore, only the pairs with discordant responses are associated with differences in the explanatory variables.

6.2.2 Details to Obtain Parameter Estimates from UGEE

The method to find $\widehat{\theta}$ is through Newton-Raphson using the pseudo-score $\widetilde{U}_n(\theta)$. For example, in a model with

$$E(f_{\mathbf{i}} \mid x_{\mathbf{i}}) = h_{\mathbf{i}}(x_{\mathbf{i}}; \theta) = \theta^{\top} g(x_{\mathbf{i}}),\ \mathbf{i} = (i_1, i_2) \in C_2^n,$$

where $x_{\mathbf{i}} = \{x_{i_1}, x_{i_2}\}$, $g(\cdot)$ is some symmetric smooth function such as the Euclidean distance where $g(x_{\mathbf{i}}) = \left[(x_{i_1} - x_{i_2})^2\right]^{1/2}$. Let

$$S_{\mathbf{i}} = f_{\mathbf{i}} - h_{\mathbf{i}}(x_{\mathbf{i}};\theta),\ D_{\mathbf{i}} = \frac{\partial}{\partial\theta} h_{\mathbf{i}}(x_{\mathbf{i}};\theta),\ V_{\mathbf{i}} = Var\left(f_{\mathbf{i}} \mid x_{\mathbf{i}}, z_{\mathbf{i}}\right) = \sigma^2 \text{ (a constant)},$$

with

$$U_n(\theta) = \sum_{\mathbf{i}\in C_2^n} U_{n,\mathbf{i}} = \sum_{\mathbf{i}\in C_2^n} D_{\mathbf{i}} V_{\mathbf{i}}^{-1} S_{\mathbf{i}} = 0,$$

we can obtain $\widehat{\theta}$ by iterating through

$$\begin{aligned}\theta^{(t+1)} - \theta^{(t)} &= \sum_{\mathbf{i}\in C_2^n} \left(D_{\mathbf{i}} V_{\mathbf{i}}^{-1} D_{\mathbf{i}}\right)^{-1} D_{\mathbf{i}} V_{\mathbf{i}}^{-1} \mathbf{S}_{\mathbf{i}} \\ &= \sum_{\mathbf{i}\in C_2^n} \left(D_{\mathbf{i}} V_{\mathbf{i}}^{-1} D_{\mathbf{i}}\right)^{-1} D_{\mathbf{i}} V_{\mathbf{i}}^{-1} \left\{ f_{\mathbf{i}} - h_{\mathbf{i}}(x_{\mathbf{i}};\theta^{(t)}) \right\}\end{aligned}$$

until convergence, where all relevant quantities of $(D_{\mathbf{i}}, V_{\mathbf{i}})$ are evaluated at the t^{th} step with $\theta^{(t)}$.

6.2.3 Proof of Theorem 3.1

Without loss of generality, consider the normalized quantity $\binom{n}{2}^{-1}\widetilde{U}_n$.

$$\begin{aligned}\binom{n}{2}^{-1}\widetilde{U}_n &= \binom{n}{2}^{-1} \sum_{\mathbf{i}\in C_2^n} \widetilde{U}_{n,\mathbf{i}}(y_{i_1}, y_{i_2}) \\ &= \binom{n}{2}^{-1} \sum_{\mathbf{i}\in C_2^n} \frac{1}{2}\left\{ U_{n,\mathbf{i}}(y_{i_1}, y_{i_2}) + U_{n,\widetilde{\mathbf{i}}}(y_{i_2}, y_{i_1}) \right\}\end{aligned}$$

Let

$$\begin{aligned}\delta_m &= \left(sign(i_1 - i_2)\delta_{12}^m(w_{\mathbf{i}}), ..., sign(i_1 - i_2)\delta_{(K_m-1)K_m}^m(w_{\mathbf{i}})\right)^\top,\ \delta = \left(\delta_1^\top, \ldots, \delta_q^\top\right)^\top, \\ r_{\mathbf{i}}(i_1, i_2) &= \left(sign(i_1 - i_2) d(x_{i_1}, x_{i_2}), (z_{i_1} - z_{i_2})^\top, \delta^\top\right)^\top = -r_{\mathbf{i}}(i_2, i_1).\end{aligned}$$

Then $h(x_{\mathbf{i}}, z_{\mathbf{i}}, w_{\mathbf{i}}; \theta) = \theta^{\top} r_{\mathbf{i}}$ for FRM with continuous or count responses,

$$
\begin{aligned}
U_{n,\mathbf{i}}(y_{i_1}, y_{i_2}) &= D_{\mathbf{i}}(y_{i_1}, y_{i_2}) V_{\mathbf{i}}^{-1}(y_{i_1}, y_{i_2}) S_{\mathbf{i}}(y_{i_1}, y_{i_2}) \\
&= D_{\mathbf{i}} V_{\mathbf{i}}^{-1} \{ f_{\mathbf{i}}(y_{i_1}, y_{i_2}) - h_{i_1 i_2}(x_{\mathbf{i}}, z_{\mathbf{i}}, w_{\mathbf{i}}; \theta) \} \\
&= r_{\mathbf{i}}(i_1, i_2) V_{\mathbf{i}}^{-1} \left\{ (y_{i_1} - y_{i_2}) - \theta^{\top} r_{\mathbf{i}}(i_1, i_2) \right\} \\
U_{n,\widetilde{\mathbf{i}}}(y_{i_2}, y_{i_1}) &= D_{\mathbf{i}}(y_{i_2}, y_{i_1}) V_{\mathbf{i}}^{-1}(y_{i_2}, y_{i_1}) S_{\mathbf{i}}(y_{i_2}, y_{i_1}) \\
&= r_{\mathbf{i}}(i_2, i_1) d(x_{i_2}, x_{i_1}) V_{\mathbf{i}}^{-1} \left\{ (y_{i_2} - y_{i_1}) - \theta^{\top} r_{\mathbf{i}}(i_2, i_1) \right\} \\
&= (-1) r_{\mathbf{i}}(i_1, i_2) V_{\mathbf{i}}^{-1} (-1) \left\{ (y_{i_1} - y_{i_2}) - \theta^{\top} r_{\mathbf{i}}(i_1, i_2) \right\} \\
&= (-1)(-1) U_{n,\mathbf{i}}(y_{i_1}, y_{i_2}) = U_{n,\mathbf{i}}(y_{i_1}, y_{i_2})
\end{aligned}
$$

For FRM with binary responses, $logit\{h(x_{\mathbf{i}}, z_{\mathbf{i}}, w_{\mathbf{i}}; \theta)\} = \theta^{\top} r_{\mathbf{i}}$.

$$
\begin{aligned}
h_{i_2 i_1}(x_{\mathbf{i}}, z_{\mathbf{i}}, w_{\mathbf{i}}; \theta) &=^{def} \pi_{i_2 i_1} = \frac{\exp(\theta^{\top} r_{\mathbf{i}}(i_2, i_1))}{1 + \exp(\theta^{\top} r_{\mathbf{i}}(i_2, i_1))} = \frac{\exp(-\theta^{\top} r_{\mathbf{i}}(i_1, i_2))}{1 + \exp(-\theta^{\top} r_{\mathbf{i}}(i_1, i_2))} \\
&= \frac{1}{1 + \exp(\theta^{\top} r_{\mathbf{i}}(i_1, i_2))} = 1 - \pi_{i_1 i_2} = 1 - h_{i_1 i_2}(x_{\mathbf{i}}, z_{\mathbf{i}}, w_{\mathbf{i}}; \theta)
\end{aligned}
$$

$$
\begin{aligned}
U_{n,\mathbf{i}}(y_{i_1},y_{i_2}) &= D_{\mathbf{i}}(y_{i_1},y_{i_2})\,V_{\mathbf{i}}^{-1}(y_{i_1},y_{i_2})\,S_{\mathbf{i}}(y_{i_1},y_{i_2}) \\
&= D_{\mathbf{i}}V_{\mathbf{i}}^{-1}\left\{f_{\mathbf{i}}(y_{i_1},y_{i_2}) - h_{i_1i_2}(x_{\mathbf{i}},z_{\mathbf{i}},w_{\mathbf{i}};\theta)\right\} \\
&= r_{\mathbf{i}}(i_1,i_2)d(\mathbf{x}_{i_1},\mathbf{x}_{i_2})\left\{I(y_{i_1}=1,y_{i_2}=0) - \pi_{i_1i_2}\right\} \\
U_{n,\widetilde{\mathbf{i}}}(y_{i_2},y_{i_1}) &= D_{\mathbf{i}}(y_{i_2},y_{i_1})\,V_{\mathbf{i}}^{-1}(y_{i_2},y_{i_1})\,S_{\mathbf{i}}(y_{i_2},y_{i_1}) \\
&= r_{\mathbf{i}}(i_2,i_1)d(\mathbf{x}_{i_2},\mathbf{x}_{i_1})\left\{I(y_{i_2}=1,y_{i_1}=0) - \pi_{i_2i_1}\right\} \\
&= (-1)r_{\mathbf{i}}(i_1,i_2)d(\mathbf{x}_{i_2},\mathbf{x}_{i_1})\left[\left\{1-I(y_{i_1}=1,y_{i_2}=0)\right\} - (1-\pi_{i_1i_2})\right] \\
&= (-1)r_{\mathbf{i}}(i_1,i_2)d(\mathbf{x}_{i_2},\mathbf{x}_{i_1})(-1)\left\{I(y_{i_1}=1,y_{i_2}=0) - \pi_{i_1i_2}\right\} \\
&= U_{n,\mathbf{i}}(y_{i_1},y_{i_2}) \\
\widetilde{U}_{n,\mathbf{i}}(y_{i_1},y_{i_2}) &= \frac{1}{2}\left\{U_{n,\mathbf{i}}(y_{i_1},y_{i_2}) + U_{n,\widetilde{\mathbf{i}}}(y_{i_2},y_{i_1})\right\} = U_{n,\mathbf{i}}(y_{i_1},y_{i_2}) = U_{n,\widetilde{\mathbf{i}}}(y_{i_2},y_{i_1})
\end{aligned}
$$

Thus, for both FRMs,

$$
\begin{aligned}
\widetilde{U}_{n,\mathbf{i}}(y_{i_1},y_{i_2}) &= \frac{1}{2}\left\{U_{n,\mathbf{i}}(y_{i_1},y_{i_2}) + U_{n,\widetilde{\mathbf{i}}}(y_{i_2},y_{i_1})\right\} \\
&= U_{n,\mathbf{i}}(y_{i_1},y_{i_2}),
\end{aligned}
$$

since the beta-diversity matrix is symmetric, i.e. $d(x_{i_1},x_{i_2}) = d(x_{i_2},x_{i_1})$. Taylor expansion gives

$$
\sqrt{n}\left(\widehat{\theta}-\theta\right) = -\left\{\frac{\partial}{\partial\theta}\widetilde{U}_n(\theta,\varphi)\right\}^{-\top}\cdot\left[\sqrt{n}\widetilde{U}_n(\theta,\varphi) + \left\{\frac{\partial}{\partial\varphi}\widetilde{U}_n(\theta,\varphi)\right\}^{\top}\cdot\sqrt{n}(\widehat{\varphi}-\varphi)\right] + o_p(1),
$$

By the assumption, $\left\{\frac{\partial}{\partial\varphi}\widetilde{U}_n(\theta,\varphi)\right\}^{\top}\cdot\sqrt{n}(\widehat{\varphi}-\varphi) = o_p(1)$. And

$$
\begin{aligned}
\frac{\partial}{\partial\theta}\widetilde{U}_n(\theta,\varphi) &= \binom{n}{2}^{-1}\sum_{\mathbf{i}\in C_2^n}\frac{\partial}{\partial\theta}\left\{-D_{\mathbf{i}}V_{\mathbf{i}}^{-1}h_{\mathbf{i}}(\theta)\right\} \to^p E\frac{\partial}{\partial\theta}h_{\mathbf{i}}(\theta)(-D_{\mathbf{i}}V_{\mathbf{i}}^{-1})^{-\top} \\
&= -E(D_{\mathbf{i}}V_{\mathbf{i}}^{-1}D_{\mathbf{i}}^{\top}) =^{def} -B^{\top},
\end{aligned}
$$

Since $\widetilde{U}_{n,\mathbf{i}}$ is a U-statistic-like quantity, it follows from the theory of multivariate U-

statistics (Kowalski and Tu, 2008a) that:

$$\begin{aligned}\sqrt{n}\widetilde{U}_n &= \sqrt{n}\binom{n}{2}^{-1}\sum_{\mathbf{i}\in C_2^n}\widetilde{U}_{n,\mathbf{i}} = \sqrt{n}\frac{2}{n}\sum_{i_1=1}^{n}E\left(\widetilde{U}_{n,\mathbf{i}} \mid y_{i_1},x_{i_1},z_{i_1},w_{i_1}\right)+o_p(1)\\ &= \sqrt{n}\frac{2}{n}\sum_{i_1=1}^{n}v_{i_1}+o_p(1)\rightarrow_d N(0,\Sigma_U),\\ \Sigma_U &= 4var(v_{i_1})\end{aligned}$$

where $o_p(1)$ denotes the stochastic version of $o(1)$. It follows that

$$\sqrt{n}\left(\widehat{\theta}-\theta\right) = \left(-\frac{\partial}{\partial\theta}\widetilde{U}_n\right)^{-\top}\sqrt{n}\widetilde{U}_n+o_p(1) = B^{-1}\sqrt{n}\frac{2}{n}\sum_{i_1=1}^{n}v_{i_1}+o_p(1)\rightarrow_d N(0,\Sigma_\theta),$$

where $\Sigma_\theta = B^{-1}\Sigma_U B^{-1}$.

