UC San Diego UC San Diego Electronic Theses and Dissertations

Title

Statistical Methods for Analyzing Between-Subject Attributes in Biomedical Research

Permalink

https://escholarship.org/uc/item/5xv4q00h

Author

Wu, Tsung-Chin

Publication Date

2024

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA SAN DIEGO

Statistical Methods for Analyzing Between-Subject Attributes in Biomedical Research

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

in

Biostatistics

by

Tsung-Chin Wu

Committee in charge:

Professor Xin Tu, Co-Chair Professor Xinlian Zhang, Co-Chair Professor Victor De Gruttola Professor Ravi Goyal Professor Bernd Schnabl

Copyright Tsung-Chin Wu, 2024 All rights reserved. The Dissertation of Tsung-Chin Wu is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2024

DEDICATION

This dissertation is dedicated to my family, whose unconditional support and encouragement have been my guiding light throughout this journey.

To my parents, for their endless love and belief in my potential; to my brother, for his silent dedication and our occasional interactions; to my wife, whose patience and understanding have been my greatest comfort; and to my son, whose smile is my greatest motivation to move forward.

I also dedicate this work to the memory of my grandparents-my grandfather, who ignited my love for mathematics, and my grandmother, who gave me the most profound and unique love in this world-whose wisdom and legacy continue to guide me.

EPIGRAPH

Statistics is the grammar of science.

Karl Pearson

The essence of statistics is to uncover simple truths from the complex and derive clear conclusions from the obscure.

Morris Wu

Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write.

H.G. Wells

Disserta	tion Approval Page	iii
Dedicati	on	iv
Epigraph	n	v
Table of	Contents	vi
List of F	igures	ix
List of T	àbles	X
Acknow	ledgements	xi
Vita		xvi
Abstract	of the Dissertation	xix
Introduc	tion	1
Chapter 1.1 1.2 1.3 1.4 1.5 1.6 1.7	1 Asymptotic Properties of Empirical Area Under Receiver Operating Characteristic Curve with Parameter Estimator of Generalized Linear Models 1.Introduction Preliminaries 1.2.1 Class Probability Model 1.2.2 ROC Curve Detailed Derivations and Theoretical Foundations 1.3.1 Mann-Whitney U-Statistic 1.3.2 Functional Delta Method Main Results of the Current Work Simulation Studies 1.5.1 Simulation Method 1.5.2 Simulation Results Discussion Acknowledgements	4 7 7 9 9 11 14 17 18 19 21 24
Chapter 2.1 2.2 2.3	 Outlier-Robust Random Forest Regression: Application to Between-Subject Attributes Introduction Background 2.2.1 Classical Random Forest Regression 2.2.2 Functional Response Models 2.2.3 Outlier-Robust Response Function Outlier-Robust Random Forest Regression for Between-Subject Attributes 	25 25 28 28 29 29 29

TABLE OF CONTENTS

2.4	Asymptotic Theory	32
	2.4.1 Preliminary Conditions	32
	2.4.2 Consistency	33
	2.4.3 Asymptotic Normality	35
2.5	Simulation Studies	36
	2.5.1 Comparison of Random Forest Regression and FRM under Parametric	
	Assumption Violations	36
	2.5.2 Assessing the Performance of Outlier-Robust Random Forest Regression	
	in the Presence of Outliers	41
2.6	Discussion	44
2.7	Acknowledgements	46
Chapter	3 Estimating Longitudinal Change in Network Transitivity: Application to	
Chapter	Viral Genetic Linkage Networks	47
31	Introduction	47
3.1	Background	49 49
5.2	3.2.1 Viral Genetic Linkage Networks	49
	3.2.1 Transitivity	50
	3.2.3 Functional Response Models	51
33	Methods	52
5.5	3 3 1 Cross-Sectional Network	52
	3.3.2 Longitudinal Networks	55
34	Simulation Studies	62
5.1	3 4 1 Simulation Method	62
	3.4.2 Simulation Results	66
35	HIV Transmission in San Diego County	68
3.6	Discussion	68
3.7	Acknowledgements	72
Chapter	4 Conclusions and Future Work	73
Append	ix A Relevant Existing Theorems	75
A.1	Stochastic Convergence	75
A.2	Two-Sample U-Statistics	76
A.3	Empirical Processes	77
	A.3.1 Brownian Bridge	78
	A.3.2 Maximum Likelihood Estimators	81
	A.3.3 Functional Delta Method	82
A.4	Trees	83
Annend	x B Proofs of the Theorems	86
R 1	Theorem on Empirical AUROC Asymptotics	86
R 2	Outlier-Robust Random Forest Regression for Retween-Subject Attributes	88
D .2	B 2.1 Asymptotic Unbiasedness	88
		00

	B.2.2	Asymptotic Normality	90
B.3	UGEE	-based FRMs for Longitudinal Networks	107
	B.3.1	Consistency of $\hat{\tau}$	108
	B.3.2	Asymptotic Normality of $\widehat{\tau}$	114
	B.3.3	Asymptotic Efficiency of $\widehat{\tau}$	116
Bibliogr	aphy		117

LIST OF FIGURES

Figure 1.1.	Comparison of the bias between $\hat{\theta}_n(\hat{\beta}_n) - \theta(\beta_0)$ and $\hat{\theta}_n(\beta_0) - \theta(\beta_0)$ across varying sample sizes.	21
Figure 1.2.	Comparison of asymptotic and empirical standard errors for $\widehat{\theta}_n(\widehat{\beta}_n)$ across varying sample sizes.	22
Figure 1.3.	Comparison of asymptotic standard deviations for $\hat{\theta}_n(\hat{\beta}_n)$ and $\hat{\theta}_n(\beta_0)$ across varying sample sizes.	23
Figure 2.1.	Comparison of predicted $P(Y_i \le Y_j X_i, X_j)$ between random forest regression and FRM with correct parametric assumptions.	40
Figure 2.2.	Comparison of predicted $P(Y_i \le Y_j X_i, X_j)$ between random forest regression and FRM under violated parametric assumptions.	41
Figure 2.3.	Comparison of predicted $E[(1 + \exp(- Y_i - Y_j))^{-1} X_i, X_j]$ between ran- dom forest regression with and without outliers in the response	43
Figure 2.4.	Comparison of predicted $P(Y_i \le Y_j X_i, X_j)$ between random forest regression with and without outliers in the response.	44
Figure 3.1.	Types of triads	50
Figure 3.2.	Temporal changes in transitivity for the San Diego molecular epidemiology data, with 95% confidence intervals.	70

LIST OF TABLES

Table 3.1.	Estimated transitivity, standard errors, and Type I errors in simulated cross- sectional networks.	66
Table 3.2.	Estimated transitivity, standard errors, and Type I errors in simulated longi- tudinal networks.	67
Table 3.3.	Power of the omnibus test for determining whether the transitivity changes over time in simulated longitudinal networks.	67
Table 3.4.	Estimated transitivity and asymptotic standard errors for the San Diego molecular epidemiology data.	69

ACKNOWLEDGEMENTS

I would like to extend my deepest gratitude to those who have supported me throughout my PhD journey. First and foremost, I would like to thank both co-chairs of my committee, Professor Xin Tu, and Professor Xinlian Zhang, for their willingness to invest a substantial amount of time providing invaluable guidance, prompt feedback, and continuous encouragement. Through Professor Xin Tu, I had the opportunity to collaborate with many researchers in the fields of biomedical and/or public health, where I learned how to contribute to different projects while working together with others. This experience has equipped me with the skills to apply statistical methods to solve real-world data problems and challenges. Professor Xin Tu also taught me how to develop an analysis plan and how to accumulate experience in selecting appropriate statistical tools for analyzing biomedical and/or public health data. Professor Xin Tu is particularly skilled at integrating different statistical methods with his research interests, which inspired me to learn how to conduct research and expand my own areas of interest. Professor Xinlian Zhang consistently identified key issues in a very short time and provided me with immediate feedback and suggestions, for which I am deeply grateful and impressed. Both professors have given me significant advice, direction, and assistance in my career planning, while also respecting and supporting my decisions. Their support and guidance have greatly contributed to my academic growth.

I am also sincerely grateful to the members of my dissertation committee, Professor Victor De Gruttola, Professor Ravi Goyal, and Professor Bernd Schnabl, for their time, effort, and thoughtful input. Professor Victor De Gruttola and Professor Ravi Goyal have deep expertise in network analysis, which allowed me to continue learning about network analysis even after leaving Taiwan. I am especially thankful to Professor Ravi Goyal for dedicating significant time to guiding me on how to write a rigorous research paper, from acknowledging past research achievements to clearly highlighting the contributions of my work—skills that will be invaluable in my academic writing. I am also grateful to Professor Victor De Gruttola for his uncompromising attention to even the less prominent yet crucial issues, which has taught me the importance of seeking truth in research. Additionally, I appreciate Professor Bernd Schnabl's patience when I sought solutions during my research, enabling me to complete our collaborative paper and inspiring two of the papers included in this dissertation. Their critical evaluations, suggestions, and constructive comments have significantly enhanced the quality of this work.

I would also like to extend my heartfelt thanks to the faculty members of the Biostatistics Department in University of California San Diego, especially Professor Armin Schwartzman, Professor Karen Messer, and Professor Florin Vaida, for their guidance in biostatistical methods, which has become the cornerstone of my work in biostatistics. Professor Loki Natarajan, you provided me with much help, guidance, and direction even before I entered the program. I was also fortunate to become a teaching assistant for one of your courses after joining the program. You taught me how to use my expertise to support students and showed me that allowing students to make mistakes is a necessary part of their learning process. The dedication of these professors to teaching and their guidance in fostering a deep understanding of biostatistical concepts have had a profound impact on my academic development and have prepared me for the research challenges in my doctoral studies.

My sincere thanks go to my colleagues and fellow graduate students in the Biostatistics Department for their companionship and for creating a stimulating and supportive environment. I am particularly thankful to Tyler Vu and Anubhav Singh Sachan, who were willing to be friend with me when I first arrived in the United States and my English was still rudimentary. Junting Ren, Charles Luo, Michael Cheung, and Chenyu Liu, who took over half of the courses with me during my graduate studies, your companionship while I was working through complex and challenging subjects is a major part of my graduate school experience. Tuo Lin, Jinyuan Liu, and Kun Yang, who have been excellent partners in research under Professor Xin Tu, your willingness to generously offer assistance and work together to understand many intriguing and challenging research findings has been invaluable on my research journey. Xinran Wang, Christian Pascual, and Yuwei Cheng, who are my true cohort, I apologize and regret not having had more opportunities to interact with you due to my completing almost all the required courses before you and entering the doctoral program during the outbreak of the pandemic. Despite this, I deeply appreciate the shared experience and wish you all continued success for the remainder of your journey. Additionally, I would like to especially acknowledge Stella Tripp, who has always gone above and beyond to handle all administrative matters, whether big or small, complex or simple.

I would also like to extend my gratitude to every collaborator I had the pleasure of working with as a Graduate Student Researcher. Special thanks are due to Professor Dilip Jeste and Professor Ellen Lee, who are my first collaborators, you introduced me to research in the field of mental health and gave me my first experience using statistical tools to assist collaborators in analyzing data and drawing conclusions. I also appreciate your recommendations during my application for the doctoral program and your financial support. Rebecca Daly, your exceptional data preprocessing skills have spared me from worrying about data quality during my analyses. I extend my deepest respect for the time and effort you have devoted. Professor Bei Gao, thank you for your patience during our collaborative project with Professor Bernd Schnabl, which also served as the motivation for two of the papers in my dissertation. Professor Lisa Eyler, David Adamowicz, and Angelina Van Dyne, thank you for allowing me to apply the statistical knowledge I have learned to the field of psychiatry even after Professor Dilip Jeste's retirement. I also appreciate your excellent use of English, which has helped to illuminate the analyses and conclusions I have worked on. Kevin Nguyen, thank you for your generous and selfless assistance with the cluster computations; without your help, I would not have been able to complete the transitivity paper. On top of that, I want to extend my deep gratitude to Paula Smith, who was always there to assist me with various matters while I was at the Stein Clinical Research Building, and whose constant bright smile made my time there truly enjoyable.

I am indebted to my family for their unconditional love and support. To my parents, Chao-Ming Wu and Chiu-Yen Lee, who have never doubted my abilities or potential and who have always shown concern for me in a way that never felt pressuring throughout this long journey of study, providing me with endless love, patience, support, and encouragement, I owe my deepest gratitude. To my brother, Tsung-Lin Wu, although you haven't provided direct support like Mom and Dad, I know that you have always been there, quietly helping and supporting me. Your two visits to the United States have added unforgettable memories to my otherwise uneventful study abroad experience and allowed me to relax as if returning to our past days together. To my wife, Yeh-Chun Ho, who has always been by my side throughout this long journey, always preparing meals and taking care of everything at home so that I do not have to worry about daily life during my busy schedule, and who has respected every decision I made, treated my stubbornness with patient tolerance, listened to all my unspoken worries and emotions, and provided me with your unconditional love and support. I am deeply grateful to have your selfless companionship at this stage of my life, and I hope that we continue to support each other and face all challenges together in the next chapter of our lives. To my son, Alec Wu, whose untroubled joy, boundless energy, and the most adorable presence in the world have given me the strength to overcome the challenges of this journey.

This dissertation is the culmination of years of hard work, and it would not have been possible without the support of these individuals. To all who have contributed to this achievement, I extend my heartfelt thanks.

Chapter 1, in part is currently being prepared for submission for publication of the material. Wu, Tsung-Chin; Lin, Tuo; Yang, Kun; Schnabl, Bernd; Zhang, Xinlian; Tu, Xin. The dissertation author was the primary investigator and author of this material.

Chapter 2, in part is currently being prepared for submission for publication of the material. Wu, Tsung-Chin; Lin, Tuo; Schnabl, Bernd; Zhang, Xinlian; Tu, Xin. The dissertation author was the primary investigator and author of this material.

Chapter 3, in part is currently being prepared for submission for publication of the material. Wu, Tsung-Chin; Nguyen, Kevin; Lin, Tuo; Liu, Jinyuan; De Gruttola, Victor; Little, Susan J.; Tu, Xin; Goyal, Ravi. The dissertation author was the primary investigator and author of this material. This work was supported in part by the National Institute of Allergy and Infectious Diseases (P30 AI036214, R01 AI147441, and R01 MH-132151) and the James B. Pendleton

Charitable Trust. This manuscript was also supported by the Health Resources and Services Administration (HRSA) of the U.S. Department of Health and Human Services (HHS) as part of award UT8HA33959. The contents are those of the author(s) and do not necessarily represent the official views of, nor an endorsement, by HRSA, HHS or the U.S. Government Health Resources and Services Administration (HRSA).

VITA

2015	Bachelor of Science and Bachelor of Science in Agriculture, National Taiwan University
2015–2016	Republic of China Army
2017	Research Assistant, National Taiwan University
2017-2018	Research Assistant, Academia Sinica
2020	Master of Science, University of California San Diego
2024	Doctor of Philosophy, University of California San Diego

PUBLICATIONS

Lin, T., Chen, R., Liu, J., **Wu, T. C.**, Gui, T. T., Li, Y., Huang, X., Yang, K., Chen, G., Chen, T., Strong, D. R., Messer, K. & Tu, X. M. (2024). On frequency and probability weights: An in-depth look at duelling weights. *International Statistical Review*.

Van Dyne, A., **Wu, T. C.**, Adamowicz, D. H., Lee, E. E., Tu, X. M., & Eyler, L. T. (2024). Longitudinal relationships between BMI and hs-CRP among people with schizophrenia. *Schizophrenia Research*, 271, 337-344.

Adamowicz, D. H., **Wu, T. C.**, Daly, R., Irwin, M. R., Jeste, D. V., Tu, X. M., Eyler, L. T., & Lee, E. E. (2024). Executive functioning trajectories and their prospective association with inflammatory biomarkers in schizophrenia and non-psychiatric comparison participants. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, *128*, Article 110866.

Liu, J., Zhang, X., Lin, T., Chen, R., Zhong, Y., Chen, T., **Wu, T. C.**, Liu, C., Huang, A., Nguyen, T. T., Lee, E. E., Jeste, D. V., & Tu, X. M. (2023). A new paradigm for high-dimensional data: Distance-based semiparametric feature aggregation framework via between-subject attributes. *Scandinavian Journal of Statistics*, *51*(2), 672-696.