6.3 S3: Supporting Information for Chapter 4: On Semi-parametric Efficiency of an Emerging Class of Distance-based Regression Models for Between-subject Attributes

6.3.1 Proofs of Theorems

1. **Proof. of Theorem 4.1**

Without loss of generality, consider the normalized quantity $\binom{n}{2}^{-1}\mathbf{U}_n$. A Taylor's series expansion gives

$$\sqrt{n}\left(\widehat{\theta}-\theta\right) = \left(-\frac{\partial}{\partial\theta}\mathbf{U}_n(\theta)\right)^{-\top}\sqrt{n}\mathbf{U}_n(\theta)+\mathbf{o}_p(1). \tag{6.11}$$

From the theory of multivariate U-statistics that (**Kowalski and Tu, 2007**),

$$\begin{aligned}\frac{\partial}{\partial\theta}\mathbf{U}_n(\theta) &= \binom{n}{2}^{-1}\sum_{\mathbf{i}\in C_2^n}\frac{\partial}{\partial\theta}(-D_{\mathbf{i}}V_{\mathbf{i}}^{-1}h_{\mathbf{i}}(\theta)) \to^p E\left(\frac{\partial}{\partial\theta}h_{\mathbf{i}}(\theta)\left(-D_{\mathbf{i}}V_{\mathbf{i}}^{-1}\right)^\top\right)\\ &= -E\left(D_{\mathbf{i}}V_{\mathbf{i}}^{-1}D_{\mathbf{i}}^\top\right) = -B,\end{aligned}$$

where $\mathbf{o}_p(1)$ denotes the stochastic version of $\mathbf{o}(1)$. Since $\mathbf{U}_{n,\mathbf{i}}$ is a U-statistic-like quantity, it again follows from the theory of multivariate U-statistics that (**Kowalski and Tu, 2007**):

$$\begin{aligned}\sqrt{n}\mathbf{U}_n &= \sqrt{n}\binom{n}{2}^{-1}\sum_{\mathbf{i}\in C_2^n}\mathbf{U}_{n,\mathbf{i}} = \sqrt{n}\frac{2}{n}\sum_{i_1=1}^{n}E\left(\mathbf{U}_{n,\mathbf{i}}\mid\mathbf{y}_{i_1},\mathbf{x}_{i_1}\right)+\mathbf{o}_p(1) \qquad (6.12)\\ &= \sqrt{n}\frac{2}{n}\sum_{i_1=1}^{n}\mathbf{v}_{i_1}+\mathbf{o}_p(1)\to_d N(\mathbf{0},\Sigma_U),\end{aligned}$$

By combining (**6.11**) and (**6.12**), we have:

$$\sqrt{n}\left(\widehat{\theta}-\theta\right) = \left(-\frac{\partial}{\partial\theta}\mathbf{U}_n\right)^{-\top}\sqrt{n}\mathbf{U}_n+\mathbf{o}_p(1) = B^{-1}\sqrt{n}\frac{2}{n}\sum_{i_1=1}^{n}\mathbf{v}_{i_1}+\mathbf{o}_p(1)\to_d N(\mathbf{0},\Sigma_\theta).$$

2. **Proof. of Theorem 4.3**

If φ and ψ are two such I.F.s, then it's easy to show that

$$E(|\varphi-\psi|)=0.$$

Thus, $\varphi=\psi$ a.s..

3. **Proof** (sketch) **of Theorem 4.4** (**Key results**).

We first give the definition of two sequences of probability measures that are contiguous: Let V_n be a sequence of random vectors and let P_{1n} and P_{0n} be sequences of probability measures with densities $p_{1n}(v_n)$ and $p_{0n}(v_n)$, respectively. The sequence of probability measures P_{1n} is contiguous to the sequence of probability measures P_{0n} if, for any sequence of events A_n defined

with respect to V_n, $P_{0n}(A_n) \to 0$ as $n \to \infty$ implies that $P_{1n}(A_n) \to 0$ as $n \to \infty$.

Consider the sequence of densities $p_{0n}(v_n) = \Pi_{\mathbf{i} \in C_2^n} p(z_{\mathbf{i}n}, \theta_0)$ and the LDGP $p_{1n}(v_n) = \Pi_{\mathbf{i} \in C_2^n} p(z_{\mathbf{i}n}, \theta_n)$, where $n^{1/2}(\theta_n - \theta_0) \to \tau$. The if we have P_{1n} and P_{0n} contiguous, we can show that $o_{P_{0n}}(1) = o_{P_{1n}}(1)$. Making use of LeCam's concept of contiguity, **Hajek (1962)** proved the asymptotic normality of such $S_\theta(f_\mathbf{i}; \theta)$ under the contiguous alternatives.

Then, by AL we have

$$n^{1/2}\left(\widehat{\beta}_n - \beta(\theta_0)\right) = \sqrt{n}\binom{n}{2}^{-1} \sum_{\mathbf{i} \in C_2^n} \varphi(f_\mathbf{i}) + \mathbf{o}_{P_{0n}}(1),$$

and by the theory of U-statistics,

$$n^{1/2}\left(\widehat{\beta}_n - \beta(\theta_n)\right) \to_d N(\mathbf{0}, 4\Sigma),$$

where

$$\Sigma = Var\left(E\left(\varphi(f_\mathbf{i}) \mid Z_i\right)\right) = E\left[E\left(\varphi(f_\mathbf{i}) \mid Z_i\right) E\left(\varphi^\top(f_\mathbf{i}) \mid Z_i\right)\right].$$

Then by adding and subtracting common terms, we obtain:

$$\begin{aligned} n^{1/2}\left(\widehat{\beta}_n - \beta(\theta_n)\right) = \sqrt{n}\binom{n}{2}^{-1} & \sum_{\mathbf{i} \in C_2^n} \left[\varphi(f_\mathbf{i}) - E_{\theta_n}\{\varphi(f_\mathbf{i})\}\right] \\ & + n^{1/2} E_{\theta_n}\{\varphi(f_\mathbf{i})\} - n^{1/2}\{\beta(\theta_n) - \beta(\theta_0)\} + o_{P_{1n}}(1), \end{aligned}$$

and also by the theory of U-statistics,

$$\begin{aligned} \sqrt{n}\binom{n}{2}^{-1} \sum_{\mathbf{i} \in C_2^n} \left[\varphi(f_\mathbf{i}) - E_{\theta_n}\{\varphi(f_\mathbf{i})\}\right] &= \sqrt{n}\frac{2}{n}\sum_{i=1}^{n} E\left(\varphi(f_\mathbf{i}) \mid Z_i\right) + \mathbf{o}_p(1) \\ &= \frac{\sqrt{n}}{n}\sum_{i=1}^{n} 2E\left(\varphi(f_\mathbf{i}) \mid Z_i\right) + \mathbf{o}_p(1) \\ &\to_d N(\mathbf{0}, 4\Sigma). \end{aligned}$$

Combining with

$$n^{1/2}\{\beta(\theta_n)-\beta(\theta_0)\}\to\Gamma(\theta_0)\tau,$$

where $\Gamma(\theta_0)=\partial\beta(\theta_0)/\partial\theta^{\top}$, and

$$n^{1/2}E_{\theta_n}\{\varphi(f_{\mathbf{i}})\}\to E_{\theta_0}\{\varphi(f_{\mathbf{i}})S_{\theta}^{T}(f,\theta_0)\}\tau.$$

which requires the Taylor expansion w.r.t the density function $p(f,\theta_n)$. Through some algebra, we then proved the theorem.

4. **Proof of Theorem 4.5**

Let $\widehat{\eta}_n$ denote a $\sqrt{n}$-consistent estimator of η, i.e.,

$$\sqrt{n}(\widehat{\eta}_n-\eta_0)=\mathbf{O}_p(1),$$

where $\mathbf{O}_p(\cdot)$ denotes stochastic boundedness.

First, note that

$$E[m(f_{\mathbf{i}},\beta_0,\eta_0)]=\mathbf{0}.$$

So

$$\int m(f_{\mathbf{i}},\beta_0,\eta_0)\,p(f_{\mathbf{i}},\beta_0,\eta_0)\,d\nu(f_{\mathbf{i}})=\mathbf{0},$$
$$\frac{\partial}{\partial\eta^{\top}}\int m(f_{\mathbf{i}},\beta_0,\eta_0)\,p(f_{\mathbf{i}},\beta_0,\eta_0)\,d\nu(f_{\mathbf{i}})=\mathbf{0}.$$

Thus,

$$\begin{aligned}
\mathbf{0} &= \int \frac{\partial m(f_{\mathbf{i}},\beta_0,\eta_0)}{\partial \eta^\top} p(f_{\mathbf{i}},\beta_0,\eta_0)\, d\nu(f_{\mathbf{i}}) \\
&+ \int m(f_{\mathbf{i}},\beta_0,\eta_0) \frac{\frac{\partial p(f_{\mathbf{i}},\beta_0,\eta_0)}{\partial \eta^\top}}{p(Z_i,\beta_0,\eta_0)} p(f_{\mathbf{i}},\beta_0,\eta_0)\, d\nu(f_{\mathbf{i}}) \\
&= \int \frac{\partial m((f_{\mathbf{i}},\beta_0,\eta_0))}{\partial \eta^\top} p(f_{\mathbf{i}},\beta_0,\eta_0)\, d\nu(f_{\mathbf{i}}) \\
&+ \int m(f_{\mathbf{i}},\beta_0,\eta_0) S_\eta^\top(f_{\mathbf{i}},\beta_0,\eta_0)\, p(f_{\mathbf{i}},\beta_0,\eta_0)\, d\nu(f_{\mathbf{i}}).
\end{aligned}$$

It follows that

$$\begin{aligned}
\mathbf{0} &= E\left[\frac{\partial m(f_{\mathbf{i}},\beta_0,\eta_0)}{\partial \eta^\top}\right] + Em(f_{\mathbf{i}},\beta_0,\eta_0) S_\eta^\top(f_{\mathbf{i}},\beta_0,\eta_0) \\
&= E\left[\frac{\partial m(f_{\mathbf{i}},\beta_0,\eta_0)}{\partial \eta^\top}\right] + E[\varphi(f_{\mathbf{i}},\beta_0,\eta_0) - E[\varphi(f_{\mathbf{i}},\beta_0,\eta_0)]] S_\eta^\top(f_{\mathbf{i}},\beta_0,\eta_0) \\
&= E\left[\frac{\partial m(f_{\mathbf{i}},\beta_0,\eta_0)}{\partial \eta^\top}\right] + E\left[\varphi(f_{\mathbf{i}},\beta_0,\eta_0) S_\eta^\top(f_{\mathbf{i}},\beta_0,\eta_0)\right].
\end{aligned}$$

by the definition of $m(f_{\mathbf{i}},\beta_0,\eta_0)$.

From fact that the influence function $\varphi(f_{\mathbf{i}})$ satisfies corollary (ii): $E\left[\varphi(f_{\mathbf{i}}) S_\eta^\top(f_{\mathbf{i}},\beta_0,\eta_0)\right] = \mathbf{0}$, we have:

$$E\left[\frac{\partial m(f_{\mathbf{i}},\beta_0,\eta_0)}{\partial \eta^\top}\right] = \mathbf{0}.$$

With similar argument by expanding w.r.t. β, we have:

$$E\left[\frac{\partial m(f_{\mathbf{i}},\beta_0,\eta_0)}{\partial \beta^\top}\right] = -\mathbf{I}_q.$$

Now expand the following equations around (β_0,η_0),

$$\sum_{\mathbf{i}\in C_2^n} m\left(f_{\mathbf{i}}, \widehat{\beta}_n, \widehat{\eta}_n\left(\widehat{\beta}_n\right)\right) = \mathbf{0},$$

we have:

$$\begin{aligned}
\mathbf{0} &= \sum_{\mathbf{i}\in C_2^n} m\left(f_{\mathbf{i}}, \widehat{\beta}_n, \widehat{\eta}_n\left(\widehat{\beta}_n\right)\right) \\
&= \sum_{\mathbf{i}\in C_2^n} m(f_{\mathbf{i}}, \beta_0, \eta_0) + \sum_{\mathbf{i}\in C_2^n} \frac{\partial}{\partial \beta^\top} m(f_{\mathbf{i}}, \beta_0, \eta_0)\left(\widehat{\beta}_n - \beta_0\right) + \\
&+ \sum_{\mathbf{i}\in C_2^n} \frac{\partial}{\partial \eta^\top} m(f_{\mathbf{i}}, \beta_0, \eta_0)\left(\widehat{\eta}_n\left(\widehat{\beta}_n\right) - \eta_0\right) + \mathbf{o}_p\left(n^{-\frac{1}{2}}\right).
\end{aligned}$$

Then

$$\begin{aligned}
\sqrt{n}\left(\widehat{\beta}_n - \beta_0\right) = &-\sqrt{n}\left[\binom{n}{2}^{-1} \sum_{\mathbf{i}\in C_2^n} \frac{\partial}{\partial \beta^\top} m(f_{\mathbf{i}}, \beta_0, \eta_0)\right]^{-1} \left[\binom{n}{2}^{-1} \sum_{\mathbf{i}\in C_2^n} m(f_{\mathbf{i}}, \beta_0, \eta_0)\right] \\
&- \sqrt{n}\left[\binom{n}{2}^{-1} \sum_{\mathbf{i}\in C_2^n} \frac{\partial}{\partial \beta^\top} m(f_{\mathbf{i}}, \beta_0, \eta_0)\right]^{-1} \\
&\left[\binom{n}{2}^{-1} \sum_{\mathbf{i}\in C_2^n} \frac{\partial}{\partial \eta^\top} m(f_{\mathbf{i}}, \beta_0, \eta_0)\left(\widehat{\eta}_n\left(\widehat{\beta}_n\right) - \eta_0\right)\right] + \mathbf{o}_p(1).
\end{aligned}$$

Since

$$\begin{aligned}
\binom{n}{2}^{-1} \sum_{\mathbf{i}\in C_2^n} \frac{\partial}{\partial \beta^\top} m(f_{\mathbf{i}}, \beta_0, \eta_0) &\longrightarrow^p E\left[\frac{\partial}{\partial \beta^\top} m(f_{\mathbf{i}}, \beta_0, \eta_0)\right] = -\mathbf{I}_q, \\
\binom{n}{2}^{-1} \sum_{\mathbf{i}\in C_2^n} \frac{\partial}{\partial \eta^\top} m(f_{\mathbf{i}}, \beta_0, \eta_0) &\longrightarrow^p E\left[\frac{\partial}{\partial \eta^\top} m(f_{\mathbf{i}}, \beta_0, \eta_0)\right] = \mathbf{0}, \\
\left(\widehat{\eta}_n\left(\widehat{\beta}_n\right) - \eta_0\right) &= \mathbf{O}_p(1),
\end{aligned}$$

it follows that

$$\begin{aligned}\sqrt{n}\left(\widehat{\beta}_n-\beta_0\right) &= \sqrt{n}\binom{n}{2}^{-1}\sum_{\mathbf{i}\in C_2^n} m\left(f_{\mathbf{i}},\beta_0,\eta_0\right)+\mathbf{o}_p\left(1\right)\\ &= \sqrt{n}\binom{n}{2}^{-1}\sum_{\mathbf{i}\in C_2^n}\left\{\varphi\left(f_{\mathbf{i}},\beta_0,\eta_0\right)-E\left[\varphi\left(f_{\mathbf{i}},\beta_0,\eta_0\right)\right]\right\}+\mathbf{o}_p\left(1\right)\\ &= \sqrt{n}\binom{n}{2}^{-1}\sum_{\mathbf{i}\in C_2^n}\varphi\left(f_{\mathbf{i}},\beta_0,\eta_0\right)+\mathbf{o}_p\left(1\right)\end{aligned}$$

5. Proof of Theorem 4.6. The set of all influence functions $\varphi\left(f_{\mathbf{i}}\right): E\left[\varphi\left(f_{\mathbf{i}}\right)S_\theta^\top\left(f_{\mathbf{i}},\theta_0\right)\right]=\Gamma\left(\theta_0\right)$ is the affine space/linear variety $\varphi^*\left(f_{\mathbf{i}}\right)+\pounds^\perp$, where $\varphi^*\left(Z\right)$ is some influence function and $\pounds^\perp$ is orthogonal to the tangent space spanned by $S_\theta\left(Z\right)$.

$\forall l\in H$, let $\varphi\left(f_{\mathbf{i}}\right)=\varphi^*\left(f_{\mathbf{i}}\right)+l\left(f_{\mathbf{i}}\right)$

$$\begin{aligned}\Gamma\left(\theta_0\right) &= E\left[\varphi\left(f_{\mathbf{i}}\right)S_\theta^\top\left(f_{\mathbf{i}},\theta_0\right)\right]=E\left[\left\{\varphi^*\left(f_{\mathbf{i}}\right)+l\left(f_{\mathbf{i}}\right)\right\}\cdot S_\theta^\top\left(f_{\mathbf{i}},\theta_0\right)\right]\\ &= E\left[\varphi^*\left(f_{\mathbf{i}}\right)S_\theta^\top\left(f_{\mathbf{i}},\theta_0\right)\right]+E\left[l\left(f_{\mathbf{i}}\right)S_\theta^\top\left(f_{\mathbf{i}},\theta_0\right)\right]\\ &= E\left[\varphi^*\left(f_{\mathbf{i}}\right)S_\theta^\top\left(f_{\mathbf{i}},\theta_0\right)\right]=\Gamma\left(\theta_0\right)\end{aligned}$$

Therefore, $E\left[l\left(f_{\mathbf{i}}\right)S_\theta^\top\left(f_{\mathbf{i}},\theta_0\right)\right]=0$, i.e. $l\in T^\perp$.

6. Proof of Theorem 4.7. The efficient influence function is given by

$$\begin{aligned}\varphi_{eff}\left(f_{\mathbf{i}}\right) &= \varphi^*\left(f_{\mathbf{i}}\right)-\Pi\left(\varphi^*\left(f_{\mathbf{i}}\right)\mid\pounds^\perp\right)\\ &= \Pi\left(\varphi^*\left(f_{\mathbf{i}}\right)\mid\pounds\right)\\ &= \Gamma\left(\theta_0\right)I^{-1}\left(\theta_0\right)S_\theta\left(f_{\mathbf{i}},\theta_0\right).\end{aligned}$$

For $\forall$ influence function $\varphi\left(f\right)$, $\varphi(f)=\varphi^*\left(f\right)+\pounds^\perp$ and $\Pi\left(\varphi^*\left(f\right)\mid\pounds^\perp\right)\in\pounds^\perp$. It follows

that

$$\varphi(f) = \varphi_{eff}(f) + l, \quad l \in \pounds^{\perp},$$
$$Var(\varphi(f)) = Var\left(\varphi_{eff}(f)\right) + Var(l), \quad \forall l \in \pounds^{\perp}.$$

By construction,

$$\varphi_{eff}(f) = \Pi(\varphi^*(f) \mid \pounds) = B_{eff} S_\theta(f, \theta_0) \text{ for some } B_{eff}.$$

Since $E\left[\varphi_{eff}(f) S_\theta^\top(f, \theta_0)\right] = \Gamma(\theta_0)$, it follows that $\varphi_{eff}(f)$ satisfies that

$$\Gamma(\theta_0) = B_{eff} E\left[S_\theta(f, \theta_0) S_\theta^\top(f, \theta_0)\right] = B_{eff} I(\theta_0),$$

with $I(\theta_0)$ the information matrix. So

$$B_{eff} = \Gamma(\theta_0) I^{-1}(\theta_0),$$
$$\varphi_{eff}(f) = \Gamma(\theta_0) I^{-1}(\theta_0) S_\theta(f, \theta_0).$$

7. Proof of Corollary 2.

Need to show: The unique efficient influence function is a scaled version of the efficient score:

$$\varphi_{eff}(f_{\mathbf{i}}, \theta_0) = E^{-1}\left(S_{eff} S_{eff}^\top\right) S_{eff}(f_{\mathbf{i}}, \theta_0)$$

By definition,

$$S_{eff}(Z, \theta_0) \perp \Lambda, \quad E\left[S_{eff} \Pi(S_\beta(Z, \theta_0) \mid \Lambda)\right] = \mathbf{0},$$

it follows that

$$E\left(S_{eff}S_{\beta}^{\top}\right)=E\left(S_{eff}S_{eff}^{\top}\right)+E\left[S_{eff}\Pi(S_{\beta}\left(Z,\theta_{0}\right)|\Lambda)\right]=E\left(S_{eff}S_{eff}^{\top}\right).$$

Let $\varphi_{eff}\left(f,\theta_{0}\right)=E^{-1}\left(S_{eff}S_{eff}^{\top}\right)S_{eff}\left(f,\theta_{0}\right)$. Then, $\varphi_{eff}\left(f,\theta_{0}\right)$ satisfies corollary (i) and (ii), since

$$\begin{aligned}
&E\left(S_{eff}S_{\beta}^{\top}\right)=E^{-1}\left(S_{eff}S_{eff}^{\top}\right)E\left(S_{eff}S_{\beta}^{\top}\right)=E^{-1}\left(S_{eff}S_{eff}^{\top}\right)E\left(S_{eff}S_{eff}^{\top}\right)=\mathbf{I}_{q},\\
&E\left(S_{eff}S_{\eta}^{\top}\right)=E^{-1}\left(S_{eff}S_{eff}^{\top}\right)\left(S_{eff}S_{\eta}^{\top}\right)=\mathbf{0}.
\end{aligned}$$

Thus, $\varphi_{eff}\left(Z,\theta_{0}\right)$ is an influence function.

Since φ_{eff} **is the unique influence function in the tangent space** $\pounds$, and both $S_{\beta}\left(f,\theta_{0}\right)\in\pounds$ and $\Pi(S_{\beta}\left(f,\theta_{0}\right)|\Lambda)\in\pounds$, we have

$$\begin{aligned}
\varphi_{eff}\left(f,\theta_{0}\right)&=E^{-1}\left(S_{eff}S_{eff}^{\top}\right)S_{eff}\left(f,\theta_{0}\right)\\
&=E^{-1}\left(S_{eff}S_{eff}^{\top}\right)\left[S_{\beta}\left(f,\theta_{0}\right)-\Pi\left(S_{\beta}\left(f,\theta_{0}\right)|\Lambda\right)\right]\\
&\in\pounds,
\end{aligned}$$

Thus $\varphi_{eff}\left(f,\theta_{0}\right)$ is the efficient influence function for RAL estimators of β.

6.3.2 Super-efficient Estimators of Between-subject Attributes

We give an example of *super-efficient estimator* for between-subject attributes that are exogenous.

Consider a sequence of normally distributed within-subject attributes $Y_{i}\sim^{i.i.d}N\left(\mu,\sigma^{2}\right)$. To construct an exogenous between-subject attributes $f_{\mathbf{i}}$, we can**,** for example, let $f_{\mathbf{i}}=Y_{i_{1}}-Y_{i_{2}}$, $\mathbf{i}=(i_{1},i_{2})\in C_{2}^{n}$, which yields $f_{\mathbf{i}}\sim^{i.d.}N\left(0,2\sigma^{2}\right)$. If interest lies in estimating $\beta=2\sigma^{2}$, either Y_{i} or $f_{\mathbf{i}}$ can be used. Since the MLE based on Y_{i} is efficient, we use its CR bound as the benchmark

to demonstrate the efficiency for estimators based on $f_{\mathbf{i}}$.

Let $\beta_0 = 2\sigma_0^2$, $\overline{Y} = (1/n)\sum_{i=1}^n Y_i$. Maximizing the log-likelihood of Y_i yields the MLE $\widehat{\beta}_Y^{\text{mle}} = (2/n)\sum_{i=1}^n \left(Y_i - \overline{Y}\right)^2$ that is AL:

$$\sqrt{n}\left(\widehat{\beta}_Y^{\text{mle}} - \beta_0\right) = n^{-1/2}\sum_{i=1}^n \varphi_{\text{mle}}(Y_i) + o_p(1),\ \varphi_{\text{mle}}(Y_i) = 2\left(Y_i - \mu_0\right)^2 - \beta_0. \tag{6.13}$$

By CLT,

$$\sqrt{n}\left(\widehat{\beta}_Y^{\text{mle}} - \beta_0\right) \longrightarrow^d N\left(0, \upsilon\right),\ \upsilon = 2\beta_0^2. \tag{6.14}$$

where υ is our benchmark, the smallest asymptotic variance among all RAL estimators of β based on Y_i.

For between-subject attributes $f_{\mathbf{i}} \sim^{i.d.} N\left(0, \beta\right)$, first consider the non-overlapping model 1, whose (parametric) efficient I.F. is found by scaling the individual score for $f_{\mathbf{i}}$ under $N\left(0, \beta\right)$:

$$\varphi_{\text{eff1}}\left(f_{\mathbf{i}}\right) = f_{\mathbf{i}}^2 - \beta_0 = \left(Y_{i_1} - Y_{i_2}\right)^2 - \beta_0, \tag{6.15}$$

which is in the equivalent class of the efficient I.F. $\psi_{\text{eff2}}\left(f_{\mathbf{i}}\right)$ for enumerated model 2, whose norm b2 equals to

$$\upsilon_2 = \left\|\psi_{\text{eff2}}\left(f_{\mathbf{i}}\right)\right\|_{b2} = \left\|\varphi_{\text{eff1}}\left(f_{\mathbf{i}}\right)\right\|_{b2} = Var\left(2E\left[\varphi_{\text{eff1}}\left(f_{\mathbf{i}}\right) \mid Y_{i_1}\right]\right) = 2\beta_0^2 = \upsilon.$$

For enumerated model 2, $\widehat{\beta}_f^{\text{eff}} = \binom{n}{2}^{-1}\sum_{\mathbf{i}\in C_2^n} f_{\mathbf{i}}^2$ is AN with variance achieving the benchmark υ, indicating that compared with Y_i, there is no loss of information in estimating β using $f_{\mathbf{i}}$, and $\widehat{\beta}_f^{\text{eff}}$ is as efficient as $\widehat{\beta}_Y^{\text{mle}}$.