Liu, J., Xu, K., **Wu, T. C.**, Yao, L., Nguyen, T. T., Jeste, D., & Zhang, X. (2023). Deciphering the 'gut–brain axis' through microbiome diversity. *General Psychiatry*, *36*(5), Article e101090.

Franz, C. E., Gustavson, D. E., Elman, J. A., Fennema-Notestine, C., Hagler Jr, D. J., Baraff, A., Tu, X. M., **Wu, T. C.**, De Anda, J., Beck, A., Kaufman, J. D., Whitsel, N., Finch, C. E., Chen, J. C., Lyons, M. J., & Kremen, W. S. (2023). Associations between ambient air pollution and cognitive abilities from midlife to early old age: Modification by APOE genotype. *Journal of Alzheimer's Disease*, *93*(1), 193-209.

Borin, J. M., Liu, R., Wang, Y., **Wu, T. C.**, Chopyk, J., Huang, L., Kuo, P., Ghose, C., Meyer, J. R., Tu, X. M., Schnabl, B., & Pride, D. T. (2023). Fecal virome transplantation is sufficient to alter fecal microbiota and drive lean and obese body phenotypes in mice. *Gut Microbes*, *15*(1), Article 2236750.

Franz, C., Tu, X., Qin, Y., Elman, J., Fennema-Notestine, C., De Anda, J., **Wu, T. C.**, & Kremen, W. (2022). Cognitive effects of PM2.5 exposure from mid-life to early old age. *Innovation in Aging*, *6*(Supplement_1), 617-618.

Bu, Y., Kurniawa, J. F., Prince, J., Nguyen, A. K., Ho, B., Sit, N. L., Pham, T., Wu, V. M., Tjhia, B., Shin, A. J., **Wu, T. C.**, Tu, X. M., Rao, R., Coleman, T. P., & Lerman, I. (2022). A flexible adhesive surface electrode array capable of cervical electroneurography during a sequential autonomic stress challenge. *Scientific Reports*, *12*(1), Article 19467.

Treichler, E. B., Palmer, B. W., **Wu, T. C.**, Thomas, M. L., Tu, X. M., Daly, R., Lee, E. E., & Jeste, D. V. (2022). Women and men differ in relative strengths in wisdom profiles: A study of 659 adults across the lifespan. *Frontiers in Psychology*, *12*, Article 769294.

Gao, B.*, **Wu, T. C.***, Lang, S., Jiang, L., Duan, Y., Fouts, D. E., Zhang, X., Tu, X. M., & Schnabl, B. (2022). Machine learning applied to omics datasets predicts mortality in patients with alcoholic hepatitis. *Metabolites*, *12*(1), Article 41. *Co-first author

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

Van Patten, R., Nguyen, T. T., Mahmood, Z., Lee, E. E., Daly, R. E., Palmer, B. W., **Wu, T. C.**, Tu, X., Jeste, D. V., & Twamley, E. W. (2022). Physical and mental health characteristics of 2,962 adults with subjective cognitive complaints. *The International Journal of Aging and Human Development*, *94*(4), 459-477.

Yadlapati, R., Kaizer, A. M., Sikavi, D. R., Greytak, M., Cai, J. X., Carroll, T. L., Gupta, S., Wani, S., Menard-Katcher, P., **Wu, T. C.**, Weissbrod, P., Vahabzadeh-Hagh A. M., Pandolfino J. E., & Chan, W. W. (2022). Distinct clinical physiologic phenotypes of patients with laryngeal symptoms referred for reflux evaluation. *Clinical Gastroenterology and Hepatology*, *20*(4), 776-786.

Lee, E. E., Govind, T., Ramsey, M., **Wu, T. C.**, Daly, R., Liu, J., Tu, X. M., Paulus, M. P., Thomas, M. L., & 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), Article 397.

Liu, J., Zhang, X., Chen, T., Wu, T. C., Lin, T., Jiang, L., Lang, S., Liu, L., Natarajan, L., Tu,

J. X., Kosciolek, T., Morton, J., Nguyen, T. T., Schnabl, B., Knight, R., Feng, C., Zhong, Y., & Tu, X. M. (2021). A semiparametric model for between-subject attributes: Applications to beta-diversity of microbiome data. *Biometrics*, *78*(3), 950-962.

Nguyen, T. T., Zhang, X., **Wu, T. C.**, Liu, J., Le, C., Tu, X. M., Knight, R., & Jeste, D. V. (2021). Association of loneliness and wisdom with gut microbial diversity and composition: An exploratory study. *Frontiers in Psychiatry*, *12*, Article 648475.

Nguyen, T. T., Lee, E. E., Daly, R. E., **Wu, T. C.**, Tang, Y., Tu, X., Van Patten, R., Jeste, D. V., & Palmer, B. W. (2020). Predictors of loneliness by age decade: Study of psychological and environmental factors in 2,843 community-dwelling Americans aged 20-69 years. *The Journal of Clinical Psychiatry*, *81*(6), Article 15111.

Gao, B., Duan, Y., Lang, S., Barupal, D., **Wu, T. C.**, Valdiviez, L., Roberts, B., Choy, Y. Y., Shen, T., Byram, G., Zhang, Y., Fan, S., Wancewicz, B., Shao, Y., Vervier, K., Wang, Y., Zhou, R., Jiang, L., Nath, S., ... & Schnabl, B. (2020). Functional microbiomics reveals alterations of the gut microbiome and host co-metabolism in patients with alcoholic hepatitis. *Hepatology Communications*, *4*(8), 1168-1182.

Van Patten, R., Nguyen, T. T., Mahmood, Z., Lee, E. E., Daly, R. E., Palmer, B. W., **Wu, T. C.**, Tu, X., Jeste, D. V., & Twamley, E. W. (2020). Physical and mental health characteristics of adults with subjective cognitive decline: A study of 3,407 people aged 18-81 years from an MTurk-based US national sample. *bioRxiv*, 2020-04.

Wu, T. C., Zhou, Z., Wang, H., Wang, B., Lin, T., Feng, C., & Tu, X. M. (2020). Advanced machine learning methods in psychiatry: An introduction. *General Psychiatry*, *33*(2), Article e100197.

Treichler, E. B., Glorioso, D., Lee, E. E., **Wu, T. C.**, Tu, X. M., Daly, R., O'Brien C., Smith J. L., & Jeste D. V. (2020). A pragmatic trial of a group intervention in senior housing communities to increase resilience. *International Psychogeriatrics*, *32*(2), 173-182.

Zhou, Z., **Wu, T. C.**, Wang, B., Wang, H., Tu, X. M., & Feng, C. (2020). Machine learning methods in psychiatry: A brief introduction. *General Psychiatry*, *33*(1), Article e100171.

Aftab, A., Lee, E. E., Klaus, F., Daly, R., **Wu, T. C.**, Tu, X., Huege, S., & Jeste, D. V. (2019). Meaning in life and its relationship with physical, mental, and cognitive functioning: A study of 1,042 community-dwelling adults across the lifespan. *The Journal of Clinical Psychiatry*, *81*(1), Article 11357.

Graham, S. A., Jeste, D. V., Lee, E. E., **Wu, T. C.**, Tu, X., Kim, H. C., & Depp, C. A. (2019). Associations between heart rate variability measured with a wrist-worn sensor and older adults' physical function: Observational study. *JMIR mHealth and uHealth*, 7(10), Article e13757.

ABSTRACT OF THE DISSERTATION

Statistical Methods for Analyzing Between-Subject Attributes in Biomedical Research

by

Tsung-Chin Wu

Doctor of Philosophy in Biostatistics

University of California San Diego, 2024

Professor Xin Tu, Co-Chair Professor Xinlian Zhang, Co-Chair

This dissertation develops new and advanced statistical methods to address complex challenges in modeling and inference for between-subject attributes in biomedical research, illustrating their applications with both real and simulated data. The work is presented in three papers, each targeting a specific set of problems within this broad theme.

Chapter 1 examines the asymptotic distribution of the empirical area under the receiver operating characteristic curve (AUROC) when parameters from class probability models are estimated by maximum likelihood estimators. Traditional methods relying on Taylor series expansions are inapplicable due to the non-differentiability of the empirical AUROC. To address this, we leverage empirical processes to derive the asymptotic distribution. We evaluate the asymptotic properties of the empirical AUROC through Monte Carlo simulations. Our findings show that the variability of the empirical AUROC is well-described by the derived asymptotic distribution, even for small sample sizes.

In Chapter 2, we address two critical issues in semiparametric and random forest regression models. First, traditional semiparametric regression models for analyzing functional responses may yield inaccurate estimates if posited parametric assumptions are violated. Second, although random forest regression offers flexibility in modeling complex relationships for functional responses, it is susceptible to outliers just as semiparametric regression for betweensubject attributes is. We propose an outlier-robust nonparametric approach using random forest regression to address both issues simultaneously. Our findings demonstrate that the proposed method offers accurate estimates even when parametric assumptions in semiparametric models are violated and/or outliers are present.

Chapter 3 tackles the statistical challenges of inferring transitivity in viral genetic linkage (VGL) networks, particularly with longitudinal data. Transitivity involves modeling connections among three individuals based on similar viral genetic sequences, which presents challenges for traditional statistical paradigms for within-subject attributes. We present a semiparametric approach leveraging functional response models to estimate transitivity, demonstrating accurate point and interval estimates of transitivity in both cross-sectional and longitudinal VGL networks. Applied to HIV surveillance data from San Diego County, our analysis reveals a significant increase in transitivity over time, indicating shorter delays between infection and diagnosis. These insights are important for assessing effects of HIV intervention and prevention programs and understanding disease dynamics.

Introduction

In modern biomedical research and statistical methodology, the development and evaluation of models that involve multi-subject-based functional responses and account for complex dependencies in such functional responses are essential for advancing our understanding of critical phenomena, ranging from diagnostic accuracy to the dynamics of infectious disease transmission. Across various biomedical and public health research domains, the traditional statistical paradigm for modeling within-subject attributes often falls short in capturing the complex relationships of interest arising in real-world data, particularly when these relationships extend beyond simple within-subject attributes to involving more complex between-subject interactions [1, 2]. This dissertation addresses these challenges through a series of studies that introduce robust and theoretically grounded methodologies for modeling such complexities.

Chapter 1 focuses on estimating the area under the receiver operating characteristic curve (AUROC), a widely used metric for assessing the performance of binary classifiers in diagnostic testing. The popular empirical AUROC estimator arising from modeling binary classifiers using class probability models presents significant statistical challenges for deriving its asymptotic properties. This study leverages empirical processes to develop an approach to derive the asymptotic properties of the empirical AUROC estimator when the parameters of the class probability model are estimated by maximum likelihood estimators (MLEs). The established asymptotic results ensure reliable statistical inference and facilitate applications of AUROC in biomedical and public health research.

Chapter 2 extends the scope of statistical modeling to scenarios where the underlying parametric assumptions of semiparametric regression models for the mean response are violated.

Semiparametric regression models for between-subjects have become increasingly popular, as they address complex functional responses involving multiple subjects to capture relationships of interest in biomedical and public health research [1, 2]. Although robust against distributional assumptions, these models require the knowledge of the correct parametric relationship between the response function and features, resulting in biased estimates and inaccurate inference when the parametric form assumptions are incorrectly specified. Additionally, data containing outliers in the response variable is a common problem for mean-based estimation in regression models [3]. To address these limitations, this study develops an outlier-robust, nonparametric approach by building upon recent asymptotic results for random forest regression, a popular machine learning method for nonparametric regression models for within-subject attributes. This proposed approach significantly improves the resilience of regression estimates to extreme values and thereby provides accurate and reliable estimates for nonparametric regression for between-subject attributes.

Chapter 3 focuses on the analysis of a network-based metric, transitivity, within the context of viral genetic linkage (VGL) networks. These networks are instrumental in understanding the transmission dynamics of infectious diseases and in evaluating the efficacy of public health interventions [4]. However, traditional statistical models are not amenable to capturing the intricate relationships defined by transitivity. By employing semiparametric functional response models (FRMs), this study introduces a novel approach to estimating transitivity and its temporal changes in growing networks over time, thereby facilitating the evaluation of HIV intervention and prevention programs.

Collectively, these studies contribute to the advancement of statistical methodologies for analyzing complex biomedical and public health research data. They underscore the necessity of sophisticated modeling techniques that transcend traditional paradigms, offering new insights and tools for both theoretical research and practical applications in biomedical and public health research.

This dissertation is organized as follows. Chapter 1 derives the asymptotic properties of

the empirical AUROC with the parameters of the class probability model estimated by the MLE. Chapter 2 introduces the outlier-robust random forest regression for between-subject attributes and establishes its asymptotic properties. It also illustrates the approach with simulated data. Chapter 3 develops the FRMs and the asymptotic properties of estimators of transitivity in both cross-sectional and longitudinal VGL networks, and illustrates its application with HIV data from San Diego County, which motivated this methodological research project. We present our conclusions and future work in Chapter 4. Appendix A states the existing theorems used in the first three chapters, and Appendix B provides proofs for the main theorems stated in the first three chapters.

Chapter 1

Asymptotic Properties of Empirical Area Under Receiver Operating Characteristic Curve with Parameter Estimator of Generalized Linear Models

1.1 Introduction

A receiver operating characteristic (ROC) curve is widely used to describe the diagnostic ability of a two-level classifier in biomedical and biostatistical research [5, 6]. It is a probability curve that plots sensitivity (true positive rate, TPR) against 1 - specificity (false positive rate, FPR) at various threshold values. The performance of the classifier can be quantified by the area under the ROC curve (AUROC). The AUROC measures the classifier's ability to distinguish between two classes, such as the presence or absence of a disease of interest. By defining one class as positive and the other as negative, the AUROC provides the probability that a classifier will rank a randomly chosen positive sample higher than a randomly chosen negative sample [7]. A higher AUROC indicates better ranking performance. An AUROC of 1 signifies perfect ranking, whereas an AUROC of 0 indicates that the classifier ranks all samples incorrectly. An AUROC of 0.5 suggests that the classifier performs no better than random guessing. Therefore, achieving an AUROC between 0.5 and 1 represents a fundamental objective in classifier modeling.

In practice, the AUROC of a classifier for a given population of interest is estimated by

calculating the empirical AUROC (see Equation 1.1 in Section 1.2.2 for details) from observed data. This empirical AUROC can be represented as a two-sample *U*-statistic, a nonparametric statistic that imposes no parametric assumptions on the class probability model (see Section 1.2.1 for details) [8, 9, 10]. According to the theory of *U*-statistics [11, 12], the asymptotic normality of the empirical AUROC can be established. However, class probabilities are typically unknown and are required to compute the empirical AUROC. Therefore, they must be estimated from the data. For two-level classification, it is common to model the relationship between subject features and their class labels using parametric models such as generalized linear models (GLMs) [13] with a link function like the logit or probit link for binary responses. Consequently, when deriving the asymptotic properties of the empirical AUROC, it is crucial to consider the predictors of class probabilities, which involves model parameters.

Traditionally, the asymptotic distribution of a statistic with estimated parameters is derived using a Taylor series expansion around the true parameters, involving differentiation of the objective function with respect to the unknown parameters. However, since the indicator functions induce discontinuities in the empirical AUROC, rendering this conventional approach inapplicable because of a differentiability issue. One common method to address this challenge is to approximate the discontinuous objective function with a smooth function. Ma and Huang [14] proposed a model parameter estimator that maximizes a sigmoid approximation to the empirical AUROC. They were able to derive the consistency and asymptotic normality of this parameter estimator due to the continuity of the sigmoid function. Similarly, Heller et al. [15] derived the asymptotic distributions of the difference in AUROC statistics from nested models by smoothing the objective function by approximating the indicator functions with the standard normal distribution. Since the model parameter estimators in these studies are derived directly from the empirical AUROC, rather than the class probability model, these smoothing-approximation-based approaches do not address the issue of AUROC with a plugin estimator from the model of interest, which is the focus of the current study. To perform valid inference, it is crucial to derive the proper asymptotic distribution of the estimator of

AUROC from class probability models, such as when comparing AUROCs between two class models. Inferences about the underlying AUROC, such as p-values and confidence intervals, may be biased without accounting for sampling variability of estimators of parameters from the class probability model, giving rise to potentially misleading conclusions in practice. Hence, understanding the correct asymptotic properties is essential for reliable statistical inference and for making valid comparisons of AUROCs between different class probability models.