The super-efficient Estimator for β

Now we introduce an example of a *super-efficient estimator* for between-subject attribute $f_{\mathbf{i}} \sim^{i.d.} N\left(0, \beta\right)$. The efficient estimator is $\widehat{\beta}_f^{\text{eff}} = \binom{n}{2}^{-1}\sum_{\mathbf{i}\in C_2^n} f_{\mathbf{i}}^2$, with $n^{1/2}\left(\widehat{\beta}_f^{\text{eff}} - \beta_0\right) \rightarrow_d$

$N\left(0, 2\beta_0^2 = \upsilon\right)$. Without loss of generality, let $\upsilon = 1$. Construct another estimator

$$\widehat{\beta}_f^s = \begin{cases} 1/\sqrt{2}, & \text{if } \left|\widehat{\beta}_f^{\text{eff}} - 1/\sqrt{2}\right| < n^{-1/4}, \\ \widehat{\beta}_f^{\text{eff}}, & \text{otherwise.} \end{cases}$$

Then it is readily shown that

$$\sqrt{n}\left(\widehat{\beta}_f^s - \beta_0\right) \longrightarrow^d \begin{cases} N(0,0), & \text{if } \beta_0 = 1/\sqrt{2}, \\ N(0, \upsilon = 1), & \text{if } \beta_0 \neq 1/\sqrt{2}. \end{cases}$$

This new estimator $\widehat{\beta}_f^s$ seems more efficient than the $\widehat{\beta}_f^{\text{eff}}$ (or $\widehat{\beta}_Y^{\text{mle}}$): at $\beta_0 = 1/\sqrt{2}$, it has a variance of 0; at other points, $\widehat{\beta}_f^s$ is as efficient as $\widehat{\beta}_f^{\text{eff}}$. This property is termed "*super-efficiency*", which is unnatural with undesirable local properties as we show now. Consider generating the data from a sequence $\beta_n = 1/\sqrt{2} + n^{-1/3}$, which converges to $1/\sqrt{2}$ as $n \longrightarrow \infty$. For the $\widehat{\beta}_f^{\text{eff}}$, we still have $n^{1/2}\left(\widehat{\beta}_f^{\text{eff}} - \beta_n\right) \longrightarrow^d N(0, \upsilon)$. However, $n^{1/2}(\widehat{\beta}_f^s - \beta_n) \longrightarrow_p -\infty$, i.e., if data are generated from this sequence $\beta_n = 1/\sqrt{2} + n^{-1/3}$, which is very close to the point where $\widehat{\beta}_f^s$ is *super-efficient* (i.e., $\beta_0 = 1/\sqrt{2}$), the local properties of this estimator would be undesirable. The super-efficiency is gained at the expense of poor estimation in a neighborhood of zero.

To rule out such estimators, we hence required an estimator to be *regular* in all our discussions.

Proof: Without loss of generality, let $\upsilon = 1$. Let $C_n = \left\{\left|\widehat{\beta}_f^{\text{eff}} - 1/\sqrt{2}\right| < n^{-1/4}\right\}$. Then,

$$\begin{aligned} \Pr(C_n) &= \Pr\left(\left|\widehat{\beta}_f^{\text{eff}} - 1/\sqrt{2}\right| < n^{-1/4}\right) \\ &= \Pr\left[-n^{-1/4} + 1/\sqrt{2} < \binom{n}{2}^{-1} \sum_{\mathbf{i} \in C_2^n} f_{\mathbf{i}}^2 < n^{-1/4} + 1/\sqrt{2}\right] \\ &= \Pr\left[\sqrt{n}\left[-n^{-1/4} - (\beta_0 - 1/\sqrt{2})\right] < \sqrt{n}\left(\widehat{\beta}_f^{\text{eff}} - \beta_0\right) < \sqrt{n}\left[n^{-1/4} - (\beta_0 - 1/\sqrt{2})\right]\right] \\ &= \Phi\left(n^{1/4} - n^{1/2}(\beta_0 - 1/\sqrt{2})\right) - \Phi\left(-n^{1/4} - n^{1/2}(\beta_0 - 1/\sqrt{2})\right). \end{aligned}$$

If $\beta_0 = 1/\sqrt{2}$, $\Pr(\widehat{\beta}_f^s = 1/\sqrt{2}) = \Pr(C_n) \longrightarrow 1 - 0 = 1$, as $n \longrightarrow \infty$, indicating that $\widehat{\beta}_f^s \longrightarrow^p \beta_0$, or $\sqrt{n}\left(\widehat{\beta}_f^s - \beta_0\right) \longrightarrow^d N(0,0)$. If $\beta_0 \neq 1/\sqrt{2}$, $-n^{1/2}(\beta_0 - 1/\sqrt{2})$ dominates inside the CDF, so $\Pr(C_n) \longrightarrow 0 - 0 = 0$ and $\widehat{\beta}_f^s = \widehat{\beta}_f^{\text{eff}}$.

With a smaller asymptotic variance at the exact point $\beta_0 = 1/\sqrt{2}$, $\widehat{\beta}_f^s$ seems more efficient than $\widehat{\beta}_f^{\text{eff}}$. However, this *super-efficiency* is unnatural with undesirable local properties. Consider a sequence $\beta_n = 1/\sqrt{2} + n^{-1/3}$, which converges to $1/\sqrt{2}$ as $n \longrightarrow \infty$. For $\widehat{\beta}_f^{\text{eff}}$,

$$n^{1/2}\left(\widehat{\beta}_f^{\text{eff}} - \beta_n\right) = n^{1/2}(\widehat{\beta}_f^{\text{eff}} - 1/\sqrt{2}) - n^{-1/6} = n^{1/2}(\widehat{\beta}_f^{\text{eff}} - 1/\sqrt{2}) + \mathbf{o}_p(1).$$

By theory of U-statistics,

$$n^{1/2}(\widehat{\beta}_f^{\text{eff}} - 1/\sqrt{2}) = n^{1/2}(\binom{n}{2}^{-1} \sum_{\mathbf{i} \in C_2^n} f_{\mathbf{i}}^2 - 1/\sqrt{2}) \longrightarrow^d N\left(\beta_0 - 1/\sqrt{2}, \upsilon\right) =^{\beta_0 = 1/\sqrt{2}} N(0, \upsilon),$$

so we still have $\sqrt{n}\left(\widehat{\beta}_f^{\text{eff}} - \beta_n\right) \longrightarrow^d N(0, \upsilon)$ if data are generated from this sequence β_n.

However, for $\widehat{\beta}_f^s$ that is "super-efficient", we have

$$\begin{aligned}
\Pr(\widehat{\beta}_f^s = 1/\sqrt{2}) &= \Pr(C_n) \\
&= \Phi\left(n^{1/4} - n^{1/2}(\beta_n - 1/\sqrt{2})\right) - \Phi\left(-n^{1/4} - n^{1/2}(\beta_n - 1/\sqrt{2})\right) \\
&= \Phi(n^{1/4} - n^{-1/6}) - \Phi(-n^{1/4} - n^{-1/6}) \longrightarrow 1 - 0 = 1 \text{ as } n \longrightarrow \infty,
\end{aligned}$$

i.e., $\Pr(\widehat{\beta}_f^s = 1/\sqrt{2}) \longrightarrow 1$, subtracting β_n and multiply by $\sqrt{n}$ on both sides yield

$$\Pr\left[\sqrt{n}\left(\widehat{\beta}_f^s - \beta_n\right) = \sqrt{n}\left(1/\sqrt{2} - \beta_n\right)\right] \longrightarrow 1.$$

Plugging in $\beta_n = 1/\sqrt{2} + n^{-1/3}$ on the RHS yields $n^{1/2}\left(1/\sqrt{2} - \beta_n\right) = -n^{1/6} \longrightarrow -\infty$ as $n \longrightarrow \infty$, so

$$\Pr(\sqrt{n}(\widehat{\beta}_f^s - \beta_n) = -\infty) \longrightarrow 1, \text{ or } \sqrt{n}(\widehat{\beta}_f^s - \beta_n) \longrightarrow_p -\infty,$$

i.e., if data are generated from a sequence $\beta_n = 1/\sqrt{2} + n^{-1/3}$, which is very close to true value of β, the local properties of this estimator would be undesirable.

6.3.3 Details about the Hilbert Space

A. Hilbert Space and Projection Theorem

We briefly review the Hilbert space and associated projection theorem that are fundamental to develop semiparametric efficiency for FRM here, for full materials, please refer to (**Rudin, Walter (1987)**).

1. Hilbert Spaces

A *Hilbert space* $\mathscr{H}$ is a complete, normed linear vector space equipped with an inner product (e.g., covariance inner product $E(\mathbf{h}^\top \mathbf{h})$). A linear subspace must contain the origin. *Completeness* means that every Cauchy sequence converges to an element of the space.

Definition. The *norm*, or "length", of a vector $\mathbf{h} \in \mathscr{H}$ is defined as $\|\mathbf{h}\| = \langle \mathbf{h}, \mathbf{h} \rangle^{1/2}$.

2. Projection Theorem for the Hilbert Space

Theorem A.1. Let $\mathscr{H}$ be a Hilbert space and $\mathscr{U}$ a linear subspace that is *closed* (i.e., contains all its limit points). Corresponding to any $\mathbf{h} \in \mathscr{H}$, there exists a unique $\mathbf{u}_0 \in \mathscr{U}$ that is closest to $\mathbf{h}$; furthermore, $\mathbf{h} - \mathbf{u}_0$ is orthogonal to $\mathscr{U}$; that is,

$$\|\mathbf{h} - \mathbf{u}_0\| \leq \|\mathbf{h} - \mathbf{u}\|, \ \langle \mathbf{h} - \mathbf{u}_0, \mathbf{u} \rangle = 0, \text{ for all } \mathbf{u} \in \mathscr{U}.$$

We refer to $\mathbf{u}_0$ as the *projection* of $\mathbf{h}$ onto the space $\mathscr{U}$, denoted by $\Pi(\mathbf{h} \mid \mathscr{U})$. Moreover, $\mathbf{u}_0$ is the only element $\mathbf{u} \in \mathscr{U}$ such that $\mathbf{h} - \mathbf{u}$ is orthogonal to $\mathscr{U}$. The condition that a Hilbert space be complete is necessary to guarantee the existence of the projection. A formal proof can be found in **Luenberger (1969).**

3. Hilbert Space for Random Vectors

Let $(\mathscr{L}, \mathscr{A}, P)$ denote a probability space, where $\mathscr{L}$ denotes the sample space, $\mathscr{A}$ the σ-algebra and P the probability measure. Let $\mathbf{Z}$ be a random vector of dimension p. Then a *linear subspace* of $\mathbb{R}^q$ is the space consisting of all q-dimensional mean-zero random functions $\mathbf{h}$ of $\mathbf{Z}$, $\mathbf{h} : \mathscr{L} \to \mathbb{R}^q$, where $\mathbf{h}(\mathbf{Z})$ is measurable and satisfies

$$\text{(i) } E\left[\mathbf{h}\left(\mathbf{Z}\right)\right] = 0, \quad \text{(ii) } E\left[\mathbf{h}^{\top}\left(\mathbf{Z}\right)\mathbf{h}\left(\mathbf{Z}\right)\right] < \infty.$$

Lemma A.1. Let $\mathbf{v}\left(\mathbf{Z}\right) = \left(v_1\left(\mathbf{Z}\right), ..., v_r\left(\mathbf{Z}\right)\right)^{\top}$ be an r-dimensional $(r \leq q)$ random function with $E\left[\mathbf{v}\left(\mathbf{Z}\right)\right] = \mathbf{0}$ and $E\left(\mathbf{v}^{\top}\mathbf{v}\right) < \infty$. Consider the linear subspace $\mathscr{U}$ spanned by $\mathbf{v}\left(\mathbf{Z}\right)$:

$$\mathscr{U} = \{\mathbf{B}\mathbf{v}; \quad \text{for any arbitrary } q \times r \text{ matrix } \mathbf{B} \text{ of real numbers}\}.$$

Assuming $E(\mathbf{v}\mathbf{v}^{\top})$ is nonsingular (i.e., positive definite), then the unique *projection* of a p-dimensional $\mathbf{h} \in \mathscr{H}$ onto $\mathscr{U}$ is

$$\Pi\{\mathbf{h}\left(\mathbf{Z}\right) \mid \mathscr{U}\} = E(\mathbf{h}\mathbf{v}^{\top})E^{-1}(\mathbf{v}\mathbf{v}^{\top})\mathbf{v}. \tag{6.16}$$

Proof: Consider the problem of finding the projection of an p-dimensional $\mathbf{h} \in \mathscr{H}$ onto $\mathscr{U}$. Such a projection $\mathbf{B}_0\mathbf{v}$ is unique and must satisfy:

$$E\left\{\left[\mathbf{h}\left(Z\right) - \mathbf{B}_0\mathbf{v}\left(Z\right)\right]^{\top}\mathbf{B}\mathbf{v}\left(Z\right)\right\} = \mathbf{0} \quad \text{for all} \quad \mathbf{B} \in \mathbb{R}^{p\times r},$$

which is equivalent to

$$E\left\{\left[\mathbf{h}\left(Z\right) - \mathbf{B}_0\mathbf{v}\left(Z\right)\right]^{\top}\mathbf{v}\left(Z\right)\right\} = \mathbf{0}.$$