In this study, we propose to leverage empirical processes to derive the asymptotic distribution of the empirical AUROC with estimated parameters from class probability models. This approach allows the empirical AUROC to achieve differentiability in Banach space, enabling the application of the functional delta method to establish its asymptotic distribution [11, 16]. Ghoudi and Rémillard [17] laid the groundwork for analyzing the asymptotic behavior of empirical processes constructed from the estimated distribution function of a non-observable random variable $\theta(Q, X)$, which depends on an observable random variable X and its unknown law Q. Their results can be utilized to generalize the Mann-Whitney statistic [18], which is closely related to the empirical AUROC, to the Mann-Whitney process. Van Der Vaart and Wellner [19] extended the results from Ghoudi and Rémillard [17] beyond indicator functions and reformulated them within the context of empirical process theory. These developments have inspired our approach to establish the asymptotic distribution of the empirical AUROC with estimated process.

We focus on the asymptotic properties of the empirical AUROC using parameter estimators derived from GLM-based class probability models. In practice, the parameters in GLM are typically estimated using maximum likelihood estimators (MLEs). Although it is possible to estimate parameters by directly maximizing the empirical AUROC [14, 15], MLEs are among the most commonly used estimators in biomedical research and thus the focus of this work.

This paper is organized as follows. Section 1.2 introduces the two-level class probability model and the probabilistic interpretation of the ROC curve. Section 1.3 reviews the existing asymptotic results for the empirical AUROC with a fixed parameter. Section 1.4 presents

our main findings on the asymptotic properties of the empirical AUROC using the MLE. In Section 1.5, we conduct simulation studies to evaluate the results of the asymptotic theories established in Section 1.4. We present our concluding remarks in the discussion in Section 1.6.

1.2 Preliminaries

1.2.1 Class Probability Model

Consider a two-class classification study where one class comprises of diseased subjects and the other comprises of non-diseased subjects. Suppose there are *n* subjects, each characterized by *d* features. Let (X_i, Y_i) , i = 1, ..., n, denote a set of independent and identically distributed (i.i.d.) samples, where $Y_i \in \{0, 1\}$ represents a binary outcome for the *i*-th subject, with $Y_i = 1$ indicating the presence of the disease and $Y_i = 0$ indicating the absence of the disease. The continuous feature vector for the *i*-th subject is denoted by $\mathbf{X}_i \in \mathbb{R}^d$.

To classify the disease status of each subject based on \mathbf{X}_i , it is essential to model the probability of Y = 1 given \mathbf{X} . Let $p_0 : \mathbb{R}^d \to [0, 1]$ denote the true conditional probability function. We model p_0 using a GLM with $P(Y = 1 | \mathbf{X}) = g^{-1}(\boldsymbol{\beta}^\top \mathbf{X})$, where $g : [0, 1] \to \mathbb{R}$ is an unknown strictly increasing link function, and $\boldsymbol{\beta} \in \mathbb{R}^d$ is a vector of unknown regression parameters. Let $\boldsymbol{\beta}_0$ represent the true parameter vector. We can express $p_0(\mathbf{X})$ as $p_0(\mathbf{X}) = P(Y = 1 | \mathbf{X}) = g^{-1}(\boldsymbol{\beta}_0^\top \mathbf{X})$.

To estimate β_0 , we employ the MLE, denoted by $\hat{\beta}_n$. For convenience, we define $\hat{p}(\mathbf{X}) = g^{-1}(\hat{\beta}_n^{\top} \mathbf{X})$ as the MLE estimator of $p_0(\mathbf{X})$. After modeling p_0 , a cutoff point is selected to classify subjects. For example, using a cutoff value of 0.5, we classify the *i*-th subject as diseased if $\hat{p}(\mathbf{X}_i) > 0.5$ and as non-diseased if $\hat{p}(\mathbf{X}_i) \leq 0.5$.

1.2.2 ROC Curve

An ROC curve can be plotted by examining the TPR and FPR at various cutoff values. For the GLM, the TPR and FPR are defined as

$$\operatorname{TPR}_{g}(c) = P(g^{-1}(\boldsymbol{\beta}^{\mathsf{T}}\mathbf{X}) > c \mid Y = 1)$$

and

$$\operatorname{FPR}_{g}(c) = P(g^{-1}(\boldsymbol{\beta}^{\mathsf{T}}\mathbf{X}) > c \mid Y = 0),$$

respectively, for a cutoff $c \in [0, 1]$. The two-dimensional space formed by the TPR and FPR is referred to as ROC space. By varying c from 1 to 0, the ROC curve $\{(\text{FPR}_g(c), \text{TPR}_g(c)) : 0 \le c \le 1\}$ is plotted in the ROC space. Let $\mathbf{X}_j^H \in \mathbb{R}^d$, $j = 1, ..., n_0$, and $\mathbf{X}_k^D \in \mathbb{R}^d$, $k = 1, ..., n_1$, denote two sequences of i.i.d. random vectors representing feature vectors from the non-diseased and diseased classes, respectively, where n_0 is the number of subjects in the non-diseased class and n_1 is the number of subjects in the diseased class ($n = n_0 + n_1$). The AUROC, denoted by θ , is obtained by integrating the ROC curve between the limits of FPR, 0 and 1:

$$\begin{split} \theta(\boldsymbol{\beta}) &= \int_0^1 \mathrm{TPR}_g(\mathrm{FPR}_g^{-1}(t)) dt = \int_1^0 \mathrm{TPR}_g(c) \,\mathrm{FPR}_g(dc) \\ &= \int_0^1 \int_0^1 I(c' > c) (1 - \mathrm{TPR}_g) (dc') (1 - \mathrm{FPR}_g) (dc) \\ &= P(g^{-1}(\boldsymbol{\beta}^\top \mathbf{X}^D) > g^{-1}(\boldsymbol{\beta}^\top \mathbf{X}^H)) \\ &= P(\boldsymbol{\beta}^\top \mathbf{X}^D > \boldsymbol{\beta}^\top \mathbf{X}^H), \end{split}$$

where $\text{FPR}_g^{-1}(t) := \inf\{c \in [0,1] : \text{FPR}_g(c) \le t\}$ and *I* is the indicator function. Note that common link functions *g* for a binary outcome include the logit and probit functions, but the AUROC is independent of the choice of link functions as long as the link function is strictly increasing.

The empirical AUROC, denoted by $\hat{\theta}_n$, is defined as

$$\widehat{\theta}_{n}(\boldsymbol{\beta}) = \frac{1}{n_{0}n_{1}} \sum_{j=1}^{n_{0}} \sum_{k=1}^{n_{1}} I(\boldsymbol{\beta}^{\mathsf{T}} \mathbf{X}_{k}^{D} > \boldsymbol{\beta}^{\mathsf{T}} \mathbf{X}_{j}^{H}).$$
(1.1)

Note that $\widehat{\theta}_n$ is a two-sample *U*-statistic that is proportional to the Mann-Whitney *U*-statistic, allowing the application of *U*-statistics theory to derive the asymptotic properties of $\widehat{\theta}_n$. The details are provided in the following section.

1.3 Detailed Derivations and Theoretical Foundations

In this section, we provide a complete set of detailed derivation and relevant results of the asymptotic distributions for the empirical AUROC with a fixed parameter β_0 . While the textbooks Van der Vaart [11] and Van der Vaart and Wellner [16] offer the fundamental theories and partial derivations, they do not provide complete and explicit derivations of the relevant results. This section aims to present these results in a comprehensive manner and provide two different approaches for this derivation. Note that the theorems and lemmas in this section are stated in Appendix A.

1.3.1 Mann-Whitney U-Statistic

From Equation (1.1), $\hat{\theta}_n$ represents a two-sample *U*-statistic with the kernel function $h(\boldsymbol{\beta}^{\top} \mathbf{X}^H; \boldsymbol{\beta}^{\top} \mathbf{X}^D) := I(\boldsymbol{\beta}^{\top} \mathbf{X}^D > \boldsymbol{\beta}^{\top} \mathbf{X}^H)$ (see Section A.2). It follows that

$$\theta(\boldsymbol{\beta}) = E\left[I(\boldsymbol{\beta}^{\top}\mathbf{X}^{D} > \boldsymbol{\beta}^{\top}\mathbf{X}^{H})\right],$$

so $\hat{\theta}_n$ is an unbiased estimator of θ for a fixed β . The statistic $n_0 n_1 \hat{\theta}_n$ is commonly referred to as the Mann-Whitney *U*-statistic [11, 12]. For convenience, we denote $\hat{\theta}_n$ as U_n in relation to the *U*-statistic.

For the kernel h, we define

$$h_{1,0}(\boldsymbol{\beta}^{\mathsf{T}}\mathbf{X}^{H}) = E\left[h(\boldsymbol{\beta}^{\mathsf{T}}\mathbf{X}^{H};\boldsymbol{\beta}^{\mathsf{T}}\mathbf{X}^{D}) \mid \boldsymbol{\beta}^{\mathsf{T}}\mathbf{X}^{H}\right],$$
$$h_{0,1}(\boldsymbol{\beta}^{\mathsf{T}}\mathbf{X}^{D}) = E\left[h(\boldsymbol{\beta}^{\mathsf{T}}\mathbf{X}^{H};\boldsymbol{\beta}^{\mathsf{T}}\mathbf{X}^{D}) \mid \boldsymbol{\beta}^{\mathsf{T}}\mathbf{X}^{D}\right]$$

and their centered versions

$$\widetilde{h}_{1,0}(\boldsymbol{\beta}^{\mathsf{T}}\mathbf{X}^{H}) = h_{1,0}(\boldsymbol{\beta}^{\mathsf{T}}\mathbf{X}^{H}) - \theta(\boldsymbol{\beta}),$$
$$\widetilde{h}_{0,1}(\boldsymbol{\beta}^{\mathsf{T}}\mathbf{X}^{D}) = h_{0,1}(\boldsymbol{\beta}^{\mathsf{T}}\mathbf{X}^{D}) - \theta(\boldsymbol{\beta}).$$

Then, the Hájek projection [11] \widehat{U}_n of $U_n - \theta$ onto the set of all functions of the form

$$\sum_{j=1}^{n_0} a_j (\boldsymbol{\beta}^\top \mathbf{X}_j^H) + \sum_{k=1}^{n_1} b_k (\boldsymbol{\beta}^\top \mathbf{X}_k^D),$$

is given by

$$\widehat{U}_n(\boldsymbol{\beta}) = \frac{1}{n_0} \sum_{j=1}^{n_0} \widetilde{h}_{1,0}(\boldsymbol{\beta}^\top \mathbf{X}_j^H) + \frac{1}{n_1} \sum_{k=1}^{n_1} \widetilde{h}_{0,1}(\boldsymbol{\beta}^\top \mathbf{X}_k^D),$$

where a_j and b_k are arbitrary measurable functions with $E\left[a_j^2(\boldsymbol{\beta}^{\mathsf{T}}\mathbf{X}_j^H)\right], E\left[b_k^2(\boldsymbol{\beta}^{\mathsf{T}}\mathbf{X}_k^D)\right] < \infty$. Since $h(\boldsymbol{\beta}^{\mathsf{T}}\mathbf{X}^H, \boldsymbol{\beta}^{\mathsf{T}}\mathbf{X}^D)$ is square-integrable, by Theorem A.4, the difference between \widehat{U}_n

and $U_n - \theta$ is asymptotically negligible. Consequently,

$$\sqrt{n}(\widehat{\theta}_n(\boldsymbol{\beta}) - \theta(\boldsymbol{\beta})) = \sqrt{n}(U_n(\boldsymbol{\beta}) - \theta(\boldsymbol{\beta})) \xrightarrow{d} \mathcal{N}\left(0, \frac{\sigma_H^2}{\lambda_0} + \frac{\sigma_D^2}{\lambda_1}\right), \tag{1.2}$$

where

$$\frac{n_{\ell}}{n} \to \lambda_{\ell} \in (0,1) \quad \text{as} \quad n_0, n_1 \to \infty, \quad \ell \in \{0,1\},$$

$$\sigma_H^2 = \operatorname{Var}\left(\widetilde{h}_{1,0}(\boldsymbol{\beta}^{\mathsf{T}} \mathbf{X}^H)\right), \quad \sigma_D^2 = \operatorname{Var}\left(\widetilde{h}_{0,1}(\boldsymbol{\beta}^{\mathsf{T}} \mathbf{X}^D)\right).$$

It follows from (1.2) that $\hat{\theta}_n$ is a \sqrt{n} -consistent estimator of θ .

1.3.2 Functional Delta Method

Although the empirical AUROC is not differentiable in the traditional calculus sense due to the presence of indicator functions, differentiability in the sense of empirical processes can still be achieved. Without loss of generality, define

$$\operatorname{TPR}(c) = P(\boldsymbol{\beta}^{\top}\mathbf{X} > c \mid Y = 1)$$

and

$$\operatorname{FPR}(c) = P(\boldsymbol{\beta}^{\mathsf{T}} \mathbf{X} > c \mid Y = 0)$$

for a cutoff $c \in \overline{\mathbb{R}} := [-\infty, \infty]$.

Let $F_0(c) = 1 - \text{FPR}(c)$ and $F_1(c) = 1 - \text{TPR}(c)$. Note that

$$\begin{aligned} \theta(\boldsymbol{\beta}) &= \int_{0}^{1} \text{TPR}(\text{FPR}^{-1}(t)) dt = \int_{\infty}^{-\infty} \text{TPR}(c) \text{FPR}(dc) \\ &= \int_{-\infty}^{\infty} (1 - F_1(c)) F_0(dc) \\ &= 1 - \int_{-\infty}^{\infty} F_1(c) F_0(dc) \\ &= F_0(c) F_1(c) \Big|_{-\infty}^{\infty} - \int_{-\infty}^{\infty} F_1(c) F_0(dc) \\ &= \int_{-\infty}^{\infty} F_0(c) F_1(dc) \\ &= \int_{0}^{1} F_0(F_1^{-1}(t)) dt, \end{aligned}$$

where $\text{FPR}^{-1}(t) := \inf\{c \in \mathbb{R} : \text{FPR}(c) \le t\}$ and $F_1^{-1}(t) := \inf\{c \in \mathbb{R} : F_1(c) \ge t\}$. Hence, the ROC curve can be expressed as a process $\{F_0(F_1^{-1}(t)) : t \in [0,1]\}$, and $\theta(\beta)$ can be written as $\phi(F_0, F_1) := \int F_0 dF_1$.

Similarly, let $\mathbb{F}_{n_0}(c) = \frac{1}{n_0} \sum_{j=1}^{n_0} I(\boldsymbol{\beta}^\top \mathbf{X}_j^H \leq c)$ denote the empirical distribution function of $\boldsymbol{\beta}^\top \mathbf{X}_1^H, \dots, \boldsymbol{\beta}^\top \mathbf{X}_{n_0}^H$ with respect to F_0 and $\mathbb{F}_{n_1}(c) = \frac{1}{n_1} \sum_{k=1}^{n_1} I(\boldsymbol{\beta}^\top \mathbf{X}_k^D \leq c)$ denote the empirical distribution function of $\boldsymbol{\beta}^\top \mathbf{X}_1^D, \dots, \boldsymbol{\beta}^\top \mathbf{X}_{n_1}^D$ with respect to F_1 . The empirical ROC curve is then represented as the process

$$\left\{ \mathbb{F}_{n_0}(\mathbb{F}_{n_1}^{-1}(t)) : t \in [0,1] \right\},\$$

where $\mathbb{F}_{n_1}^{-1}(t) := \inf\{c \in \mathbb{R} : \mathbb{F}_1(c) \ge t\}$. Additionally, $\widehat{\theta}_n(\beta)$ can be expressed as

$$\int \mathbb{F}_{n_0}(\mathbb{F}_{n_1}^{-1}(t))dt = \int \mathbb{F}_{n_0}d\mathbb{F}_{n_1} = \phi(\mathbb{F}_{n_0},\mathbb{F}_{n_1}).$$

Let D[a,b] denote the Banach space of all cadlag functions (right continuous with left limits) $z : [a,b] \to \mathbb{R}$ on an interval $[a,b] \subseteq \overline{\mathbb{R}}$, equipped with the uniform norm. Let $BV_M[a,b]$ represent the set of all cadlag functions $z : [a,b] \to [-M,M] \subset \mathbb{R}$ with total variation bounded by *M*. Then, ϕ can be viewed as a map from the domain $BV_1[-\infty, \infty] \times BV_1[-\infty, \infty] \subset D[-\infty, \infty] \times BV_1[-\infty, \infty]$ into \mathbb{R} , equipped with the product norm. From Lemma A.11, it follows that ϕ is Hadamard differentiable [11, 16] at (F_0, F_1) . Specifically, there exists a continuous linear map $\phi'_{(F_0,F_1)} : D[-\infty, \infty] \times BV_1[-\infty, \infty] \to \mathbb{R}$ such that

$$\frac{\phi(F_0 + tH_{0t}, F_1 + tH_{1t}) - \phi(F_0, F_1)}{t} \to \phi'_{(F_0, F_1)}(H_0, H_1) \quad \text{as} \quad t \to 0,$$

for all converging sequences $(H_{0t}, H_{1t}) \rightarrow (H_0, H_1)$ such that

$$(F_0 + tH_{0t}, F_1 + tH_{1t}) \in \mathrm{BV}_1[-\infty, \infty] \times \mathrm{BV}_1[-\infty, \infty].$$

From the same lemma, the derivative $\phi'_{(F_0,F_1)}$ is given by

$$\phi'_{(F_0,F_1)}(H_0,H_1) = \int F_0 dH_1 + \int H_0 dF_1 = F_0 H_1 \Big|_{-\infty}^{\infty} - \int (H_1)_- dF_0 + \int H_0 dF_1,$$

where $(H_1)_{-}$ denotes the left-continuous version of a cadlag function H_1 .