Therefore, we have $E(\mathbf{h}\mathbf{v}^{\top}) = \mathbf{B}_0 E(\mathbf{v}\mathbf{v}^{\top})$. Assuming $E(\mathbf{v}\mathbf{v}^{\top})$ is nonsingular (i.e., positive

definite, or PD),

$$\mathbf{B}_0 = E(\mathbf{h}\mathbf{v}^\top)E^{-1}(\mathbf{v}\mathbf{v}^\top).$$

Hence, the unique projection is

$$\Pi(\mathbf{h}(Z) \mid \mathscr{U}) = E(\mathbf{h}\mathbf{v}^\top)E^{-1}(\mathbf{v}\mathbf{v}^\top)\mathbf{v}.$$

4. q-replicating linear subspaces

For the r-dimensional ($r \leq q$) random function $\mathbf{v}(\mathbf{Z}) = (v_1(\mathbf{Z}), ..., v_r(\mathbf{Z}))^\top$ with $E[\mathbf{v}(\mathbf{Z})] = \mathbf{0}$ and $E\left(\mathbf{v}^\top\mathbf{v}\right) < \infty$. The linear subspace $\mathscr{U}$ spanned by $\mathbf{v}(\mathbf{Z})$:

$$\mathscr{U} = \{\mathbf{B}\mathbf{v}; \quad \text{for any arbitrary } q \times r \text{ matrix } \mathbf{B} \text{ of real numbers}\}$$

is a *q-replicating linear subspace*, since we can define $\mathscr{U}^{(1)} = \{\mathbf{b}^\top \mathscr{K}[\mathbf{v}(Z_\mathbf{i})]; \quad \text{for an arbitrary } r\text{-dimensional constant vector } \mathbf{b}^{r\times 1}\}$ and have $\mathscr{U} = \left[\mathscr{U}^{(1)}\right]^q$. This special form of subspace that we consider throughout allows the *multivariate Pythagoras* to hold, which yield some nice properties as we demonstrate below.

Theorem A.1. indicates

$$\|\mathbf{h}(\mathbf{Z})\|^2 = \|\Pi(\mathbf{h} \mid \mathscr{U})\|^2 + \|\mathbf{h} - \Pi(\mathbf{h} \mid \mathscr{U})\|^2 \geq \|\Pi[\mathbf{h}(\mathbf{Z_i}) \mid \mathscr{U}]\|^2, \tag{6.17}$$

suggesting that in the Hilbert space $\mathscr{H}$, the norm (or the squared distance to the origin) of any element $\mathbf{h}(\mathbf{Z})$ is always larger than or equal to that of its projection onto the subspace $\mathscr{U}$. For $q = 1$ where $h(\mathbf{Z})$ being a scalar, this is equivalent to

$$Var(h(\mathbf{Z})) = Var[\Pi(h \mid \mathscr{U})] + Var[h - \Pi(h \mid \mathscr{U})] \geq Var[\Pi(h \mid \mathscr{U})].$$

But in general for $q > 1$, to compare variance matrix of two q-dimensional functions

$\mathbf{h}_1, \mathbf{h}_2 \in \mathscr{H}$, we evaluate whether $Var(\mathbf{h}_1) - Var(\mathbf{h}_2)$ is nonnegative or nonpositive definite, we say the variance matrix $Var(\mathbf{h}_1) \leq Var(\mathbf{h}_2)$ iff $Var(\mathbf{h}_1) - Var(\mathbf{h}_2)$ is nonnegative definite. With $\mathscr{U}$ being a q-replicating linear space allows us to have below holding true for a q-dimensional $\mathbf{h}(\mathbf{Z_i})$ with $q > 1$ (which in general only hold for $q = 1$) :

$$Var(\mathbf{h}) = Var[\Pi(\mathbf{h} \mid \mathscr{U})] + Var[\mathbf{h} - \Pi(\mathbf{h} \mid \mathscr{U})],$$

indicating that the variance matrix of any element $\mathbf{h}$ is larger than (i.e., the difference being nonnegative definite) the variance matrix of the projection $\Pi(\mathbf{h} \mid \mathscr{U})$, or the variance matrix of the residual after projection $[\mathbf{h} - \Pi(\mathbf{h} \mid \mathscr{U})]$. This is useful when we construct efficient score functions using projection, and we no longer distinguish between the Hilbert space of random functions with $q = 1$ or $q > 1$.

B. Hilbert Space for Between-subject Attributes

1. Inner Prodcut 2 and Norm b2

For the *norm b2* of the between-subject attributes that encompass an FRM form for the correlated $\mathbf{h}(\mathbf{Z_i})$'s, we equipped the Hilbert space $\mathscr{H}_b^{(q)}$ with

$$\langle \mathbf{h}_1(\mathbf{Z_i}), \mathbf{h}_2(\mathbf{Z_i}) \rangle_{b2} = E\left\{ 2E\left[\mathbf{h}_1^{\top}(\mathbf{Z_i}) \mid \mathbf{Z}_{i_1} \right] \cdot 2E\left[\mathbf{h}_2(\mathbf{Z_i}) \mid \mathbf{Z}_{i_1} \right] \right\},$$
$$\|\mathbf{h}(\mathbf{Z_i})\|_{b2} = \langle \mathbf{h}(\mathbf{Z_i}), \mathbf{h}(\mathbf{Z_i}) \rangle_{b2}^{1/2} = E^{1/2}\left\{ 2E\left[\mathbf{h}^{\top}(\mathbf{Z_i}) \mid \mathbf{Z}_{i_1} \right] \cdot 2E\left[\mathbf{h}(\mathbf{Z_i}) \mid \mathbf{Z}_{i_1} \right] \right\}.$$

It is readily checked that this definition of inner product 2 satisfies conditions 1) - 3) below,

1). $\langle \mathbf{h}_1(\mathbf{Z_i}), \mathbf{h}_2(\mathbf{Z_i}) \rangle_{b2} = \langle \mathbf{h}_2(\mathbf{Z_i}), \mathbf{h}_1(\mathbf{Z_i}) \rangle_{b2}$,

2). $\langle a\mathbf{h}_1(\mathbf{Z_i}), \mathbf{h}_2(\mathbf{Z_i}) \rangle_{b2} = a \langle \mathbf{h}_1(\mathbf{Z_i}), \mathbf{h}_2(\mathbf{Z_i}) \rangle_{b2}$,

3). $\langle \mathbf{h}_1(\mathbf{Z_i}) + \mathbf{h}_2(\mathbf{Z_i}), \mathbf{h}_3(\mathbf{Z_i}) \rangle_{b2} = \langle \mathbf{h}_1(\mathbf{Z_i}), \mathbf{h}_2(\mathbf{Z_i}) \rangle_{b2} + \langle \mathbf{h}_1(\mathbf{Z_i}), \mathbf{h}_3(\mathbf{Z_i}) \rangle_{b2}$,

4). $\langle \mathbf{h}(\mathbf{Z_i}), \mathbf{h}(\mathbf{Z_i}) \rangle_{b2} \geq 0$, $\langle \mathbf{h}(\mathbf{Z_i}), \mathbf{h}(\mathbf{Z_i}) \rangle_{b2} = 0$ iff $E[\mathbf{h}(\mathbf{Z_i}) \mid \mathbf{Z}_{i_1}] = \mathbf{0}$ a.s..

For 4), we have that if $E[\mathbf{h}(\mathbf{Z_i}) \mid \mathbf{Z}_{i_1}] = \mathbf{0}$ a.s., then $\|\mathbf{h}(\mathbf{Z_i})\|_{b2}^2 = \langle \mathbf{h}(\mathbf{Z_i}), \mathbf{h}(\mathbf{Z_i}) \rangle_{b2} = 0$. Conversely, $\langle \mathbf{h}(\mathbf{Z_i}), \mathbf{h}(\mathbf{Z_i}) \rangle_{b2} = 0$ implies that for all $1 \leq s \leq q$,

$$E\{E[h_s(\mathbf{Z_i}) \mid \mathbf{Z}_{i_1}] E[h_s(\mathbf{Z_i}) \mid \mathbf{Z}_{i_1}]\} = E\left\{E^2[h_s(\mathbf{Z_i}) \mid \mathbf{Z}_{\mathbf{i}i_1}]\right\} = 0,$$

we then have:

$$E[h_s(\mathbf{Z_i}) \mid \mathbf{Z}_{i_1}] = 0 \; a.s. \text{ for all } 1 \leq s \leq q, \text{ i.e., } E[\mathbf{h}(\mathbf{Z_i}) \mid \mathbf{Z}_{i_1}] = \mathbf{0} \; a.s..$$

Thus,

$$\langle \mathbf{h}(\mathbf{Z_i}), \mathbf{h}(\mathbf{Z_i}) \rangle_{b2} = 0 \text{ iff } E[\mathbf{h}(\mathbf{Z_i}) \mid \mathbf{Z}_{i_1}] = \mathbf{0} \; a.s..$$

In general, $\langle \mathbf{h}(\mathbf{Z_i}), \mathbf{h}(\mathbf{Z_i}) \rangle_{b2} = 0$ does not imply $\mathbf{h}(\mathbf{Z_i}) = \mathbf{0}$ $a.s.$. To see this, consider a counterexample

$$Z_{i_1}, Z_{i_2} \sim N(1,1), \quad h(Z_\mathbf{i}) = h(Z_{i_1}, Z_{i_2}) = (1 - Z_{i_1})(1 - Z_{i_2}).$$

Then, $h(Z_\mathbf{i}) = h(Z_{i_1}, Z_{i_2})$ is summetric and although $\langle h(Z_\mathbf{i}), h(Z_\mathbf{i}) \rangle_{b2} = \|h(Z_\mathbf{i})\|_{b2}^2 = 0$, since

$$E[h(Z_\mathbf{i}) \mid Z_{i_1}] = (1 - Z_{i_1}) E(1 - Z_{i_2}) = 0 \; a.s.,$$

in general,

$$h(Z_\mathbf{i}) \neq 0 \; a.s.,$$

i.e., here $\|h(Z_\mathbf{i})\|_{b2}^2 = 0$ iff $E[h(Z_\mathbf{i}) \mid Z_{i_1}] = 0$ a.s., but $\|h(Z_\mathbf{i})\|_{b2}^2 = 0$ does not imply $h(Z_\mathbf{i}) = 0$ a.s..

Thus, unlike the origin of $\mathscr{H}_b$ under the inner product 1, the origin of $\mathscr{H}_b$ under inner product 2 is not the equivalence class of $\mathbf{h}(\mathbf{Z_i})$ with $\mathbf{h}(\mathbf{Z_i}) = \mathbf{0}$ a.s., but a larger equivalence class consisting of functions $\mathbf{h}(\mathbf{Z_i})$ such that $E[\mathbf{h}(\mathbf{Z_i}) \mid \mathbf{Z}_{i_1}] = \mathbf{0}$ a.s..

6.3.4 Detailed Simulation Settings

Within-subject Regression

Without loss of generality, we include one continuous covariate $X_i \sim^{i.i.d} U(a,b)$, with $U(a,b)$ denoting a uniform distribution over (a,b). Given X_i, $Y_i \sim NB\left(\tau, h_i(\beta^\top X_i)\right)$, with $h_i(\beta^\top X_i) = \exp(\beta_0 + \beta_1 X_i)$, and $NB(\tau,\mu)$ denoting a Negative Binomial with mean μ and dispersion paramater τ.

To demonstrate the local efficiency of GEE for count responses, we first simulate $X_i \sim^{i.i.d} Unif(a,b)$. With

$$E(Y_i \mid x_i) = \exp(\beta_0 + \beta_1 x_i) = h_i(\beta^\top X_i),\ \beta = (\beta_0, \beta_1),$$

we then simulate overdispersed $Y_i \sim NB(\tau, h_i(\beta^\top X_i))$.

We then estimated β using *1)* parametric MLEs from *a)* Negative Binomial (NB) and *b)* Poisson distributions, respectively. To compare, we applied the *2)* semiparametric GEE for within-subject attributes, with working variance assumptions from *a)* NB and *b)* Poisson, respectively. Specifically, the working variance of NB is $Var(Y_i \mid x_i) = \mu_i(\beta)/p_i(\beta)$, where $p_i(\beta) = \tau/(\tau + \mu_i(\beta))$. In the simulation, this additional parameter τ was estimated from the sample.

Let $\widehat{\beta}^{(m)}$ denote the estimator of β and $\widehat{\Sigma}_\beta^{(m)}$ the asymptotic variance from the mth MC iteration, $\widehat{\beta}$ and $\widehat{\Sigma}_\beta^{(asymp)}$ denote the sample mean of $\widehat{\beta}^{(m)}$ and $\widehat{\Sigma}_\beta^{(m)}$, respectively, and let $\widehat{\Sigma}_\beta^{(emp)}$ denote the sample variance of $\widehat{\beta}^{(m)}$. We can then assess the asymptotic performances by comparing asymptotic and empirical variances from $\widehat{\Sigma}_\beta^{(asymp)}$ and $\widehat{\Sigma}_\beta^{(emp)}$. We set $\tau = 10$, $\beta_0 = 3$, $\beta_1 = 3$, $a = 0$, $b = 1$ and report the parameter estimators (Est.), their asymptotic (Asy.) and empirical (Emp.) variances under different sample sizes in **Supplemental Table 1**.

********** **Supplemental Table 1** goes here **********

Supplemental Table 1 shows that for within-subject attributes, MLEs and GEEs all

yield unbiased estimators. As the true data was generated from NB, the MLE from NB reaches the CR bound and hence is the most efficient. Note that except for the Poisson MLE, all other methods yield close asymptotic and empirical variances as expected, this indicates the drawback of parametric MLE, i.e., if the fitted parametric model deviates from the true data, the obtained MLE will be incorrect. While semiparametric models alleviate this issue as we impose no parametric assumption. The two GEE estimators are both unbiased, yet the one with the correct working variance assumption of NB has smaller variances and reaches the semiparametric bound.

Between-subject regression

Now we conduct a similar simulation for between-subject attributes, to demonstrate the local efficiency of UGEE for count responses. We first simulate $X_i \sim^{i.i.d} Unif(a,b)$, then construct $X_{\mathbf{i}} = X_{i_1} + X_{i_2}$. Let

$$E(f_{\mathbf{i}} \mid x_{\mathbf{i}}) = \exp(\beta_0 + \beta_1 x_{\mathbf{i}}) = h_{\mathbf{i}}(\beta),\ \beta = (\beta_0, \beta_1),$$

we can simulate overdispersed $f_{\mathbf{i}} \sim NB(\tau, h_{\mathbf{i}}(\beta))$ following a Negative Binomial distribution with mean $h_{\mathbf{i}}(\beta)$ and dispersion paramater τ (or the shape parameter of the gamma mixing distribution).

We then estimate β using

1) **The working-MLE of Negative Binomial through** $f_{\mathbf{i}}$;

2) **Semiparametric UGEE with**

$$\mathbf{U}_n(\beta) = \sum_{\mathbf{i} \in C_2^n} \mathbf{D}_{\mathbf{i}}^\top V_{\mathbf{i}}^{-1} S_{\mathbf{i}},\ S_{\mathbf{i}} = f_{\mathbf{i}} - h_{\mathbf{i}},\ \mathbf{D}_{\mathbf{i}} = \frac{\partial}{\partial \beta^\top} h_{\mathbf{i}}(\beta).$$

For the unknown $V_{\mathbf{i}}$, we respectively chose

a) **the true variance of Negative Binomial for** $f_{\mathbf{i}}$. Let

$$V_{\mathbf{i}} = Var\left(f_{\mathbf{i}} \mid x_{\mathbf{i}}\right) = \frac{h_{\mathbf{i}}(\beta)}{p_{\mathbf{i}}(\beta)}, \quad p_{\mathbf{i}}(\beta) = \frac{\tau}{\tau + h_{\mathbf{i}}(\beta)}.$$

The optimal UGEE for estimating β then becomes

$$\mathbf{U}_n\left(\beta\right) = \sum_{\mathbf{i} \in C_2^n} \mathbf{D}_{\mathbf{i}}^{\top} V_{\mathbf{i}}^{-1} S_{\mathbf{i}} = \sum_{\mathbf{i} \in C_2^n} \mathbf{X}_{\mathbf{i}}^{\top} p_{\mathbf{i}}(\beta)\left[f_{\mathbf{i}} - h_{\mathbf{i}}(\beta)\right] = 0, \tag{6.18}$$

yielding the asymptotic variance of

$$\begin{aligned} Var(\beta) &= \mathbf{B}^{-1} 4 Var\left[\mathbf{X}_{\mathbf{i}}^{\top} p_{\mathbf{i}}(\beta)\left\{f_{\mathbf{i}} - h_{\mathbf{i}}(\beta)\right\} \mid f_{i_1}, \mathbf{X}_{i_1}\right] \mathbf{B}^{-1}, \\ \mathbf{B} &= E\left[\mathbf{X}_{\mathbf{i}}^{\top} p_{\mathbf{i}}(\beta) h_{\mathbf{i}}(\beta) E\left\{h_{\mathbf{i}}(\beta)\right\}\right]. \end{aligned}$$

In the simulation, we estimate τ from the sample.

b) **a (wrong) variance assumption of** *Poisson* **for** θ **through** $f_{\mathbf{i}}$. Let

$$V_{\mathbf{i}} = Var\left(f_{\mathbf{i}} \mid x_{\mathbf{i}}\right) = h_{\mathbf{i}}(\beta).$$

The optimal UGEE for estimating β is

$$\mathbf{U}_n\left(\beta\right) = \sum_{\mathbf{i} \in C_2^n} \mathbf{D}_{\mathbf{i}}^{\top} V_{\mathbf{i}}^{-1} S_{\mathbf{i}} = \sum_{\mathbf{i} \in C_2^n} \mathbf{X}_{\mathbf{i}}^{\top}\left\{f_{\mathbf{i}} - h_{\mathbf{i}}(\beta)\right\} = 0, \tag{6.19}$$

yielding the asymptotic variance of

$$Var(\beta) = E\left[\mathbf{X}_{\mathbf{i}}^{\top} h_{\mathbf{i}}(\beta) \mathbf{X}_{\mathbf{i}}\right]^{-1} 4 Var\left[\mathbf{X}_{\mathbf{i}}^{\top}\left\{f_{\mathbf{i}} - h_{\mathbf{i}}(\beta)\right\} \mid f_{i_1}, \mathbf{X}_{i_1}\right] E\left[\mathbf{X}_{\mathbf{i}}^{\top} h_{\mathbf{i}}(\beta) \mathbf{X}_{\mathbf{i}}\right]^{-1}.$$

c) **a bad (wrong) variance assumption (Constant) through** $f_{\mathbf{i}}$. Let

$$V_{\mathbf{i}} = Var\left\{f_{\mathbf{i}} \mid x_{\mathbf{i}}\right\} = C.$$

The optimal UGEE for estimating β is now

$$\mathbf{U}_n(\beta) = \sum_{\mathbf{i}\in C_2^n} \mathbf{D}_\mathbf{i}^\top V_\mathbf{i}^{-1} S_\mathbf{i} = \sum_{\mathbf{i}\in C_2^n} C^{-1}\mathbf{X}_\mathbf{i}^\top h_\mathbf{i}(\beta)\{f_\mathbf{i} - h_\mathbf{i}(\beta)\} = 0, \tag{6.20}$$

yielding the asymptotic variance of

$$Var(\beta) = E\left[\mathbf{X}_\mathbf{i}^\top h_\mathbf{i}(\beta)^2 \mathbf{X}_\mathbf{i}\right]^{-1} 4Var\left[\mathbf{X}_\mathbf{i}^\top h_\mathbf{i}(\beta)\{f_\mathbf{i} - h_\mathbf{i}(\beta)\} \mid f_{i_1}, \mathbf{X}_{i_1}\right] E\left[\mathbf{X}_\mathbf{i}^\top h_\mathbf{i}(\beta)^2 \mathbf{X}_\mathbf{i}\right]^{-1}.$$

In the simulation, we used $C = \widehat{Var}(f_\mathbf{i})$.

We set $\tau = 10, \beta_0 = 3, \beta_1 = 3, a = 0, b = 1$ in all our similations.

6.3.5 Examples of Efficient Estimators for FRM

Binary Responses

Let $f_\mathbf{i}$ denote an *exogenous* outcome. For example, let $f_\mathbf{i} = Z_i - Z_j$, $Z_i \sim^{i.i.d} Bern(p)$.

0. MLE of p through Z_i

The MLE of p can be estimated via maximizing the log-likelihood $l_n(z_i; p) = \sum_{i=1}^n z_i \log(p) + (1-z_i)\log(1-p)$, which yields the MLE $\widehat{p}_n^{\text{mle}}(Z_i) = (1/n)\sum_{i=1}^n Z_i$.

This MLE $\widehat{p}_n^{\text{mle}}(Z_i)$ is AL, with the expansion

$$n^{1/2}\left(\widehat{p}_n^{\text{mle}}(Z_i) - p_0\right) = n^{-1/2}\sum_{i=1}^n (Z_i - p_0) + o_p(1), \quad \varphi_{p_0}^{\text{mle}}(Z_i; p_0) = Z_i - p_0.$$

By CLT, $n^{1/2}\left(\widehat{p}_n^{\text{mle}}(Z_i) - p_0\right) \longrightarrow^d N\left(0, \Sigma_{p_0}^{\text{mle}} = (1/n)p_0(1-p_0)\right)$. Thus, $\Sigma_{p_0}^{\text{mle}}$ is the variance of MLE $\widehat{p}_n^{\text{mle}}(Z_i)$ and hence the CR bound.

Now assume we are interested in estimating $\theta = 2p(1-p)$, then its MLE is simply the plug-in estimator $\widehat{\theta}_n^{\text{mle}}(Z) = 2\overline{Z}(1-\overline{Z}) = (1/n^2)\sum_{i=1}^n \sum_{j=1}^n 2Z_i(1-Z_j)$.

1. MLE of $\theta = 2p(1-p)$ through Z_i

It turns out that

$$Var(\overline{Z}(1-\overline{Z})) = \frac{1}{n}(\mu_4 - \mu_2^2) + O(n^{-2}).$$

where μ_k is the kth central moment of Z_i. Therefore, the asymptotic variance of $\widehat{\theta}_n^{\text{mle}}(Z)$ is

$$\Sigma_{\theta_0}^{\text{mle}}(Z_i) = \frac{4}{n}(\mu_4 - \mu_2^2) = \frac{1}{n} 2\theta_0(1-2\theta_0). \tag{6.21}$$

2. MLE of $\theta = 2p(1-p)$ through $f_{\mathbf{i}}$

With $\mathbf{i} = (i,j) \in C_2^n$, $f_{\mathbf{i}} = Z_i - Z_j$ and $f_{\mathbf{i}}^2$ respectively simplify to

$$f_{\mathbf{i}} = \begin{cases} 1, & \text{if } Z_i = 1, Z_j = 0, \\ -1, & \text{if } Z_i = 0, Z_j = 1, \\ 0, & \text{if } Z_i = Z_j, \end{cases} \quad f_{\mathbf{i}}^2 = \begin{cases} 1, & \text{w.p. } 2p(1-p), \\ 0, & \text{w.p. } 1-2p(1-p). \end{cases}$$

Hence, $f_{\mathbf{i}}^2$ follows $Bern(\theta)$, $\theta = 2p(1-p)$.

Then the log-likelihood for a single observation of $f_{\mathbf{i}}^2$ is

$$l_1\left(f_{\mathbf{i}}^2; p\right) = f_{\mathbf{i}}^2 \log(2p(1-p)) + (1-f_{\mathbf{i}}^2)\log(1-2p(1-p)),$$

or reparameterized with $\theta = 2p(1-p)$ yields $l_1\left(f_{\mathbf{i}}^2; \theta\right) = f_{\mathbf{i}}^2 \log(\theta) + (1-f_{\mathbf{i}}^2)\log(1-\theta)$.

Here we define the MLE θ as $\widehat{\theta}_n^{\text{mle}}(f_{\mathbf{i}}^2)$ and its I.F. for the $\mathbf{i}$th observation of $f_{\mathbf{i}}^2$:

$$\widehat{\theta}_n^{\text{mle}}(f_{\mathbf{i}}^2) = \binom{n}{2}^{-1} \sum_{\mathbf{i} \in C_2^n} f_{\mathbf{i}}^2, \ \varphi_{\theta_0}^{\text{mle}}(f_{\mathbf{i}}; \theta_0) = f_{\mathbf{i}}^2 - \ \theta_0.$$

To obtain its asymptotic properties, we adopt the theory of U-statistics,

$$\begin{aligned} n^{1/2}\left(\widehat{\theta}_n^{\text{mle}}-\theta_0\right) &= \sqrt{n}\binom{n}{2}^{-1}\sum_{\mathbf{i}\in C_2^n}\varphi_{\theta_0}^{\text{mle}}\left(f_{\mathbf{i}}^2;\theta_0\right) \\ &= \sqrt{n}\frac{1}{n}\sum_{i=1}^{n}2E\left(\varphi_{\theta_0}^{\text{mle}}\left(f_{\mathbf{i}}^2;\theta_0\right)\mid Z_i\right)+o_p(1) \\ &\to_d N(\mathbf{0},4\Sigma),\ \Sigma = E\left[E\left(\varphi_{\theta_0}^{\text{mle}}\left(f_{\mathbf{i}}^2;\theta_0\right)\mid Z_i\right)^2\right], \end{aligned}$$

where

$$\begin{aligned} E\left(\varphi_{\theta_0}^{\text{mle}}\left(f_{\mathbf{i}}^2;\theta_0\right)\mid Z_i\right) &= E\left[\left((Z_i-Z_j)^2-\ \theta_0\right)\mid Z_i\right] \\ &= (Z_i^2-2p_0Z_i+p_0)-\ \theta_0. \end{aligned}$$

We thus have $\Sigma=\frac{1}{2}\theta_0(1-2\theta_0)$, and the asymptotic variance of $\varphi_{\theta_0}^{\text{mle}}\left(f_{\mathbf{i}}^2;\theta_0\right)$ is

$$\Sigma_{\theta_0}^{\text{mle}}(f_{\mathbf{i}}^2)=\frac{1}{n}4\Sigma=\frac{1}{n}2\theta_0(1-2\theta_0), \tag{6.22}$$

which is the same as that obtained from MLE through Z_i in (6.21). Hence, there is no loss of information in estimating θ using the between-subject attrubites $f_{\mathbf{i}}$.