By Donsker's theorem (see Theorem A.7) and Slutsky's theorem (see Theorem A.3),

$$\sqrt{n}(\mathbb{F}_0 - F_0, \mathbb{F}_1 - F_1) \xrightarrow{d} \left(\frac{\mathbb{G}_{F_0}}{\sqrt{\lambda_0}}, \frac{\mathbb{G}_{F_1}}{\sqrt{\lambda_1}} \right),$$

where \mathbb{G}_{F_0} and \mathbb{G}_{F_1} are independent Brownian bridges (see Section A.3.1). It follows that the asymptotic distribution of the empirical AUROC can alternatively be derived using the functional delta method (see Theorem A.10):

$$\begin{split} \sqrt{n}(\theta_n(\boldsymbol{\beta}) - \theta(\boldsymbol{\beta})) &= \sqrt{n}(\phi(\mathbb{F}_{n_0}, \mathbb{F}_{n_1}) - \phi(F_0, F_1)) \\ & \stackrel{d}{\to} \phi'_{(F_0, F_1)}\left(\frac{\mathbb{G}_{F_0}}{\sqrt{\lambda_0}}, \frac{\mathbb{G}_{F_1}}{\sqrt{\lambda_1}}\right) = -\frac{1}{\sqrt{\lambda_1}}\int (\mathbb{G}_{F_1})_{-}dF_0 + \frac{1}{\sqrt{\lambda_0}}\int \mathbb{G}_{F_0}dF_1. \end{split}$$

By Theorem A.8, we have $-\int (\mathbb{G}_{F_1})_{-}dF_0 \sim (\mathbb{G}_{F_1})_{-}(F_0)$ and $\int \mathbb{G}_{F_0}dF_1 \sim -\mathbb{G}_{F_0}(F_1)$, where $\mathbb{G}_G(F)$ denotes a *G*-Brownian bridge process indexed by the function *F*. Thus, we obtain

$$\sqrt{n}(\widehat{\theta}_n(\boldsymbol{\beta}) - \theta(\boldsymbol{\beta})) \xrightarrow{d} \frac{(\mathbb{G}_{F_1})_-(F_0)}{\sqrt{\lambda_1}} - \frac{\mathbb{G}_{F_0}(F_1)}{\sqrt{\lambda_0}}.$$

The limit distribution of $\sqrt{n}(\hat{\theta}_n(\boldsymbol{\beta}) - \theta(\boldsymbol{\beta}))$ can be deduced from its finite-dimensional distributions as follows:

$$\sqrt{n}(\widehat{\theta}_{n}(\boldsymbol{\beta}) - \theta(\boldsymbol{\beta})) \xrightarrow{d} \mathcal{N}\left(0, \frac{\operatorname{Var}\left(F_{0}(\boldsymbol{\beta}^{\top}\mathbf{X}^{D})\right)}{\lambda_{1}} + \frac{\operatorname{Var}\left(F_{1}(\boldsymbol{\beta}^{\top}\mathbf{X}^{H})\right)}{\lambda_{0}}\right).$$
(1.3)

The result (1.3) obtained from the empirical processes argument matches the result shown in (1.2), which is derived using the *U*-statistics theory.

1.4 Main Results of the Current Work

In this study, we extend the existing asymptotic analysis from $\hat{\theta}_n(\beta_0)$, as presented in Section 1.3, to the more complex case of $\hat{\theta}_n(\hat{\beta}_n)$. The key innovation of our work lies in addressing the challenge posed by the sampling variability inherent in the estimator $\hat{\beta}_n$, as opposed to the fixed parameter β_0 . To establish the asymptotic properties of $\hat{\theta}_n(\hat{\beta}_n)$, we begin by demonstrating the consistency of $\hat{\theta}_n(\hat{\beta}_n)$.

Theorem 1.1. If $\widehat{\boldsymbol{\beta}}_n \xrightarrow{p} \boldsymbol{\beta}_0$, then $\widehat{\theta}_n(\widehat{\boldsymbol{\beta}}_n) \xrightarrow{p} \theta(\boldsymbol{\beta}_0)$ as $n \to \infty$.

Proof. From Section 1.3, we have $\widehat{\theta}_n(\beta_0) \xrightarrow{p} \theta(\beta_0)$ as $n \to \infty$. Hence, it suffices to prove that $\widehat{\theta}_n(\widehat{\beta}_n) \xrightarrow{p} \widehat{\theta}_n(\beta_0)$ for any n if $\widehat{\beta}_n \xrightarrow{p} \beta_0$. Note that $\widehat{\theta}_n$ is a step function with at most $n_0 \times n_1$ (countable) discontinuity points, so the discontinuity set of $\widehat{\theta}_n$ has measure zero. Therefore, by the continuous mapping theorem (see Theorem A.1), $\widehat{\theta}_n(\widehat{\beta}_n) \xrightarrow{p} \widehat{\theta}_n(\beta_0)$ for any n if $\widehat{\beta}_n \xrightarrow{p} \beta_0$.

Thus, for any $\epsilon > 0$,

$$P\left(|\widehat{\theta}_{n}(\widehat{\boldsymbol{\beta}}_{n}) - \theta(\boldsymbol{\beta}_{0})| \ge \epsilon\right) \le P\left(|\widehat{\theta}_{n}(\widehat{\boldsymbol{\beta}}_{n}) - \widehat{\theta}_{n}(\boldsymbol{\beta}_{0})| + |\widehat{\theta}_{n}(\boldsymbol{\beta}_{0}) - \theta(\boldsymbol{\beta}_{0})| \ge \epsilon\right)$$
$$\le P\left(|\widehat{\theta}_{n}(\widehat{\boldsymbol{\beta}}_{n}) - \widehat{\theta}_{n}(\boldsymbol{\beta}_{0})| \ge \epsilon/2\right) + P\left(|\widehat{\theta}_{n}(\boldsymbol{\beta}_{0}) - \theta(\boldsymbol{\beta}_{0})| \ge \epsilon/2\right)$$
$$\to 0 \quad \text{as } n \to \infty \text{ and } \widehat{\boldsymbol{\beta}}_{n} \xrightarrow{p} \boldsymbol{\beta}_{0}.$$

From Theorem 1.1, we see that $\widehat{\theta}_n(\widehat{\beta}_n)$ is a consistent estimator of $\theta(\beta_0)$, similar to $\widehat{\theta}_n(\beta_0)$. We next demonstrate the asymptotic normality of $\widehat{\theta}_n(\widehat{\beta}_n)$.

Theorem 1.2. If $\widehat{\beta}_n$ is the MLE for β_0 , then

$$\sqrt{n}(\widehat{\theta}_n(\widehat{\boldsymbol{\beta}}_n) - \theta(\boldsymbol{\beta}_0)) \xrightarrow{d} \mathcal{N}\left(0, \frac{\operatorname{Var}\left(F_0(\boldsymbol{\beta}_0^{\mathsf{T}} \mathbf{X}^D)\right)}{\lambda_1} + \frac{\operatorname{Var}\left(F_1(\boldsymbol{\beta}_0^{\mathsf{T}} \mathbf{X}^H)\right)}{\lambda_0}\right) \quad as \ n \to \infty$$

where $\frac{n_{\ell}}{n} \rightarrow \lambda_{\ell} \in (0, 1)$ as $n_0, n_1 \rightarrow \infty$.

Proof. We begin by decomposing $\sqrt{n}(\widehat{\theta}_n(\widehat{\beta}_n) - \theta(\beta_0))$ into three terms:

$$\sqrt{n}(\widehat{\theta}_n(\widehat{\boldsymbol{\beta}}_n) - \widehat{\theta}_n(\boldsymbol{\beta}_0) - \theta(\widehat{\boldsymbol{\beta}}_n) + \theta(\boldsymbol{\beta}_0)) + \sqrt{n}(\widehat{\theta}_n(\boldsymbol{\beta}_0) - \theta(\boldsymbol{\beta}_0)) + \sqrt{n}(\theta(\widehat{\boldsymbol{\beta}}_n) - \theta(\boldsymbol{\beta}_0)).$$
(1.4)

The second term captures the variability associated with the empirical estimator of AUROC, the third term accounts for the variability introduced by the MLE, and the first term represents the residual.

Focusing on the first term in (1.4), we note that
$$\begin{split} &\sqrt{n}(\widehat{\theta}_{n}(\widehat{\boldsymbol{\beta}}_{n}) - \widehat{\theta}_{n}(\boldsymbol{\beta}_{0})) \\ &= \frac{\sqrt{n}}{n_{0}n_{1}} \sum_{j=1}^{n_{0}} \sum_{k=1}^{n_{1}} \left(I(\widehat{\boldsymbol{\beta}}_{n}^{\top} \mathbf{X}_{k}^{D} > \widehat{\boldsymbol{\beta}}_{n}^{\top} \mathbf{X}_{j}^{H}) - I(\boldsymbol{\beta}_{0}^{\top} \mathbf{X}_{k}^{D} > \boldsymbol{\beta}_{0}^{\top} \mathbf{X}_{j}^{H}) \right) \\ &= \frac{\sqrt{n}}{n_{0}n_{1}} \sum_{j=1}^{n_{0}} \sum_{k=1}^{n_{1}} \left(I(\widehat{\boldsymbol{\beta}}_{n}^{\top} (\mathbf{X}_{j}^{H} - \mathbf{X}_{k}^{D}) < 0) - I(\boldsymbol{\beta}_{0}^{\top} (\mathbf{X}_{j}^{H} - \mathbf{X}_{k}^{D}) < 0) \right) \\ &= \frac{1}{n_{0}n_{1}} \sum_{j=1}^{n_{0}} \sum_{k=1}^{n_{1}} \left(-\delta(\boldsymbol{\beta}_{0}^{\top} (\mathbf{X}_{j}^{H} - \mathbf{X}_{k}^{D})) (\mathbf{X}_{j}^{H} - \mathbf{X}_{k}^{D})^{\top} \sqrt{n}(\widehat{\boldsymbol{\beta}}_{n} - \boldsymbol{\beta}_{0}) + o_{p}(1) \right), \end{split}$$
(1.5)

where $\delta(\boldsymbol{\beta}_0^{\top}(\mathbf{X}^H - \mathbf{X}^D))$ denotes the Dirac delta function, which is zero almost everywhere except at $\boldsymbol{\beta}_0^{\top}\mathbf{X}^H = \boldsymbol{\beta}_0^{\top}\mathbf{X}^D$, where it is theoretically infinite. The last equation in Equation (1.5) results from a linear approximation approach. Although the indicator function is indeed nondifferentiable at the point where its argument is zero and the Dirac delta function is not a classical derivative, it is used here as a tool to approximate the impact of small deviations within an asymptotic framework. While $\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)$ converges in distribution to a normal distribution (see Theorem A.9), the Dirac delta function is highly localized. Therefore, for almost all realizations, $\delta(\boldsymbol{\beta}_0^{\top}(\mathbf{X}^H - \mathbf{X}^D))$ is zero unless $\boldsymbol{\beta}_0^{\top}\mathbf{X}^H = \boldsymbol{\beta}_0^{\top}\mathbf{X}^D$, which implies that $\delta(\boldsymbol{\beta}_0^{\top}(\mathbf{X}^H - \mathbf{X}^D)) = 0$ almost surely. By Theorem A.2, Slutsky's theorem, and the continuous mapping theorem, Equation (1.5) converges to 0 in probability. Consequently, the first term in (1.4) can be rewritten as:

$$\sqrt{n}(\widehat{\theta}_n(\widehat{\beta}_n) - \widehat{\theta}_n(\beta_0) - \theta(\widehat{\beta}_n) + \theta(\beta_0)) = -\sqrt{n}(\theta(\widehat{\beta}_n) - \theta(\beta_0)) + o_p(1).$$
(1.6)

Notice that Equation (1.6) implies that the sum of the first and the third terms in (1.4) converges to 0 in probability. Hence, $\sqrt{n}(\hat{\theta}_n(\hat{\beta}_n) - \theta(\beta_0))$ converges to the second term in (1.4) in probability:

$$\sqrt{n}(\widehat{\theta}_n(\widehat{\boldsymbol{\beta}}_n) - \theta(\boldsymbol{\beta}_0)) = \sqrt{n}(\widehat{\theta}_n(\boldsymbol{\beta}_0) - \theta(\boldsymbol{\beta}_0)) + o_p(1).$$

From the argument in Section 1.3.2, the second term in (1.4) converges in distribution to a Gaussian process:

$$\sqrt{n}(\widehat{\theta}_n(\boldsymbol{\beta}_0) - \theta(\boldsymbol{\beta}_0)) \xrightarrow{d} \frac{(\mathbb{G}_{F_1})_-(F_0)}{\sqrt{\lambda_1}} - \frac{\mathbb{G}_{F_0}(F_1)}{\sqrt{\lambda_0}}$$

Thus, the limit distribution of $\sqrt{n}(\hat{\theta}_n(\hat{\beta}_n) - \theta(\beta_0))$ can be deduced from the finitedimensional distributions of this Gaussian process:

$$\sqrt{n}(\widehat{\theta}_n(\widehat{\boldsymbol{\beta}}_n) - \theta(\boldsymbol{\beta}_0)) \xrightarrow{d} \mathcal{N}\left(0, \frac{\operatorname{Var}\left(F_0(\boldsymbol{\beta}_0^{\top} \mathbf{X}^D)\right)}{\lambda_1} + \frac{\operatorname{Var}\left(F_1(\boldsymbol{\beta}_0^{\top} \mathbf{X}^H)\right)}{\lambda_0}\right).$$

Surprisingly, the variability introduced by the MLE of class probability models diminishes as the sample size increases. The main cause of this result is from the use of indicator functions in this particular case. If the indicator functions were replaced by functions that possess non-zero derivatives on a set of non-zero measure, Equation (1.5) would converge to a non-degenerate distribution. In such a scenario, the first term in (1.4) would converge to 0 in probability, and the entire expression in (1.4) would converge to a distribution determined by the joint distribution of the second and third terms in (1.4) [19].

1.5 Simulation Studies

In this section, we evaluate the validity of the asymptotic results presented in Section 1.4 using simulated data. We investigate the bias between $\hat{\theta}_n(\hat{\beta}_n)$ and $\theta(\beta_0)$ and compare the asymptotic variance of $\hat{\theta}_n(\hat{\beta}_n)$ with its empirical variance across varying sample sizes. The analyses are conducted using code developed on the R software platform [20].

1.5.1 Simulation Method

We perform M = 1000 Monte Carlo (MC) simulation replicates for each sample size. We report results for 20 sample sizes, starting from n = 100 and increasing in increments of 100 to n = 2000, to evaluate the trend of asymptotic properties. For each MC replication, we simulate data (\mathbf{X}_i, Y_i), i = 1, ..., n, from the following marginal and conditional distributions:

$$\mathbf{X}_{i} = \begin{pmatrix} x_{1i} \\ x_{2i} \end{pmatrix} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \quad \boldsymbol{\mu} = \begin{pmatrix} 2 \\ 1 \end{pmatrix}, \quad \boldsymbol{\Sigma} = \begin{pmatrix} 3 & 0 \\ 0 & 1 \end{pmatrix},$$
$$Y_{i} \sim \text{Bernoulli}(p_{0}(\mathbf{X}_{i})), \quad p_{0}(\mathbf{X}_{i}) = \frac{1}{1 + \exp\left(-(\beta_{0} + x_{1i}\beta_{1} + x_{2i}\beta_{2})\right)}, \quad \boldsymbol{\beta}_{0} = \begin{pmatrix} \beta_{0} \\ \beta_{1} \\ \beta_{2} \end{pmatrix} = \begin{pmatrix} -3 \\ 0.2 \\ 1 \end{pmatrix}.$$

Given the simulated data, we apply logistic regression (a GLM with the logit link) to obtain the MLE $\hat{\beta}_n$ of β_0 . The asymptotic variance of $\hat{\theta}_n(\hat{\beta}_n)$ is estimated using the consistent estimator provided below:

$$\frac{\widehat{\sigma}_D^2}{n_1} + \frac{\widehat{\sigma}_H^2}{n_0},\tag{1.7}$$

where

$$\begin{aligned} \widehat{\sigma}_{D}^{2} &= \operatorname{Var}\left(\mathbb{F}_{n_{0}}(\widehat{\boldsymbol{\beta}}_{n}^{\mathsf{T}}\mathbf{X}^{D})\right) = \frac{1}{n_{1}-1} \sum_{k=1}^{n_{1}} \left(\frac{1}{n_{0}} \sum_{j=1}^{n_{0}} I(\widehat{\boldsymbol{\beta}}_{n}^{\mathsf{T}}\mathbf{X}_{j}^{H} < \widehat{\boldsymbol{\beta}}_{n}^{\mathsf{T}}\mathbf{X}_{k}^{D}) - \widehat{\theta}_{n}(\widehat{\boldsymbol{\beta}}_{n})\right)^{2}, \\ \widehat{\sigma}_{H}^{2} &= \operatorname{Var}\left(\mathbb{F}_{n_{1}}(\widehat{\boldsymbol{\beta}}_{n}^{\mathsf{T}}\mathbf{X}^{H})\right) = \frac{1}{n_{0}-1} \sum_{j=1}^{n_{0}} \left(\frac{1}{n_{1}} \sum_{k=1}^{n_{1}} I(\widehat{\boldsymbol{\beta}}_{n}^{\mathsf{T}}\mathbf{X}_{k}^{D} \le \widehat{\boldsymbol{\beta}}_{n}^{\mathsf{T}}\mathbf{X}_{j}^{H}) - \left(1 - \widehat{\theta}_{n}(\widehat{\boldsymbol{\beta}}_{n})\right)\right)^{2} \\ &= \frac{1}{n_{0}-1} \sum_{j=1}^{n_{0}} \left(\frac{1}{n_{1}} \sum_{k=1}^{n_{1}} I(\widehat{\boldsymbol{\beta}}_{n}^{\mathsf{T}}\mathbf{X}_{j}^{H} < \widehat{\boldsymbol{\beta}}_{n}^{\mathsf{T}}\mathbf{X}_{k}^{D}) - \widehat{\theta}_{n}(\widehat{\boldsymbol{\beta}}_{n})\right)^{2}. \end{aligned}$$

We compare the estimated asymptotic variance (1.7) with the empirical variance:

$$\frac{1}{M-1}\sum_{m=1}^{M} \left(\widehat{\theta}_{n}^{(m)}(\widehat{\boldsymbol{\beta}}_{n}) - \overline{\widehat{\theta}}_{n}(\widehat{\boldsymbol{\beta}}_{n})\right)^{2},$$

where $\overline{\hat{\theta}}_n(\widehat{\boldsymbol{\beta}}_n) = \frac{1}{M} \sum_{m=1}^M \widehat{\theta}_n^{(m)}(\widehat{\boldsymbol{\beta}}_n)$ and $\widehat{\theta}_n^{(m)}(\widehat{\boldsymbol{\beta}}_n)$ denotes the *m*-th MC replication. To demonstrate that the variability introduced by the MLE decreases as the sample size increases, we also calculate the estimated asymptotic variance of $\widehat{\theta}_n(\boldsymbol{\beta}_0)$ across these 20 sample sizes. This variance is computed using the consistent estimator given in (1.7), with $\boldsymbol{\beta}_0$ substituted for $\widehat{\boldsymbol{\beta}}_n$. We present the results in terms of standard errors (SEs) rather than variances to facilitate interpretation, as SEs are in the same units as the biases. Additionally, we use standard deviations (SDs) of $\sqrt{n}(\widehat{\theta}_n(\widehat{\boldsymbol{\beta}}_n) - \theta(\boldsymbol{\beta}_0))$ and $\sqrt{n}(\widehat{\theta}_n(\boldsymbol{\beta}_0) - \theta(\boldsymbol{\beta}_0))$ instead of SEs (without multiplying \sqrt{n}) when comparing the asymptotic variances of $\widehat{\theta}_n(\widehat{\boldsymbol{\beta}}_n)$ and $\widehat{\theta}_n(\boldsymbol{\beta}_0)$ to eliminate the effect of sample size, thereby more clearly revealing the reduction in variability induced by MLE estimation.

1.5.2 Simulation Results

It can be shown from Figure 1.1 that as the sample size increased, the bias $\hat{\theta}_n(\hat{\beta}_n) - \theta(\beta_0)$ generally decreased. The consistency of the estimator $\hat{\theta}_n(\hat{\beta}_n)$ is supported by this trend, as a consistent estimator should exhibit bias reduction with an increasing sample size. When compared to $\hat{\theta}_n(\beta_0) - \theta(\beta_0)$, this bias is typically larger, suggesting the presence of variation induced by estimating the MLE. In terms of variance analysis, Figure 1.2 shows that the asymptotic and empirical SEs are aligned, indicating reliable estimation under both theoretical conditions and practical simulation scenarios. The differences between the asymptotic and empirical SEs tended to decrease as the sample size increased, which is consistent with the theoretical expectation that the asymptotic variance approximation improves with larger sample sizes. The decrease in the discrepancy between the asymptotic SDs of $\hat{\theta}_n(\hat{\beta}_n)$ and $\hat{\theta}_n(\beta_0)$, as shown in Figure 1.3, indicates that the impact of variation due to the MLE weakened as the sample size grew. It is noteworthy that the SDs of the empirical AUROC with MLE are consistently smaller than the SDs of the empirical AUROC with the true parameter. This implies that the covariance between the empirical AUROC estimator and the MLE-based AUROC estimator is negative, according to (1.4) and the following:

$$Var(A+B) = Var(A) + Var(B) + 2Cov(A, B) < Var(A)$$

$$\Rightarrow Cov(A, B) < -\frac{1}{2}Var(B) \le 0,$$

where $A = \sqrt{n}(\hat{\theta}_n(\beta_0) - \theta(\beta_0))$ and $\sqrt{n}(\hat{\theta}_n(\hat{\beta}_n) - \hat{\theta}_n(\beta_0))$. Negative covariance indicates that the empirical AUROC and MLE-based AUROC estimators are inversely related in their variability. Specifically, when empirical AUROC estimates are high, their MLE-based counterparts tend to be lower, and vice versa. Based on the results shown in Figure 1.1, we can conclude that although the MLE-based AUROC has greater bias compared to the empirical AUROC with the true parameter, the smaller asymptotic variance of empirical AUROC with the estimated parameter compared to that with the true parameter is a common phenomenon [21].

It is important to note that the decreased discrepancy in SDs is not directly attributed to increased sample size. As $SE = SD/\sqrt{n}$, the discrepancy between the two SEs naturally decreases as sample size increases due to the \sqrt{n} term. However, if the discrepancy between the two SDs also decreases, this suggests a reduction in the sampling variability of $\hat{\beta}_n$, which is independent of the effect of sample size. If the variation from the MLE were not negligible (i.e., did not converge to 0 in probability), one would expect a roughly constant discrepancy between the asymptotic standard deviations of $\hat{\theta}_n(\hat{\beta}_n)$ and that of $\hat{\theta}_n(\beta_0)$ regardless of sample size. Neither Figure 1.1 nor Figure 1.2 provides conclusive evidence regarding whether this variation is negligible, as the reduction in discrepancy for both bias and standard errors could be attributed to either increasing sample size or MLE estimation effects. Figure 1.3 shows the plot of the two SDs as a function of sample size. Unlike Figure 1.1 and Figure 1.2, this plot shows the change of the variability of the two AUROCs independent of the impact of sample size.



Figure 1.1. Comparison of the bias between $\widehat{\theta}_n(\widehat{\beta}_n) - \theta(\beta_0)$ and $\widehat{\theta}_n(\beta_0) - \theta(\beta_0)$ across varying sample sizes.

These results demonstrate how sample size affects the reliability of parameter estimates in the context of AUROC evaluation. Larger sample sizes reduce the variation associated with the MLE estimator, revealing how sample size impacts the stability of the estimates and supporting the robustness of the asymptotic variance.

1.6 Discussion

In this paper, we examined the asymptotic properties of empirical AUROC by accounting for the sampling variability of the MLE from the GLM-based class probability model and derived the consistency and asymptotic normality of the AUROC. We focused on all continuous variables



Figure 1.2. Comparison of asymptotic and empirical standard errors for $\hat{\theta}_n(\hat{\beta}_n)$ across varying sample sizes.

in the feature vector in the class probability model. If the feature vector contains categorical variables, one-hot encoding [22] can be applied to represent them using dummy variables so that the established asymptotic properties can be applied to this setting as well.

While it is important to acknowledge the inherent limitation of the asymptotic results– their dependency on large sample size–our comparison of asymptotic and empirical SEs provides valuable insight. As shown in Figure 1.2, the asymptotic SEs appear to approximate the empirical SEs reasonably well overall even for smaller sample sizes. This suggests that the asymptotic SEs offer a reliable approximation even for small to moderate sample sizes.

Future research could address the limitation that the established asymptotic properties



Figure 1.3. Comparison of asymptotic standard deviations for $\hat{\theta}_n(\hat{\beta}_n)$ and $\hat{\theta}_n(\beta_0)$ across varying sample sizes.

for MLE-based AUROC estimators are applicable only to binary classification. Hand and Till [23] extended the definition to multi-class scenarios by averaging the AUROC values for each pair of classes. Building on this work, our derivation could be extended to this multi-class AUROC. Another direction for future research could involve exploring the use of other estimators, including those from regularization techniques such as ridge regression [24] and Lasso [25]. Models built with these techniques may also be evaluated using AUROC as a performance metric.

In conclusion, this study enhances our understanding of how sampling variability of the MLE influences the asymptotic properties of empirical AUROCs based on GLM class probability

models. The findings may extend beyond MLE to all consistent estimators such as those from semiparametric GLM to improve the asymptotic variance of empirical AUROC with these estimators.

1.7 Acknowledgements

Chapter 1, in part is currently being prepared for submission for publication of the material. Wu, Tsung-Chin; Lin, Tuo; Yang, Kun; Schnabl, Bernd; Zhang, Xinlian; Tu, Xin. The dissertation author was the primary investigator and author of this material.

Chapter 2

Outlier-Robust Random Forest Regression: Application to Between-Subject Attributes

2.1 Introduction

Between-subject attributes refer to characteristics or variables that differ among individual subjects. Inferences about parameters defined by between-subject attributes are of interest in many biomedical studies. Examples of quantities that relate to between-subject attributes are beta-diversity, which measures the distance between taxonomic counts of two individuals in microbiome studies [1]; viral genetic linkage, defined by the distance between genetic sequences of two individuals in epidemiological studies [2]; and the area under the receiver operating characteristic curve, used to evaluate the diagnostic accuracy of biomarkers [14]. For between-subject attributes, the relationships between the *i*-th and *j*-th subjects and the *i*-th and *k*-th subjects introduce dependencies between these attributes. Statistical models such as generalized linear models [13] fail to take dependencies among outcomes involving multiple subjects into account, which presents a major challenge when modeling these attributes. Consequently, theories that establish asymptotic properties, such as the central limit theorem and the law of large numbers, become inapplicable. Models with functional responses involving multiple subjects are necessary for appropriate analysis in order to address this issue.

Semiparametric regression models are typically used to handle functional responses involving multiple subjects. For instance, Liu, Zhang, et al. [1] proposed a semiparametric regression approach using functional response models (FRMs) [12] to analyze beta-diversity. These models, however, may provide inaccurate estimates if the underlying parametric assumptions for the mean response are violated. An alternative to this limitation is offered by nonparametric methods using random forest regression [26], which model complex relationships between functional responses and features of between-subject attributes without relying on strict parametric assumptions. This approach allows for more flexible and robust modeling of between-subject attributes [27].

Despite its advantages, random forest regression is not immune to the influence of outliers in the dependent variable [28, 29, 30], similar to other regression methods that handle continuous responses, such as linear models. As shown in M. Brence and Brown [28], unbounded outliers and heteroscedastic conditions can affect the performance of the traditional random forest algorithm, which relies on mean-based predictions and mean-squared error (MSE). J. R. Brence [29] proposed a variant that uses median-based predictions and mean absolute deviation and demonstrated improved performance with datasets containing outliers. Li and Martin [30] introduced a general framework for incorporating robust loss functions into forest-type regression and showed increased insensitivity to outliers. Inspired by this framework, we propose a new approach to employ the response function $h(Y_i, Y_i) = I(Y_i \le Y_i)$, where Y_i and Y_i represent the outcomes of the data. This is inherently less sensitive to outliers. Chen et al. [3] has demonstrated that this outlier-insensitive response function performs effectively against outliers within the framework of the FRMs. However, this FRM-based semiparametric approach relies on a correct (parametric) specification of the mean response function to provide valid inference. In this study, we extend this approach to a nonparametric mean response function using the random forest framework to ensure more reliable modeling of between-subject attributes in the presence of extreme values.

Strong predictive performance has been demonstrated by machine learning methods,

including random forests, which are widely utilized for modeling high-dimensional data [26]. However, many of these methods lack theoretical justification for their asymptotic properties. Recent work by Wager and Athey [31] has established the consistency and asymptotic normality of random forests in nonparametric regressions. Furthermore, T. Lin [32] extended these results by demonstrating the asymptotic properties of estimators for the Mann-Whitney-Wilcoxon type of causal effect, marking it the very first successful attempt to extend random forest regression from within-subject attributes to between-subject attributes.

In this study, we introduce an outlier-robust nonparametric approach for modeling between-subject attributes by building upon the work of T. Lin [32] to develop our outlier-robust random forest regression setting by extending estimates calculated from within-leaf to between-leaves. As in T. Lin [32], we adapt the classical random forest regression [26], originally designed for within-subject attributes, to accommodate modeling between-subject attributes. By incorporating outlier-robust adjustments, we propose a different extension to the between-subject attributes for random forest. This newly developed method improves resilience against extreme values in our setting to provide more accurate and reliable estimates.

This paper is organized as follows. Section 2.2 provides a brief introduction to classical random forest regression, FRMs, and the outlier-robust response function employed in this study. Section 2.3 introduces the outlier-robust random forest regression for between-subject attributes. In Section 2.4, we establish asymptotic properties of our proposed random forest regression. Section 2.5 presents simulation studies to evaluate the performance of the proposed random forest regression under deviations from parametric assumptions of the semiparametric model and to assess the robustness of the model against outliers. In Section 2.6, we provide our concluding remarks and discuss future directions.