3. Semiparametric UGEE of $\theta=2p(1-p)$ through $f_{\mathbf{i}}$

Since $f_{\mathbf{i}}^2\sim Bern(\theta)$, we can construct the FRM with $E\left(f_{\mathbf{i}}^2\right)=\theta$ and let

$$S_{\mathbf{i}}=f_{\mathbf{i}}^2-\theta,\ \ D_{\mathbf{i}}=\frac{\partial}{\partial\theta}\theta,\quad V_{\mathbf{i}}=Var\left(f_{\mathbf{i}}^2\right)=\theta(1-\theta),\quad \mathbf{i}=(i,j)\in C_2^n,$$

then define the UGEE estimator of θ as the solution to the estimating equation:

$$U_n(\theta)=\sum_{\mathbf{i}\in C_2^n}U_{n,\mathbf{i}}(\theta)=\sum_{\mathbf{i}\in C_2^n}D_{\mathbf{i}}V_{\mathbf{i}}^{-1}S_{\mathbf{i}}=\sum_{\mathbf{i}\in C_2^n}\frac{1}{\theta(1-\theta)}\left(f_{\mathbf{i}}^2-\theta\right)=0,$$

i.e. we obtain UGEE estimator $\widehat{\theta}_n^{\text{ugee}}$ and its I.F. for the $\mathbf{i}$th observation $f_{\mathbf{i}}^2$:

$$\widehat{\theta}_n^{\text{ugee}}(f_{\mathbf{i}}^2) = \binom{n}{2}^{-1} \sum_{\mathbf{i}\in C_2^n} f_{\mathbf{i}}^2,\ \varphi_{\theta_0}^{\text{ugee}}(f_{\mathbf{i}};\theta_0) = f_{\mathbf{i}}^2 - \theta_0.$$

Its asymptotic properties are again obtained by the theory of U-statistics,

$$\begin{aligned} n^{1/2}\left(\widehat{\theta}_n^{\text{ugee}} - \theta_0\right) &= \sqrt{n}\frac{1}{n}\sum_{i=1}^{n} 2E\left(\varphi_{\theta_0}^{\text{ugee}}\left(f_{\mathbf{i}}^2;\theta_0\right) \mid Z_i\right) + o_p(1) \\ &\to_d N(\mathbf{0}, 4\Sigma), \end{aligned}$$

where we also have the asymptotic variance of $\varphi_{\theta_0}^{\text{ugee}}\left(f_{\mathbf{i}}^2;\theta_0\right)$ as

$$\Sigma_{\theta_0}^{\text{ugee}}(f_{\mathbf{i}}^2) = 4\Sigma = \frac{1}{n} 2\theta_0(1-2\theta_0). \tag{6.23}$$

which is the same as that obtained from MLE through Z_i in (6.21) and that of MLE through $f_{\mathbf{i}}$ in (6.22), i.e., this semiparametric UGEE estimator for binary response reaches the efficiency bound for estimating θ.

Count Responses

Example 3. **Count Responses**

Consider an exogenous $f_{\mathbf{i}}$ with $f_{\mathbf{i}} = Z_i - Z_j$, where $Z_i \sim^{i.i.d} Pois(\lambda)$.

1. MLE of λ through Z_i

The MLE of λ can be estimated via maximizing the log-likelihood $l_n(z_i;\lambda)$ of the within-subject attributes Z_i, which yields the MLE $\widehat{\lambda}_n^{\text{mle}}(Z_i) = (1/n)\sum_{i=1}^n Z_i$ that is AL with

$$n^{1/2}\left(\widehat{\lambda}_n^{\text{mle}}(Z_i) - \lambda_0\right) = n^{-1/2}\sum_{i=1}^{n}(Z_i - \lambda_0) + o_p(1),\ \varphi_{\lambda_0}^{\text{mle}}(Z_i;\lambda_0) = Z_i - \lambda_0.$$

CLT yields $n^{1/2}\left(\widehat{\lambda}_n^{\text{mle}}(Z_i) - \lambda_0\right) \longrightarrow^d N\left(0, \Sigma_{\lambda_0}^{\text{mle}} = E_{\lambda_0}\left(\varphi_{\lambda_0}^{\text{mle}}\varphi_{\lambda_0}^{\text{mle}}\right)\right)$. Thus, $\Sigma_{\lambda_0}^{\text{mle}} = (1/n)\lambda_0$ is the variance of MLE $\widehat{\lambda}_n^{\text{mle}}(Z_i)$, hence the CR bound for estimating λ.

Suppose we are interest in estimating $\theta = 2\lambda$, then the variance its MLE $\widehat{\theta}_n^{\text{mle}}(Z_i) =$

$(2/n)\sum_{i=1}^{n} Z_i$ is

$$\Sigma_{\theta_0}^{\text{mle}} = \frac{1}{n} 4\lambda_0. \tag{6.24}$$

2. MLE of $\theta = 2\lambda$ through $g_{\mathbf{i}} = Z_i + Z_j$

Here with $\mathbf{i} = (i, j) \in C_2^n$, $g_{\mathbf{i}} = Z_i + Z_j \sim Poisson(\theta)$, the log-likelihood for a single observation $g_{\mathbf{i}}$ is

$$l_1(g_{\mathbf{i}}; \theta) = \log(\theta) g_{\mathbf{i}} - \theta - \log(g_{\mathbf{i}}!).$$

Here we define the MLE θ as $\widehat{\theta}_n^{\text{mle}}(g_{\mathbf{i}})$ and its I.F. for the $\mathbf{i}$th observation $g_{\mathbf{i}}$:

$$\widehat{\theta}_n^{\text{mle}}(g_{\mathbf{i}}) = \binom{n}{2}^{-1} \sum_{\mathbf{i} \in C_2^n} g_{\mathbf{i}}, \ \varphi_{\theta_0}^{\text{mle}}(f_{\mathbf{i}}; \theta_0) = g_{\mathbf{i}} - \theta_0.$$

The asymptotic properties of $\widehat{\theta}_n^{\text{mle}}(g_{\mathbf{i}})$ can be obtained via the theory of U-statistics,

$$\begin{aligned} n^{1/2}\left(\widehat{\theta}_n^{\text{mle}} - \theta_0\right) &= \sqrt{n}\frac{1}{n}\sum_{i=1}^{n} 2E\left(\varphi_{\theta_0}^{\text{mle}}(g_{\mathbf{i}}; \theta_0) \mid Z_i\right) + o_p(1) \\ &\to_d N(\mathbf{0}, 4\Sigma), \ \Sigma = Var\left(E\left(\varphi_{\theta_0}^{\text{mle}}(g_{\mathbf{i}}; \theta_0) \mid Z_i\right)\right). \end{aligned}$$

By the construction of $g_{\mathbf{i}} = Z_i + Z_j$,

$$E\left(\varphi_{\theta_0}^{\text{mle}}(g_{\mathbf{i}}; \theta_0) \mid Z_i\right) = E\left([(Z_i + Z_j) - \theta_0] \mid Z_i\right) = Z_i - \lambda_0,$$

Σ simplifies to

$$\Sigma = E\left[E\left(\varphi_{\theta_0}^{\text{mle}}(g_{\mathbf{i}}; \theta_0) \mid Z_i\right)^2\right] = E\left[(Z_i - \lambda_0)^2\right] = \lambda_0.$$

Thus, the asymptotic variance of $\varphi_{\theta_0}^{\text{mle}}(g_{\mathbf{i}}; \theta_0)$ is

$$\Sigma_{\theta_0}^{mle}(g_{\mathbf{i}}) = \frac{1}{n} 4\lambda_0 \tag{6.25}$$

which is the same as that obtained from MLE through Z_i in (6.24), indicating no loss of information in estimating θ using between-subject attrubites $f_{\mathbf{i}}$.

3. Semiparametric UGEE of $\theta = 2\lambda$ through $g_{\mathbf{i}} = Z_i + Z_j$

With $g_{\mathbf{i}} \sim Poisson(\theta)$, we construct the FRM $E(g_{\mathbf{i}}) = \theta$ and let

$$S_{\mathbf{i}} = g_{\mathbf{i}} - \theta, \;\; D_{\mathbf{i}} = \frac{\partial}{\partial \theta}\theta, \quad V_{\mathbf{i}} = Var(g_{\mathbf{i}}) = \theta, \quad \mathbf{i} = (i,j) \in C_2^n.$$

Define the UGEE estimator of $\theta = 2\lambda$ as the solution to the estimating equation

$$U_n(\theta) = \sum_{\mathbf{i} \in C_2^n} U_{n,\mathbf{i}} = \sum_{\mathbf{i} \in C_2^n} D_{\mathbf{i}} V_{\mathbf{i}}^{-1} S_{\mathbf{i}} = \sum_{\mathbf{i} \in C_2^n} \frac{1}{\theta}(g_{\mathbf{i}} - \theta) = 0,$$

then the UGEE estimator $\widehat{\theta}_n^{\text{ugee}}$ and its I.F. for the $\mathbf{i}$th observation $g_{\mathbf{i}}$ are

$$\widehat{\theta}_n^{\text{ugee}}(g_{\mathbf{i}}) = \binom{n}{2}^{-1} \sum_{\mathbf{i} \in C_2^n} g_{\mathbf{i}}, \; \varphi_{\theta_0}^{\text{ugee}}(g_{\mathbf{i}}; \theta_0) = g_{\mathbf{i}} - \theta_0.$$

The asymptotic properties of $\widehat{\theta}_n^{\text{ugee}}(g_{\mathbf{i}})$ are again obtained by the theory of U-statistics,

$$\begin{aligned} n^{1/2}\left(\widehat{\theta}_n^{\text{ugee}} - \theta_0\right) &= \sqrt{n}\frac{1}{n}\sum_{i=1}^{n} 2E\left(\varphi_{\theta_0}^{\text{ugee}}(g_{\mathbf{i}}; \theta_0) \mid Z_i\right) + o_p(1) \\ &\to_d N(\mathbf{0}, 4\Sigma), \; \Sigma = Var\left(E\left(\varphi_{\theta_0}^{\text{ugee}}(g_{\mathbf{i}}; \theta_0) \mid Z_i\right)\right). \end{aligned}$$

With $E\left(\varphi_{\theta_0}^{\text{ugee}}(g_{\mathbf{i}}; \theta_0) \mid Z_i\right) = Z_i - \lambda_0$, the asymptotic variance of $\varphi_{\theta_0}^{\text{ugee}}(g_{\mathbf{i}}; \theta_0)$ is also

$$\Sigma_{\theta_0}^{\text{ugee}}(g_{\mathbf{i}}) = \frac{1}{n} 4\lambda_0, \tag{6.26}$$

which is the same as that obtained from MLE through Z_i in (6.24), and that obtained from MLE through $g_{\mathbf{i}}$ in (6.25). Hence, this semiparametric UGEE estimator for count response reaches the efficiency bound for estimating θ.

Supplemental Table 1

Method	Assumption	β_0			β_1		
		$n = 100$					
		Est.	Variance		Est.	Variance	
			Asy.	Emp.		Asy.	Emp.
MLE	NB	3.0008	0.0051	0.0058	2.9992	0.0143	0.0160
	Pois	2.9995	0.0008	0.0087	3.0010	0.0014	0.0242
			Asy.	Emp.		Asy.	Emp.
GEE	NB	3.0008	0.0051	0.0058	2.9993	0.0142	0.0160
	Pois	2.9995	0.0080	0.0083	3.0010	0.0213	0.0224
		$n = 300$					
		Est.	Variance		Est.	Variance	
			Asy.	Emp.		Asy.	Emp.
MLE	NB	2.9977	0.0017	0.0017	3.0035	0.0047	0.0048
	Pois	2.9990	0.0003	0.0025	3.0012	0.0005	0.0069
			Asy.	Emp.		Asy.	Emp.
GEE	NB	2.9977	0.0017	0.0017	3.0034	0.0047	0.0048
	Pois	2.9990	0.0027	0.0025	3.0012	0.0072	0.0069
		$n = 500$					
		Est.	Variance		Est.	Variance	
			Asy.	Emp.		Asy.	Emp.
MLE	NB	3.0004	0.0010	0.0010	2.9992	0.0028	0.0028
	Pois	2.9993	0.0002	0.0016	3.0009	0.0003	0.0044
			Asy.	Emp.		Asy.	Emp.
GEE	NB	3.0004	0.0010	0.0010	2.9992	0.0028	0.0028
	Pois	2.9993	0.0016	0.0016	3.0009	0.0043	0.0044

Bibliography

Agresti, A. (2003a). *Categorical data analysis*, volume 482. John Wiley & Sons.

Agresti, A. (2003b). *Categorical data analysis*, volume 482. John Wiley & Sons.

Aitchison, J. (1989). Measures of location of compositional data sets. *Mathematical Geology* **21,** 787–790.

Aldous, J. L., Pond, S. K., Poon, A., Jain, S., Qin, H., Kahn, J. S., Kitahata, M., Rodriguez, B., Dennis, A. M., Boswell, S. L., et al. (2012). Characterizing hiv transmission networks across the united states. *Clinical Infectious Diseases* **55,** 1135–1143.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)* **57,** 289–300.

Bickel, P. J. (1982). On adaptive estimation. *The Annals of Statistics* pages 647–671.

Bickel, P. J., Ritov, Y., and Ryden, T. (1998). Asymptotic normality of the maximum-likelihood estimator for general hidden markov models. *The Annals of Statistics* **26,** 1614–1635.

Carabotti, M., Scirocco, A., Maselli, M. A., and Severi, C. (2015). The gut-brain axis: interactions between enteric microbiota, central and enteric nervous systems. *Annals of gastroenterology: quarterly publication of the Hellenic Society of Gastroenterology* **28,** 203.

Chabris, C. F., Lee, J. J., Benjamin, D. J., Beauchamp, J. P., Glaeser, E. L., Borst, G., Pinker, S., and Laibson, D. I. (2013). Why it is hard to find genes associated with social science traits: Theoretical and empirical considerations. *American journal of public health* **103,** S152–S166.

Chen, R., Chen, T., Lu, N., Zhang, H., Wu, P., Feng, C., and Tu, X. (2014). Extending the mann–whitney–wilcoxon rank sum test to longitudinal regression analysis. *Journal of Applied Statistics* **41,** 2658–2675.

Chen, T., Kowalski, J., Chen, R., Wu, P., Zhang, H., Feng, C., and Tu, X. M. (2016). Rank-preserving regression: a more robust rank regression model against outliers. *Statistics in medicine* **35,** 3333–3346.

Cho, I. and Blaser, M. J. (2012). The human microbiome: at the interface of health and disease. *Nature Reviews Genetics* **13,** 260–270.

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.

Dubitzky, W., Wolkenhauer, O., Yokota, H., and Cho, K.-H. (2013). *Encyclopedia of systems biology*. Springer Publishing Company, Incorporated.

Durack, J. and Lynch, S. V. (2019a). The gut microbiome: relationships with disease and opportunities for therapy. *Journal of experimental medicine* **216,** 20–40.

Durack, J. and Lynch, S. V. (2019b). The gut microbiome: Relationships with disease and opportunities for therapy. *Journal of Experimental Medicine* **216,** 20–40.

Eddelbuettel, D., François, R., Allaire, J., Ushey, K., Kou, Q., Russel, N., Chambers, J., and Bates, D. (2011). Rcpp: Seamless r and c++ integration. *Journal of statistical software* **40,** 1–18.

Goslee, S. C. and Urban, D. L. (2007). The ecodist package for dissimilarity-based analysis of ecological data. *Journal of Statistical Software* **22,** 1–19.

Hájek, J. (1968). Asymptotic normality of simple linear rank statistics under alternatives. *The Annals of Mathematical Statistics* pages 325–346.

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of the american statistical association* **69,** 383–393.

Hemerik, J., Goeman, J. J., et al. (2018). False discovery proportion estimation by permutations: confidence for significance analysis of microarrays. *Journal of the Royal Statistical Society Series B* **80,** 137–155.

Hoeffding, W. and Robbins, H. (1948). The central limit theorem for dependent random variables. *Duke Mathematical Journal* **15,** 773–780.

Hoff, P. D. and Niu, X. (2012). A covariance regression model. *Statistica Sinica* pages 729–753.

Holmes, E., Li, J. V., Athanasiou, T., Ashrafian, H., and Nicholson, J. K. (2011). Understanding the role of gut microbiome–host metabolic signal disruption in health and disease. *Trends in*

Microbiology **19,** 349–359.

Ichimura, H. and Newey, W. K. (2015). The influence function of semiparametric estimators. *arXiv preprint arXiv:1508.01378* .

Kennedy, P. (2003). *A guide to econometrics*. MIT press.

Kowalski, J., Hao, S., Chen, T., Liang, Y., Liu, J., Ge, L., Feng, C., and Tu, X. (2018). Modern variable selection for longitudinal semi-parametric models with missing data. *Journal of Applied Statistics* **45,** 2548–2562.

Kowalski, J. and Tu, X. M. (2008a). *Modern applied U-statistics*, volume 714. John Wiley & Sons.

Kowalski, J. and Tu, X. M. (2008b). *Modern applied U-statistics*, volume 714. John Wiley & Sons.

Kruskal, J. B. and Wish, M. (1978). Multidimensional scaling (quantitative applications in the social sciences). *Beverly Hills* .

Lang, S., Duan, Y., Liu, J., Torralba, M. G., Kuelbs, C., Ventura-Cots, et al. (2020). Human intestinal mycobiome in alcoholic liver disease targeted loci. https://www.ncbi.nlm.nih.gov/bioproject/PRJNA517994.

Lang, S., Duan, Y., Liu, J., Torralba, M. G., Kuelbs, C., Ventura-Cots, M., Abraldes, J. G., Bosques-Padilla, F., Verna, E. C., Brown Jr, R. S., et al. (2020). Intestinal fungal dysbiosis and systemic immune response to fungi in patients with alcoholic hepatitis. *Hepatology* **71,** 522–538.

LeCam, L. (1953). On some asymptotic properties of maximum likelihood estimates and related bayes estimates. *Univ. California Pub. Statist.* **1,** 277–330.

Legendre, P. and Gallagher, E. D. (2001). Ecologically meaningful transformations for ordination of species data. *Oecologia* **129,** 271–280.

Li, X., Kane, M., Zhang, Y., Sun, W., Song, Y., Dong, S., Lin, Q., Zhu, Q., Jiang, F., Zhao, H., et al. (2021). Circadian rhythm analysis using wearable device data: Novel penalized machine learning approach. *Journal of Medical Internet Research* **23,** e18403.

Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73,** 13–22.

Lichstein, J. W. (2007). Multiple regression on distance matrices: a multivariate spatial analysis

tool. *Plant Ecology* **188,** 117–131.

Lin, T., Chen, T., Liu, J., and Tu, X. M. (2021). Extending the mann-whitney-wilcoxon rank sum test to survey data for comparing mean ranks. *Statistics in Medicine* **40,** 1705–1717.

Lin, Z., Kong, D., and Wang, L. (2021). Causal inference on non-linear spaces: Distribution functions and beyond. *arXiv e-prints* pages arXiv–2101.

Liu, J., Zhang, X., Chen, T., Wu, T., Lin, T., Jiang, L., Lang, S., Liu, L., Natarajan, L., Tu, J., et al. (2021). A semiparametric model for between-subject attributes: Applications to beta-diversity of microbiome data. *Biometrics* .

Lozupone, C. and Knight, R. (2005). Unifrac: a new phylogenetic method for comparing microbial communities. *Applied and environmental microbiology* **71,** 8228–8235.

Lozupone, C., Lladser, M. E., Knights, D., Stombaugh, J., and Knight, R. (2011). Unifrac: an effective distance metric for microbial community comparison. *The ISME journal* **5,** 169–172.

Lu, N., Chen, T., Wu, P., Gunzler, D., Zhang, H., He, H., and Tu, X. (2014). Functional response models for intraclass correlation coefficients. *Journal of Applied Statistics* **41,** 2539–2556.

Luenberger, D. G. (1997). *Optimization by vector space methods*. John Wiley & Sons.

Mantel, N. (1967). The detection of disease clustering and a generalized regression approach. *Cancer research* **27,** 209–220.

Martinato, M., Lorenzoni, G., Zanchi, T., Bergamin, A., Buratin, A., Azzolina, D., Gregori, D., et al. (2021). Usability and accuracy of a smartwatch for the assessment of physical activity in the elderly population: Observational study. *JMIR mHealth and uHealth* **9,** e20966.

McArdle, B. H. and Anderson, M. J. (2001). Fitting multivariate models to community data: a comment on distance-based redundancy analysis. *Ecology* **82,** 290–297.

McGraw, K. O. and Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological methods* **1,** 30.

Moon, K. R., van Dijk, D., Wang, Z., Chen, W., Hirn, M. J., Coifman, R. R., Ivanova, N. B., Wolf, G., and Krishnaswamy, S. (2017). Phate: a dimensionality reduction method for visualizing trajectory structures in high-dimensional biological data. *BioRxiv* page 120378.

Morton, J. T., Marotz, C., Washburne, A., Silverman, J., Zaramela, L. S., Edlund, A., et al. (2019). Establishing microbial composition measurement standards with reference frames.

Nature Communications **10,** 1–11.

National Academies of Sciences, E. and Medicine (2018). *Environmental Chemicals, the Human Microbiome, and Health Risk: A Research Strategy*. The National Academies Press, Washington, DC.

Nelder, J. A. and Wedderburn, R. W. (1972). Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)* **135,** 370–384.

Newey, W. K. (1990). Semiparametric efficiency bounds. *Journal of applied econometrics* **5,** 99–135.

Nguyen, T. T., Kosciolek, T., Maldonado, Y., Daly, R. E., Martin, A. S., et al. (2019). Differences in gut microbiome composition between persons with chronic schizophrenia and healthy comparison subjects. *Schizophrenia research* **204,** 23–29.

Nguyen, T. T., Zhang, X., Wu, T.-C., Liu, J., Le, C., Tu, X. M., Knight, R., and Jeste, D. V. (2021a). Association of loneliness and wisdom with gut microbial diversity and composition: An exploratory study. *Frontiers in Psychiatry* **12,** 395.

Nguyen, T. T., Zhang, X., Wu, T.-C., Liu, J., Le, C., Tu, X. M., Knight, R., and Jeste, D. V. (2021b). Association of loneliness and wisdom with gut microbial diversity and composition: An exploratory study. *Frontiers in psychiatry* **12,** 395.

Oksanen, J., Blanchet, F. G., Kindt, R., Legendre, P., Minchin, P. R., O'hara, R., et al. (2013). Package 'vegan'. *Community Ecology Package, Version* **2,** 1–295.

Perezgonzalez, J. D. (2015). Fisher, neyman-pearson or nhst? a tutorial for teaching data testing. *Frontiers in psychology* page 223.

R Development Core Team (2012). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0, http://www.R-project.org.

Roberts, D. W. (2017). Distance, dissimilarity, and mean–variance ratios in ordination. *Methods in Ecology and Evolution* **8,** 1398–1407.

Robinson, P. M. (1988). Root-n-consistent semiparametric regression. *Econometrica: Journal of the Econometric Society* pages 931–954.

Sehgal, V., Singh, S., and Smithson, R. (1987). Nearest points and some fixed point theorems for weakly compact sets. *Journal of mathematical analysis and applications* **128,** 108–111.

Shrout, P. E. and Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin* **86,** 420.

Sklar, M. (1959). Fonctions de repartition an dimensions et leurs marges. *Publ. inst. statist. univ. Paris* **8,** 229–231.

Sørensen, T. J. (1948). *A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons*. I kommission hos E. Munksgaard.

Tang, W., He, H., and Tu, X. M. (2012a). *Applied categorical and count data analysis*. CRC Press.

Tang, W., He, H., and Tu, X. M. (2012b). *Applied categorical and count data analysis*. CRC Press.

Tchetgen, E. J. T. and Shpitser, I. (2012). Semiparametric theory for causal mediation analysis: efficiency bounds, multiple robustness, and sensitivity analysis. *Annals of statistics* **40,** 1816.

Team, R. C. (2017). R core team (2017). r: A language and environment for statistical computing. *R Found. Stat. Comput. Vienna, Austria* .

Thas, O., Neve, J. D., Clement, L., and Ottoy, J.-P. (2012). Probabilistic index models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **74,** 623–671.

Tibshirani, R. (2011). Regression shrinkage and selection via the lasso: a retrospective. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **73,** 273–282.

Tsiatis, A. (2007). *Semiparametric theory and missing data*. Springer Science & Business Media.

Tsiatis, A. A. (2006). Semiparametric theory and missing data.

Tsiatis, A. A. and Ma, Y. (2004). Locally efficient semiparametric estimators for functional measurement error models. *Biometrika* **91,** 835–848.

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.

Tvedebrink, T. (2010). Overdispersion in allelic counts and θ-correction in forensic genetics. *Theoretical Population Biology* **78,** 200–210.

Van der Vaart, A. W. (2000). *Asymptotic statistics*, volume 3. Cambridge university press.

Walter, R. (1987). Real and complex analysis.

Ware Jr, J. E. and Sherbourne, C. D. (1992). The mos 36-item short-form health survey (sf-36): I. conceptual framework and item selection. *Medical care* pages 473–483.

Wu, P., Gunzler, D., Lu, N., Chen, T., Wymen, P., and Tu, X. M. (2014a). Causal inference for community-based multi-layered intervention study. *Statistics in Medicine* **33,** 3905–3918.

Wu, P., Gunzler, D., Lu, N., Chen, T., Wymen, P., and Tu, X. M. (2014b). Causal inference for community-based multi-layered intervention study. *Statistics in medicine* **33,** 3905–3918.

Zhang, Y., Zhou, H., Zhou, J., and Sun, W. (2017). Regression models for multivariate count data. *Journal of Computational and Graphical Statistics* **26,** 1–13.