2.2 Background

2.2.1 Classical Random Forest Regression

Consider a sample of *n* subjects and *d* features in the study. Let $Z_i = (X_i, Y_i)$, where i = 1, ..., n, denote an independent and identically distributed (i.i.d.) training sample with continuous responses $Y_i \in \mathbb{R}$ and continuous feature vectors $X_i \in [0, 1]^d$. Let $x \in [0, 1]^d$ represent a test point, and define the true conditional mean function

$$\mu(x) = E[Y \mid X = x].$$

Let $T(x;\xi,Z_1,...,Z_n)$ be a regression tree using the classification and regression tree (CART) methodology [33] constructed using $\{Z_1,...,Z_n\}$, where $\xi \sim \Xi$ introduces auxiliary randomness from some unknown distribution Ξ (see Wager and Athey [31] for details). The tree *T* can be used to estimate μ at *x* through the following process. First, the feature space is recursively split until it is partitioned into a set of leaves *L*, with each containing a small number of training subsamples. For a given test point *x*, the estimator $\hat{\mu}(x)$ is determined by identifying the leaf L(x) that contains *x* and setting

$$\widehat{\mu}(x) = \frac{1}{|\{i : X_i \in L(x)\}|} \sum_{i \in \{i : X_i \in L(x)\}} Y_i.$$

By aggregating multiple trees, we can form a random forest estimator and estimate μ nonparametrically by averaging estimates from trees trained on all possible subsamples of size *s* from the training data (without replacement), while marginalizing over the auxiliary noise ξ :

$$\operatorname{RF}(x; Z_1, \dots, Z_n) = {\binom{n}{s}}^{-1} \sum_{(i_1, \dots, i_s) \in \binom{S_n}{s}} E_{\xi \sim \Xi} [T(x; \xi, Z_{i_1}, \dots, Z_{i_s})],$$

where $S_n := \{1, ..., n\}.$

2.2.2 Functional Response Models

Using FRMs is one approach of using semiparametric regression to model betweensubject attributes. Let $Z_i = (X_i, Y_i)$, where i = 1, ..., n, be an i.i.d. random sample, with X_i as a $d \times 1$ vector of independent variables and Y_i as a response. The definition of the general form of an FRM is as follows [12]:

$$E[h(Y_{i_1}, \dots, Y_{i_q}) \mid X_{i_1}, \dots, X_{i_q}] = f(X_{i_1}, \dots, X_{i_q}; \beta), \quad (i_1, \dots, i_q) \in \binom{S_n}{q},$$
(2.1)

where *h* is a real-valued function, *f* is a smooth function with continuous second-order derivatives, and β is a *d* × 1 vector of unknown parameters.

2.2.3 Outlier-Robust Response Function

If there are outliers (extremely large in absolute values) in the response, we make the model less vulnerable to outliers in the response by using the response function

$$h(Y_i, Y_j) = I(Y_i \le Y_j), \quad (i, j) \in \binom{S_n}{2},$$

where *I* denotes the indicator function. Note that *h* depends solely on the ranks of the responses Y_i and Y_j . Unlike $|Y_i - Y_j|$, which varies with changes in Y_i and Y_j , *h* remains invariant as long as the rank order of Y_i and Y_j is preserved. Thus, models employing *h* are rank-preserving and robust to outliers [3].

2.3 Outlier-Robust Random Forest Regression for Between-Subject Attributes

Under the setting described in Section 2.2.1, consider two test points $x^{(1)}, x^{(2)} \in [0, 1]^d$. We define the conditional mean outlier-robust functional response between two subjects with feature vectors at $(x^{(1)}, x^{(2)})$ as

$$\begin{aligned} \theta(x^{(1)}, x^{(2)}) &= E\left[h(Y_i, Y_j) \mid (X_i, X_j) = (x^{(1)}, x^{(2)})\right] \\ &= E\left[I(Y_i \le Y_j) \mid (X_i, X_j) = (x^{(1)}, x^{(2)})\right] \\ &= P(Y_i \le Y_j \mid (X_i, X_j) = (x^{(1)}, x^{(2)})), \quad (i, j) \in \binom{S_n}{2}. \end{aligned}$$

To extend classical random forest regression to an outlier-robust random forest regression for between-subject attributes in our setting, we propose the following tree-growing procedure, which employs the double-sample trees method [31]. This method involves dividing the training sample into two parts: one part is used to estimate the conditional mean function θ , while the other is used to determine the splits when growing the trees.

Procedure 2.1. (*Outlier-Robust Regression Tree for Between-Subject Attributes*) *Input: Training samples* $\{Z_1, ..., Z_n\}$ *and a minimum leaf size k.*

- 1. Draw a random subsample of size s from $\{Z_1, \ldots, Z_n\}$ without replacement.
- Divide this random subsample into two disjoint sets I and J with sizes |I| = ⌊s/2⌋ and |J| = ⌈s/2⌉, respectively, where ⌊·⌋ denotes the floor function and ⌈·⌉ denotes the ceiling function. The notation for the sets I and J follows the conventions used in Wager and Athey [31].
- 3. Grow a tree using recursive partitioning (the algorithm used in CART trees). For each node, evaluate potential splits by randomly selecting a feature vector from the whole sample, $I \cup J$, and calculating the MSE of predictions for each candidate split using responses in the J-sample only. The best split is chosen based on minimizing the weighted MSE of the child nodes. Splits are constrained so that each leaf of the tree must contain at least k I-sample observations.

- 4. Identify the leaves $L(x^{(1)})$ and $L(x^{(2)})$ that contain $x^{(1)}$ and $x^{(2)}$, respectively.
- 5. Estimate θ using only the responses from the *I*-sample by calculating

$$\widehat{\theta}_{tree}(x^{(1)}, x^{(2)}) = \begin{cases} \frac{1}{|L^{(1)}||L^{(2)}|} \sum_{\substack{i \in L^{(1)} \\ j \in L^{(2)}}} I(Y_i \le Y_j) & \text{if } L(x^{(1)}) \neq L(x^{(2)}), \\ \\ \frac{1}{|L^{(1)}|(|L^{(1)}|-1)} \sum_{\substack{i,j \in L^{(1)} \\ i \neq j}} I(Y_i \le Y_j) & \text{if } L(x^{(1)}) = L(x^{(2)}), \end{cases}$$

$$(2.2)$$

where $L^{(1)} = \{i : X_i \in L(x^{(1)})\}$ and $L^{(2)} = \{i : X_i \in L(x^{(2)})\}.$

A random forest estimate of $\theta(x^{(1)}, x^{(2)})$ can then be obtained by averaging all $\binom{n}{s}$ possible tree estimates $\widehat{\theta}_{\text{tree}}(x^{(1)}, x^{(2)})$ derived from trees built using Procedure 2.1, while integrating over the auxiliary noise ξ :

$$\widehat{\theta}_{\text{forest}}(x^{(1)}, x^{(2)}; Z_1, \dots, Z_n) = \binom{n}{s}^{-1} \sum_{(i_1, \dots, i_s) \in \binom{S_n}{s}} E_{\xi \sim \Xi}[\widehat{\theta}_{\text{tree}}(x^{(1)}, x^{(2)}; \xi, Z_{i_1}, \dots, Z_{i_s})].$$
(2.3)

However, computing $\hat{\theta}_{\text{forest}}$ directly for large *n* is often impractical. Instead, we generally approximate $\hat{\theta}_{\text{forest}}$ using Monte Carlo averaging:

$$\widehat{\theta}(x^{(1)}, x^{(2)}; Z_1, \dots, Z_n) = \frac{1}{B} \sum_{b=1}^B \widehat{\theta}_{\text{tree}}(x^{(1)}, x^{(2)}; \xi_b^*, Z_{b1}^*, \dots, Z_{bs}^*),$$
(2.4)

where $\{Z_{b1}^*, \dots, Z_{bs}^*\}$ are subsamples of size *s* drawn from $\{Z_1, \dots, Z_n\}$, ξ_b^* is a random sample from Ξ , and *B* denotes the number of Monte Carlo replicates. This method provides a practical alternative for computing random forest estimators, as $\binom{n}{s}$ is large and computing $\hat{\theta}_{\text{forest}}$ directly is not only computationally expensive but also infeasible with large *n*. The effects of choosing a finite number of Monte Carlo samples *B* on the accuracy of the approximation have been discussed by Mentch and Hooker [34].

Notice that from Equation (2.2), $\hat{\theta}_{\text{tree}}(x^{(1)}, x^{(2)})$ is always equal to 0.5 if $x^{(1)}$ and $x^{(2)}$ are in the same leaf, but $\hat{\theta}_{\text{forest}}(x^{(1)}, x^{(2)})$ and $\hat{\theta}(x^{(1)}, x^{(2)})$ may not be 0.5 since $x^{(1)}$ and $x^{(2)}$ may not always be in the same leaf across all different trees.

2.4 Asymptotic Theory

Our asymptotic theory for outlier-robust random forest regression extends the work of Wager and Athey [31] on the asymptotic analysis of random forest regression for within-subject attributes. Although earlier research has explored the convergence and consistency of within-subject random forests [35, 36, 37], the seminal work of Wager and Athey [31] is crucial in demonstrating their asymptotic normality. Building on this, our research extends these results by establishing the consistency and asymptotic normality of outlier-robust random forest regression for between-subject attributes within our context.

2.4.1 Preliminary Conditions

Before delving into the details, we first outline the preliminary conditions required to establish our asymptotic results. The following condition, referred to as honesty (Definition 2.2) and originating from Wager and Athey [31], must be satisfied by each tree.

Definition 2.2. A tree is considered honest if, for each training sample Z_i , the response Y_i is used exclusively either to estimate the true conditional mean function θ using Equation (2.2) or to determine the placement of splits, but not for both purposes.

Growing trees using Procedure 2.1 always satisfies the honesty condition, as the responses in the \mathcal{I} -sample are used exclusively for estimation, while those in the \mathcal{J} -sample are used solely for splitting.

As the sample size n increases, it is essential that the range of the feature values within the leaves becomes more refined (i.e., smaller) in each dimension of the feature space to ensure consistency in the random forest regression. We follow the approach from Wager and Athey [31], which builds upon the methodology introduced by Meinshausen and Ridgeway [38] and involves incorporating randomness into the process of selecting variables for tree splits, to achieve this. Specifically, a variable is chosen with a probability of at least π/d , where $0 < \pi \le 1$ (Definition 2.3). This randomness induced by choosing splits for features is contained in the auxiliary random variable ξ .

Definition 2.3. A tree is considered a random-split tree if, at every step where a feature is selected for splitting during the tree-growing process, the probability of choosing the *j*-th feature, after marginalizing over the random variable ξ , is at least π/d for some $0 < \pi \le 1$, for all j = 1, ..., d.

Next, we adopt the regularity condition (Definition 2.4) from Wager and Athey [31] to control the shape of the tree leaves, ensuring an appropriate number of observations in each leaf for accurate estimation, and the symmetry condition (Definition 2.5) to apply classical tools in proving asymptotic normality.

Definition 2.4. A tree grown by recursive partitioning is α -regular for some $\alpha > 0$ if the tree in *Procedure 2.1 satisfies the following condition for the I-sample: At each split, at least a fraction of* α *of the available training samples are allocated to each side of the split and, moreover, the tree is fully grown until there are between k and* 2k - 1 *observations in each leaf for some* $k \in \mathbb{N}$.

Definition 2.5. A tree is symmetric if the (possibly randomized) estimated values $\hat{\theta}_{tree}$ defined in Equation (2.2) do not depend on the order in which the training samples are indexed.

Note that our targeted response $I(Y_i \le Y_j)$ changes value if the order of Y_i and Y_j is switched. Therefore, to satisfy Definition 2.5, the order can be permuted only within leaves, not between leaves.

2.4.2 Consistency

To establish the consistency of outlier-robust random forest estimators, it is sufficient to show their asymptotic unbiasedness with vanishing variance. We achieve this by deriving asymptotic unbiasedness through bounding the bias of regression trees, using an approach that relies on the Lipschitz continuity of the true conditional mean function and honesty, following the framework provided by Wager and Athey [31]. Specifically, we leverage Lemma 1 in Wager and Athey [31] (restated in Lemma A.12 in Appendix A.4 of the Relevant Existing Theorems section) as a tool to bound the bias of a single regression tree. Since a random forest is the sample mean of independently generated trees, its bias is of the same order as that of a single tree. The consistency of the outlier-robust random forest estimator is stated in the following theorem, which extends their Theorem 3.2 to fit our nonparametric regression for between-subject attributes.

Theorem 2.6. Suppose that we have n i.i.d. training samples $Z_i = (X_i, Y_i) \in [0, 1]^d \times \mathbb{R}$. Assume further that the features X_i are sampled i.i.d. from Unif $([0, 1]^d)$, and that the function

$$\theta(x^{(1)}, x^{(2)}) = E[h(Y_i, Y_j) \mid (X_i, X_j) = (x^{(1)}, x^{(2)})], \quad (i, j) \in \binom{S_n}{2}$$

is Lipschitz continuous, where $h(Y_i, Y_j) = I(Y_i \le Y_j)$. Let T be an honest, α -regular with $\alpha \le 0.2$, and random-split tree according to Definitions 2.2, 2.3, and 2.4, and let $\widehat{\theta}_{forest}(x^{(1)}, x^{(2)})$ be the estimate for $\theta(x^{(1)}, x^{(2)})$ given by a random forest with base learner T and a subsample size s. Then, for sufficiently large s, the bias of the random forest estimator at $(x^{(1)}, x^{(2)})$ can be bounded by

$$\left| E[\widehat{\theta}_{forest}(x^{(1)}, x^{(2)})] - \theta(x^{(1)}, x^{(2)}) \right| \le K\sqrt{2d} (1+2d) s^{-\frac{1}{2} \frac{\log((1-\alpha)^{-1})}{\log(\alpha^{-1})} \frac{\pi}{d}},$$

where *K* is the Lipschitz constant for $\theta(x^{(1)}, x^{(2)})$.

Proof. See Appendix B.2.1 for a proof.

Note that, although Theorem 2.6 assume uniformity for the feature vectors X_i , this assumption is primarily for simplifying the presentation. The results are applicable to any continuous distribution, thanks to the probability integral transformation [12].

2.4.3 Asymptotic Normality

In this section, we prove the asymptotic normality of outlier-robust random forest regression (Theorem 2.7). To avoid any confusion, Theorem 2.7 is stated using notation that explicitly indicates the dependence of $\hat{\theta}_{\text{forest, }n}$ and s_n on n. Although inspired by the Theorem 3.1 of Wager and Athey [31], this theorem presents new results that involve a major overhaul of their original framework to account for the specific setting of the outlier-robust nonparametric regression.

Theorem 2.7. Suppose that we have n i.i.d. training samples $Z_i = (X_i, Y_i) \in [0, 1]^d \times \mathbb{R}$. Assume further that the features X_i are sampled i.i.d. from Unif $([0, 1]^d)$, and that the function

$$\theta(x^{(1)}, x^{(2)}) = E[h(Y_i, Y_j) \mid (X_i, X_j) = (x^{(1)}, x^{(2)})], \quad (i, j) \in \binom{S_n}{2}$$

is Lipschitz continuous, where $h(Y_i, Y_j) = I(Y_i \le Y_j)$. Additionally, suppose for $(i, j) \in \binom{S_n}{2}$,

$$\operatorname{Var}(h(Y_i, Y_j) \mid X_i = x^{(1)}) > 0,$$

and

$$E\left[\left|E[h(Y_i, Y_j) \mid Y_i] - E[h(Y_i, Y_j) \mid X_i = x^{(1)}]\right|^{2+\delta} \middle| X_i = x^{(1)}\right] \le M$$

for some positive constants δ and M, uniformly across all $x^{(1)} \in [0,1]^d$. Given this datagenerating process, let T be an honest, α -regular with $\alpha \leq 0.2$, and symmetric random-split tree according to Definitions 2.2, 2.3, 2.4, and 2.5, and let $\hat{\theta}_{forest, n}(x^{(1)}, x^{(2)})$ be the estimate for $\theta(x^{(1)}, x^{(2)})$ given by a random forest with base learner T and a subsample size s_n . If the subsample size s_n scales as

$$s_n \approx n^{\gamma} \quad for \ some \quad \gamma_{\min} := 1 - \left(1 + \frac{\log\left(\alpha^{-1}\right)}{\log\left((1-\alpha)^{-1}\right)} \frac{d}{\pi}\right)^{-1} < \gamma < 1, \tag{2.5}$$

then the random forest estimators are asymptotically normal:

$$\frac{\widehat{\theta}_{forest,n}(x^{(1)}, x^{(2)}) - \theta(x^{(1)}, x^{(2)})}{\sigma_n(x^{(1)}, x^{(2)})} \xrightarrow{d} \mathcal{N}(0, 1) \quad for \ a \ sequence \quad \sigma_n(x^{(1)}, x^{(2)}) \to 0.$$
(2.6)

Proof. See Appendix B.2.2 for a proof.

When the subsampling rate meets the criterion in (2.5), the bias decreases at a faster rate than the variance, thereby establishing the asymptotic normality result (2.6) as stated in Theorem 2.7.

2.5 Simulation Studies

In Section 2.5.1, we assess the performance of the proposed random forest regression by comparing it with the FRM-based semiparametric model approach, introduced in Section 2.2.2, under the scenarios where the parametric mean response function is incorrectly specified. In Section 2.5.2, we evaluate the performance of the outlier-robust random forest regression on data containing outliers. The analyses are conducted using the code developed on the R software platform [20] and the generalized random forests from the grf package [31, 39], which support honest estimation as defined in Definition 2.2.

2.5.1 Comparison of Random Forest Regression and FRM under Parametric Assumption Violations

2.5.1.1 Simulation Methods

To assess the performance of random forest regression versus FRM-based semiparametric regression under violations of parametric assumptions in the latter approach, we first simulate

data from the following linear model:

$$Y_{i} = X_{1i} + \varepsilon_{i}, \quad \begin{pmatrix} X_{1i} \\ X_{2i} \end{pmatrix}^{i.i.d.} \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right),$$

$$\varepsilon_{i} \stackrel{i.i.d.}{\sim} \mathcal{N}\left(0, \sigma_{\varepsilon}^{2}\right), \quad \sigma_{\varepsilon}^{2} = 1/2, \quad 1 \le i \le n = 500.$$

$$(2.7)$$

Subsequently, to examine the impact of violations of parametric assumptions on model performance, we simulate data from the following model using the hyperbolic tangent function:

$$Y_{i} = \tanh(X_{1i}) + \varepsilon_{i}, \quad \begin{pmatrix} X_{1i} \\ X_{2i} \end{pmatrix}^{i.i.d.} \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right),$$

$$\varepsilon_{i} \stackrel{i.i.d.}{\sim} \mathcal{N}\left(0, \sigma_{\varepsilon}^{2}\right), \quad \sigma_{\varepsilon}^{2} = 1/2, \quad 1 \le i \le n = 500.$$

$$(2.8)$$

Note that in both models, Y_i is only related to X_{1i} and is unrelated to X_{2i} ; however, both covariates will be included in the models when fitting them. Adding an additional feature can increase the flexibility of FRM-based semiparametric regression and facilitate convergence when the parametric assumption for the conditional mean function is violated.

For the random forest regression, we utilized Equation (2.4) to construct the random forest estimator for $\theta(X_i, X_j)$, where $X_i = (X_{1i}, X_{2i})$. The forest was grown using 100 trees. Bootstrap methods [40], with a bootstrap sample size of 1000, were applied to construct 95% confidence intervals (CIs) for the random forest estimates.

For the FRM, consider the following classic linear model:

$$Y_i = X_i \beta + \varepsilon_i, \quad \beta \in \mathbb{R}^2, \quad \varepsilon_i \stackrel{i.i.d.}{\sim} \mathcal{N}\left(0, \sigma_{\varepsilon}^2\right), \quad 1 \le i \le n.$$

$$(2.9)$$

Thus, we have

$$E[I(Y_i \le Y_j) \mid X_i, X_j] = P\left(\varepsilon_i - \varepsilon_j \le -(X_i - X_j)\beta \mid X_i, X_j\right)$$
$$= P\left(\frac{1}{\sqrt{2}\sigma_{\varepsilon}}(\varepsilon_i - \varepsilon_j) \le -\frac{1}{\sqrt{2}\sigma_{\varepsilon}}X_i - X_j)\beta \mid X_i, X_j\right)$$
$$= \Phi\left(-\frac{1}{\sqrt{2}\sigma_{\varepsilon}}(X_i - X_j)\beta\right), \quad (i, j) \in \binom{S_n}{2}, \tag{2.10}$$

where Φ denotes the cumulative distribution function of the standard normal distribution. Notice that since we simulate the data with $\sigma_{\varepsilon}^2 = 1/2$ according to Equations (2.7) and (2.8), Equation (2.10) can be simplified to

$$E[I(Y_i \le Y_j) \mid X_i, X_j] = \Phi\left(-(X_i - X_j)\beta\right), \quad (i, j) \in \binom{S_n}{2}.$$
(2.11)

to avoid scaling in β when modeling. By setting $h(Y_i, Y_j) = I(Y_i \le Y_j)$ and $f(X_i, X_j; \beta) = \Phi(-(X_i - X_j)\beta)$, it is evident that the model in Equation (2.11) is an FRM defined in Equation (2.1).

To obtain estimates of $P(Y_i \le Y_j \mid X_i, X_j)$, we first estimate β using the *U*-statistics based generalized estimating equations (UGEE) for FRM [12]. The UGEE for the FRM in Equation (2.11) is given by:

$$U_n(\beta) = \sum_{(i,j) \in \binom{S_n}{2}} D_{ij}(\beta)^\top V_{ij}(\beta)^{-1} S_{ij}(\beta),$$

where

$$\begin{split} S_{ij}(\beta) &= h(Y_i, Y_j) - f(X_i, X_j; \beta), \\ D_{ij}(\beta) &= \frac{d}{d\beta} f(X_i, X_j; \beta), \\ V_{ij}(\beta) &= f(X_i, X_j; \beta) (1 - f(X_i, X_j; \beta)). \end{split}$$

We compute the estimate $\widehat{\beta}$ by solving $U_n(\beta) = 0$ using the Gauss-Newton method [41]. Estimates of $P(Y_i \le Y_j \mid X_i, X_j)$ are then obtained by substituting $\widehat{\beta}$ for β into Equation (2.11).

Additionally, a consistent estimator $\widehat{\Sigma}_{\beta}$ of the asymptotic variance of $\sqrt{n}(\widehat{\beta} - \beta)$ can be constructed by replacing β with $\widehat{\beta}$ in the expression for the asymptotic variance, as derived in Chen et al. [3]. Specifically,

$$\widehat{\Sigma}_{\beta} = \widehat{B}^{-1} \widehat{\Sigma}_U \widehat{B}^{-1},$$

where

$$\begin{split} \widehat{\Sigma}_{U} &= \frac{4}{n} \sum_{i=1}^{n} \widehat{u}_{i} \widehat{u}_{i}^{\top}, \quad \widehat{u}_{i} = \frac{1}{n-1} \sum_{j \in S_{n} \setminus \{i\}} D_{ij}(\widehat{\beta})^{\top} V_{ij}(\widehat{\beta})^{-1} S_{ij}(\widehat{\beta}), \\ \widehat{B} &= \binom{n}{2}^{-1} \sum_{(i,j) \in \binom{S_{n}}{2}} D_{ij}(\widehat{\beta})^{\top} V_{ij}(\widehat{\beta})^{-1} D_{ij}(\widehat{\beta}). \end{split}$$

We can therefore construct 95% CIs for the FRM estimates using the above variance estimates.

2.5.1.2 Simulation Results

To present the results effectively, we plot estimates of $P(Y_i \le Y_j | X_i, X_j)$ by fixing $X_j =$ (2, 0) and varying X_{1i} (ranging from -2 to 1.98), while keeping $X_{2i} = 0$ fixed, as it should not contribute any predictive power according to Equations (2.7) and (2.8). When the parametric form of the model assumption aligns with the data generation process, FRM accurately estimates the true $P(Y_i \le Y_j | X_i, X_j)$, as shown in Figure 2.1. Although random forest regression also predicts the true probability reasonably well, a discernible difference remains between the performances of the random forest and FRM, with larger bias and CIs of the former and the smaller bias and CIs of the latter. Conversely, when the model assumption does not align with the data, Figure 2.2 illustrates that the random forest captured the true values with relatively small bias whereas FRM failed to accurately capture these values and exhibited larger bias.

The nonlinear characteristics of the hyperbolic tangent function became apparent when X_{1i} got closer to $X_{1j} = 2$, leading to poorer FRM performance. In contrast, the random forest regression continued to perform well even when the nonlinear trend is apparent. The CIs of the random forest regression were generally larger than those of FRM, indicating that while nonparametric models offer greater robustness to model regression relationships with less bias estimates, these advantage comes at the cost of increased variance.



Figure 2.1. Comparison of predicted $P(Y_i \le Y_j | X_i, X_j)$ between random forest regression and FRM with correct parametric assumptions.



Figure 2.2. Comparison of predicted $P(Y_i \le Y_j | X_i, X_j)$ between random forest regression and FRM under violated parametric assumptions.

2.5.2 Assessing the Performance of Outlier-Robust Random Forest Regression in the Presence of Outliers

2.5.2.1 Simulation Methods

To assess the robustness of the proposed random forest regression against outliers, we first simulate data based on the model specified in Equation (2.7) and then we introduce outliers by replacing the 50 (10% of *n*) largest Y_i values with new outlying observations. Specifically, we order the simulated Y_i values from smallest to largest:

$$Y_{(1)} \le Y_{(2)} \le \ldots \le Y_{(500)}$$

We generate 50 outliers $\{U_1, U_2, \dots, U_{50}\}$ from a uniform distribution, Unif(10, 100), and order them in ascending order as:

$$U_{(1)} \le U_{(2)} \le \ldots \le U_{(50)}.$$

These ordered outliers replace the largest 50 Y_i values, such that

$$Y_{(451)} \to U_{(1)}, \quad Y_{(452)} \to U_{(2)}, \quad \dots, \quad Y_{(500)} \to U_{(50)}.$$

With the exception of the top 10% of the largest observations, which have been replaced with outliers, the dataset thus remains identical to the one generated from the model in Equation (2.7). We then compare the performance of two random forest regression models with one using the sigmoid response function:

$$h(Y_i, Y_j) = \frac{1}{1 + \exp(-|Y_i - Y_j|)}$$

and the other using $h(Y_i, Y_j) = I(Y_i \le Y_j)$ on both the originally simulated data and the data containing outliers. The sigmoid response function approaches 1 when $|Y_i - Y_j|$ is large and tends to 0.5 when $|Y_i - Y_j|$ is small, allowing us to compare its performance with the response function $I(Y_i \le Y_j)$. As discussed in Section 2.2.3, the sigmoid response function is expected to be sensitive to outliers, as $|Y_i - Y_j|$ varies with changes in Y_i and Y_j .

2.5.2.2 Simulation Results

As in Section 2.5.1.2, we present results by fixing $X_j = (2,0)$ and plotting the estimate for varying values of X_{1i} (ranging from -2 to 1.98), while keeping $X_{2i} = 0$ constant. As shown in Figure 2.3, it is evident that outliers significantly impacted estimates for the random forest model with the sigmoid response function. Notably, as we replace the top 10% of the largest Y_i values generated from Equation (2.7), $X_{1j} = 2$ should correspond to an outlier in our simulated data. Consequently, we observe that the estimated values remained close to 1 when X_{1i} is near X_{1j} , indicating that outliers biased the estimates, as $|Y_i - Y_j|$ is much larger in the presence of outliers than in their absence. In contrast, Figure 2.4 shows that estimates from the random forest model with the response function $I(Y_i \le Y_j)$ were not affected by outliers, as the order of Y_i remained preserved even with the outliers. Therefore, our proposed outlier-robust random forest approach effectively captured the regression relationship of interest, while remaining robust to outliers in the response variable.



Figure 2.3. Comparison of predicted $E[(1 + \exp(-|Y_i - Y_j|))^{-1}|X_i, X_j]$ between random forest regression with and without outliers in the response.



Figure 2.4. Comparison of predicted $P(Y_i \le Y_j | X_i, X_j)$ between random forest regression with and without outliers in the response.

2.6 Discussion

In this paper, we proposed an outlier-robust nonparametric regression approach for modeling between-subject attributes by leveraging the work of random forest regression for within-subject attributes by Wager and Athey [31]. Our method is designed to address two critical issues in current statistical practices: the parametric assumption in FRM-based semiparametric regression models for between-subject attributes and robust response functions against outliers within random forest regression models.

The primary contributions of our work are the development of 1) a class of random forest regression models for between-subject attributes to address the limitations of FRM-based

semiparametric regression models, 2) a response function that mitigates the impact of outliers and enhances the robustness of random forest models in the presence of outliers, and 3) asymptotic properties for the random forest regression models. Our simulation results demonstrate that while random forest regression generally performs well, particularly when the parametric assumption of the FRM-based semiparametric model is violated, its estimates are significantly impacted by outliers when using a response function that is not robust against outliers. In contrast to this, even in the presence of extreme outliers, the proposed rank-based response function effectively resolves this issue and yields reliable estimates. The asymptotic results extend the work by T. Lin [32] on random forest estimators for the Mann-Whitney-Wilcoxon type causal models to the current random forest regression models.

The nonparametric nature of our approach, which allows for greater flexibility and robustness in modeling complex regression relationships between variables without making any parametric assumptions on the mean response, is one of its key strengths. Our method provides more reliable estimates by focusing on outlier robustness, especially in datasets where outliers are present. This robustness is particularly valuable in biomedical and psychosocial research, where outlying observations are not only fairly common, but also should not be discarded, as they are not simply the results of errors in data collection.

Some limitations must be taken into account when using our proposed methods. Although estimations can be more flexible and less susceptible against outliers, they may also exhibit higher variance compared to their semiparametric counterpart. A parametric or semiparametric model should be preferable when responses and predictors have a clear parametric relationship, such as in the simulation example shown in Figure 2.1, since it not only reduces variances but also facilitates interpretation of estimates. Additionally, when there are not enough data points in the leaves to support complex interactions, the model can have trouble capturing the underlying relationships, which can result in poor performance because of data sparsity. Another consideration is that random forest regression can sometimes overfit the data, particularly when the number of trees is excessively large or if individual trees are too deep. Methods such as pruning may help mitigate this issue. These limitations emphasize the importance of carefully evaluating the context and characteristics of the data when applying random forest models in general and our proposed method in particular.

Future research could explore numerous paths to expand upon our findings. One possible approach is to combine our outlier-robust method with other nonparametric techniques to take advantage of these models. Changing the response functions to adapt to different types of data or research questions is also a plausible way to further modify the models. Due to the generalization of our model framework, it not only works on outlier-robust estimation but can also be extended to other types of between-subject attribute problems.

To summarize, our outlier-robust method offers improved accuracy and robustness when handling outliers, making it a significant advancement in random forest regression for biomedical research. Our study offers a solution to address important weaknesses in current methods for nonparametric regression, which helps build more reliable statistical tools for analyzing complex, real-world data.

2.7 Acknowledgements

Chapter 2, in part is currently being prepared for submission for publication of the material. Wu, Tsung-Chin; Lin, Tuo; Schnabl, Bernd; Zhang, Xinlian; Tu, Xin. The dissertation author was the primary investigator and author of this material.

Chapter 3

Estimating Longitudinal Change in Network Transitivity: Application to Viral Genetic Linkage Networks

3.1 Introduction

Connections among individuals drive innovation [42], spread infectious diseases [43], and influence policy decisions [44]. Mathematically, our connections can be represented as a network, where individuals are nodes and connections between two individuals are edges [45]. The structure of the network has been shown to influence the rate of spread of these processes [4]. Therefore, there is a need to understand the network structure and how the structure changes over time. A phenomenon in social behavior that has been shown to influence processes operating on networks is the tendency of individuals to gather, interact, and form cohesive groups [46]. One metric that can be used to quantify this tendency is the level of transitivity in a network, that is, the propensity of two individuals that share a common neighbor to also be connected [47].

Estimates of transitivity have the potential to provide key insights into infectious disease epidemics. Epidemic simulation models have shown that transitivity has the potential to provide insight about epidemic dynamics–and thereby provide evidence of effectiveness of mitigation strategies [4]. Recently, there has been research investigating transitivity in networks constructed based on viral molecular sequence data–referred to as viral genetic linkage (VGL) networks [48].

VGL networks are constructed by connecting individuals with similar viral genetic sequences [49]. Because of ongoing evolution of the virus–within hosts over time and across hosts–the sequences of pairs of individuals who are closer in a transmission chain and are sequenced more closely in time are more likely to be more similar than those of pairs who are further away in the chain or in times of sequencing. Therefore, the level of transitivity can provide insight into the delay between infection and diagnosis (diagnosis delay) [48]. Evaluating diagnosis delay is important for the estimation of incidence [50, 51]. It also provides a metric for evaluating the performance of public health infrastructure with regard to the goal of rapid testing and diagnosing individuals. Doing so can both decrease disease spread [52] and improve the outcomes of infected individuals through the timely provision of treatment [53].

Inference on the level of transitivity in a network, as well as changes in transitivity over time, poses statistical challenges. Transitivity involves modeling the connections among three individuals, which typically cannot be accommodated by traditional statistical models such as generalized linear models (GLMs) [13], generalized estimating equations (GEEs) [54], or generalized linear mixed-effect models (GLMMs) [55]. Moreover, relationships among subjects (such as transitivity) cannot be modeled using the predominant paradigm for within-subject attributes, i.e., models that focus on relationships between an individual's characteristics and their outcome.

This issue becomes more challenging when assessing changes over time. A unique aspect of modeling VGL networks is that they are a type of growing network where links and nodes are added but never removed. As new individuals are sequenced over time, the VGL network grows. Therefore, estimates of transitivity–which are based on all subjects sequenced by that time–change, creating difficulty for inference about transitivity. For example, when comparing transitivity between two time points in a longitudinal study, the transitivity at the later time point involves all subjects sequenced at both time points, making the network at the earlier time a subgraph of the network at the later time. Traditional statistical models for within-subject attributes do not apply to transitivity, even for cross-sectional data, because of

the interlocked and correlated transitivity outcomes. To overcome this difficulty, we leverage semiparametric functional response models (FRMs) [12] to estimate transitivity and model its changes over time. FRMs have been applied to similar between-subject outcomes, such as microbiome beta-diversity data [1]. We develop a new approach to address the challenge of evolving samples when modeling changes in transitivity in VGL networks.

This paper is organized as follows. Section 3.2.1 introduces VGL in greater detail as we use VGL networks as the primary area of interest; however, the statistical methods developed can be applied to many applications. In Section 3.2.2, we introduce notation and formally define transitivity. In Section 3.3.1, we develop the proposed model and its asymptotic properties of model parameters for cross-sectional networks. The extension to longitudinal networks is presented in Section 3.3.2. In Section 3.4, we illustrate both the cross-sectional and longitudinal approaches using simulated data. The simulation study provides information on the performance of the proposed approach for varying network sizes. Section 3.5 presents an analysis using HIV data from San Diego County. Finally, concluding remarks are addressed in the discussion presented in Section 3.6.

3.2 Background

3.2.1 Viral Genetic Linkage Networks

Let $V = \{v_1, ..., v_n\}$ be a set of viral genetic sequences. The pairwise distance between these sequences are denoted by D_{ij} , where $(i, j) \in \binom{S_n}{2}$, $S_n := \{1, ..., n\}$, and $\binom{S}{q}$ represents the set of all *q*-combinations from set *S*. Additionally, we define $S_{\infty} := \{1, 2, ...\}$. Note that $D_{ij} = D_{ji}$. We consider two individuals to be linked if the pairwise distance between their viral genetic sequences is less than some given threshold *c*. Based on *V*, we can construct a VGL network *G*. Since the direction of transmission is generally unknown, we assume that *G* is an undirected network.

3.2.2 Transitivity

We use the global clustering coefficient [56], denoted as θ , as our metric for transitivity, which is defined as the ratio of the number of closed triads (or triangles) to the total number of two-path (either open or closed) in the network. A triad consists of three individuals connected by either two (open triad) or three (closed triad) links [56]. Therefore, a triad among $\{v_i, v_j, v_k\}$ is open if there are exactly two linked pairs such as (v_i, v_j) and (v_i, v_k) , and closed if all three links $(v_i, v_j), (v_i, v_k)$, and (v_j, v_k) are present; see Figure 3.1 for an illustration.



Figure 3.1. Types of triads.

In this paper, our interest is in infinite population inference, that is, we estimate θ based on a network consisting of a subset of individuals *n* from an infinite population [57]. We model the viral genetic sequences $\{v_1, \ldots, v_n\}$ as random variables. The distances D_{ij} are then considered random variables that denote the pairwise distances between these sequences. The transitivity θ can be defined as follows:

$$\theta = \frac{3 \times P\left(\max\{D_{ij}, D_{ik}, D_{jk}\} \le c\right)}{P\left(\max\{D_{ij}, D_{ik}\} \le c\right) + P\left(\max\{D_{ij}, D_{jk}\} \le c\right) + P\left(\max\{D_{ik}, D_{jk}\} \le c\right)},$$
(3.1)

where $(i, j, k) \in \binom{S_{\infty}}{3}$.

We estimate θ using a ratio of two *U*-statistics, which are a class of statistics that is defined as the average value of a kernel function calculated over all possible subsets of a fixed

size (size 3 in our case) of a given set [11]. The kernel functions associated with our two *U*-statistics are the following:

$$h_1(v_i, v_j, v_k) = I\left(\max\left\{D_{ij}, D_{ik}, D_{jk}\right\} \le c\right),$$

$$h_2(v_i, v_j, v_k) = \frac{1}{3}\left(I\left(\max\left\{D_{ij}, D_{ik} \le c\right\}\right) + I\left(\max\left\{D_{ij}, D_{jk} \le c\right\}\right) + I\left(\max\left\{D_{ik}, D_{jk} \le c\right\}\right)\right).$$

Therefore, the estimate of θ can be presented as follows:

$$\widehat{\theta} = \frac{\binom{n}{3}^{-1} \sum_{(i,j,k) \in \binom{S_n}{3}} h_1(v_i, v_j, v_k)}{\binom{n}{3}^{-1} \sum_{(i,j,k) \in \binom{S_n}{3}} h_2(v_i, v_j, v_k)}.$$

In this study, we leverage FRMs to facilitate inferences on θ .

3.2.3 Functional Response Models

To estimate transitivity, we leverage a class of semiparametric FRMs. Let $(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)$ be *n* pairs of independently and identically distributed (i.i.d.) random samples, where \mathbf{x}_i is a $p \times 1$ vector of independent variables and y_i is a response. Then, the general form of an FRM can be defined as follows:

$$E\left[h(y_{i_1},\ldots,y_{i_q}) \mid \mathbf{x}_{i_1},\ldots,\mathbf{x}_{i_q}\right] = f(\mathbf{x}_{i_1},\ldots,\mathbf{x}_{i_q};\boldsymbol{\beta}), \quad (i_1,\ldots,i_q) \in \binom{S_n}{q}, \quad (3.2)$$

where *h* is a real-valued function, *f* is a smooth function that has continuous second-order derivatives, and β is a $p \times 1$ vector of unknown parameters. While we recognize the limitation of FRMs assuming i.i.d. random sampling of nodes–which may not fully capture the network dependence structure–their semiparametric nature offers significant advantages over fully parametric models. Despite this limitation, FRMs provide a valuable alternative to modeling network data, especially
in contexts where interactions can be effectively captured under the i.i.d. assumption through random sampling. A detailed discussion is in Section 3.6.

3.3 Methods

We develop a regression modeling approach using FRMs and its asymptotic properties for conducting inference on transitivity for both cross-sectional and longitudinal network data. We start with cross-sectional network.

3.3.1 Cross-Sectional Network

Let

$$\mathbf{h}_{ijk} = \begin{pmatrix} h_1(v_i, v_j, v_k) \\ h_2(v_i, v_j, v_k) \end{pmatrix}.$$

We define ϕ to be the expectation of $h_2(v_i, v_j, v_k)$, i.e.,

$$\phi = E\left[h_2(v_i, v_j, v_k)\right], \quad (i, j, k) \in \binom{S_n}{3}.$$
(3.3)

Using Equations (3.1) and (3.3), the expectation of $h_1(v_i, v_j, v_k)$ is the following:

$$E\left[h_1(v_i, v_j, v_k)\right] = \theta E\left[h_2(v_i, v_j, v_k)\right] = \theta \phi.$$
(3.4)

To apply the FRM framework in Equation (3.2) and ensure that the range restrictions for θ and ϕ ($0 \le \theta, \phi \le 1$) are satisfied, we can employ the expit, or inverse logit, transformation as follows:

$$\theta = \frac{\exp\left(\beta_1\right)}{\exp\left(\beta_1\right) + 1}, \quad \phi = \frac{\exp\left(\beta_2\right)}{\exp\left(\beta_2\right) + 1},\tag{3.5}$$

where $\beta_1, \beta_2 \in \mathbb{R}$. Consequently, estimating θ corresponds to estimating β_1 via the expit

transformation.

Substituting the formulas for θ and ϕ in Equation (3.5) into Equations (3.3) and (3.4), we can write an FRM in the form of Equation (3.2) using the responses (v_i, v_j, v_k) and the parameters $\boldsymbol{\beta} = (\beta_1, \beta_2)^{\top}$:

$$E\left[\mathbf{h}_{ijk}\right] = \mathbf{f}(\boldsymbol{\beta}) = \begin{pmatrix} f_1(\boldsymbol{\beta}) \\ f_2(\boldsymbol{\beta}) \end{pmatrix}, \quad (i, j, k) \in \begin{pmatrix} S_n \\ 3 \end{pmatrix}, \quad (3.6)$$
$$f_1(\boldsymbol{\beta}) = \frac{\exp\left(\beta_1\right)}{\exp\left(\beta_1\right) + 1} \frac{\exp\left(\beta_2\right)}{\exp\left(\beta_2\right) + 1}, \quad f_2(\boldsymbol{\beta}) = \frac{\exp\left(\beta_2\right)}{\exp\left(\beta_2\right) + 1}.$$

For inference about β , we leverage the *U*-statistics based generalized estimating equations (UGEE) for FRM [12]. The UGEE for the FRM above has the form:

$$\mathbf{U}_{n}(\boldsymbol{\beta},\boldsymbol{\alpha}) = \sum_{(i,j,k)\in\binom{S_{n}}{3}} \mathbf{D}_{ijk}(\boldsymbol{\beta})^{\mathsf{T}} \mathbf{V}_{ijk}(\boldsymbol{\beta},\boldsymbol{\alpha})^{-1} \mathbf{S}_{ijk}(\boldsymbol{\beta}), \qquad (3.7)$$

where

$$\begin{split} \mathbf{S}_{ijk}(\boldsymbol{\beta}) &= \mathbf{h}_{ijk} - \mathbf{f}(\boldsymbol{\beta}), \\ \mathbf{D}_{ijk}(\boldsymbol{\beta}) &= \frac{d}{d\boldsymbol{\beta}} \mathbf{f}(\boldsymbol{\beta}), \\ \mathbf{V}_{ijk}(\boldsymbol{\beta}, \boldsymbol{\alpha}) &= \mathbf{A}(\boldsymbol{\beta})^{1/2} \mathbf{R}(\boldsymbol{\alpha}) \mathbf{A}(\boldsymbol{\beta})^{1/2}, \\ \mathbf{A}(\boldsymbol{\beta}) &= \begin{pmatrix} f_1(\boldsymbol{\beta}) \left(1 - f_1(\boldsymbol{\beta})\right) & 0 \\ 0 & f_2(\boldsymbol{\beta}) \left(1 - f_2(\boldsymbol{\beta})\right) \end{pmatrix}, \end{split}$$

and $\mathbf{R}(\alpha)$ represents a working correlation matrix depending on a $r \times 1$ vector of unknown parameters α . As in the case of GEE for within-subject attributes, choices for the working correlation $\mathbf{R}(\alpha)$ include independent, exchangeable, and autoregressive structures, among others [54]. Within the current setting of binary responses, we may impose the Fréchet bounds on the elements of $\mathbf{R}(\alpha)$, so they logically satisfy the bounds of the variances of binary outcomes such as modeling $\mathbf{R}(\alpha)$ through odds ratios with logistic models and estimate α and β jointly [58]. The choice of \mathbf{R} will not affect consistency and asymptotic distribution of $\hat{\beta}$, even if none of the elements of \mathbf{R} satisfy the Fréchet bounds as long as the estimator for α is \sqrt{n} -consistent [59]. It will only affect the efficiency of $\hat{\beta}$ [21]. For this reason, we choose independent working correlation to reduce the computational burden for estimating α .

Although similar in appearance, Equation (3.7) is not a GEE for GLM, since the summands for $\mathbf{U}_n(\boldsymbol{\beta}, \boldsymbol{\alpha})$, i.e., $\mathbf{D}_{ijk}(\boldsymbol{\beta})^\top \mathbf{V}_{ijk}(\boldsymbol{\beta}, \boldsymbol{\alpha})^{-1} \mathbf{S}_{ijk}(\boldsymbol{\beta})$, are not independent. However, like GEE, the UGEE estimator $\hat{\boldsymbol{\beta}}$ has the same asymptotic properties, as summarized in Theorem 3.1 [1, 60].

Theorem 3.1. Let $\hat{\beta}$ denote the estimate of the parameter β obtained by solving the UGEE in Equation (3.7). Under mild regularity conditions and assuming that $\hat{\alpha}$ is \sqrt{n} -consistent, we have the following:

- 1. $\widehat{\boldsymbol{\beta}}$ is consistent.
- 2. $\widehat{\beta}$ is asymptotically normal:

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}) \xrightarrow{d} \mathcal{N}(0,\boldsymbol{\Sigma}_{\boldsymbol{\beta}}),$$

where

$$\Sigma_{\boldsymbol{\beta}} = \mathbf{B}^{-1} \Sigma_{U} \mathbf{B}^{-1},$$

$$\Sigma_{U} = 3^{2} \operatorname{Var}(\mathbf{u}_{i}), \quad \mathbf{u}_{i} = E \left[\mathbf{D}_{ijk}(\boldsymbol{\beta})^{\top} \mathbf{V}_{ijk}(\boldsymbol{\beta}, \boldsymbol{\alpha})^{-1} \mathbf{S}_{ijk}(\boldsymbol{\beta}) \mid v_{i} \right],$$

$$\mathbf{B} = E \left[\mathbf{D}_{ijk}(\boldsymbol{\beta})^{\top} \mathbf{V}_{ijk}(\boldsymbol{\beta}, \boldsymbol{\alpha})^{-1} \mathbf{D}_{ijk}(\boldsymbol{\beta}) \right], \quad (i, j, k) \in \binom{S_{n}}{3}.$$

3. $\widehat{\beta}$ is asymptotically efficient.

Proof. See Appendix B.3, or Liu, Zhang, et al. [1] and Liu, T. Lin, et al. [60] for a proof.

By assuming independent working correlation, we compute $\hat{\beta}$ by solving $U_n(\beta) = 0$ using the Gauss-Newton method [41]. A consistent estimator of Σ_{β} can then be constructed by substituting $\hat{\beta}$ for β in the results of Theorem 3.1. Specifically,

$$\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}} = \widehat{\mathbf{B}}^{-1} \widehat{\boldsymbol{\Sigma}}_U \widehat{\mathbf{B}}^{-1},$$

where

$$\widehat{\boldsymbol{\Sigma}}_{U} = \frac{9}{n} \sum_{i=1}^{n} \widehat{\boldsymbol{u}}_{i} \widehat{\boldsymbol{u}}_{i}^{\mathsf{T}}, \quad \widehat{\boldsymbol{u}}_{i} = \binom{n-1}{2}^{-1} \sum_{(j,k) \in \binom{S_{n} \setminus \{i\}}{2}} \boldsymbol{D}_{ijk}(\widehat{\boldsymbol{\beta}})^{\mathsf{T}} \mathbf{V}_{ijk}(\widehat{\boldsymbol{\beta}})^{-1} \mathbf{S}_{ijk}(\widehat{\boldsymbol{\beta}}),$$
$$\widehat{\boldsymbol{B}} = \binom{n}{3}^{-1} \sum_{(i,j,k) \in \binom{S_{n}}{3}} \boldsymbol{D}_{ijk}(\widehat{\boldsymbol{\beta}})^{\mathsf{T}} \mathbf{V}_{ijk}(\widehat{\boldsymbol{\beta}})^{-1} \boldsymbol{D}_{ijk}(\widehat{\boldsymbol{\beta}}).$$

Since our goal is to estimate θ , we can readily use the marginal asymptotic distribution of $\hat{\beta}$ for inference about β_1 to obtain the asymptotic distribution of $\hat{\theta}$. A consistent estimate of the asymptotic variance of $\sqrt{n}(\hat{\theta} - \theta)$ can be derived by applying the Delta method, as shown below:

$$\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}} = \left(\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}}\right)_{1,1} \cdot \left(\frac{d}{d\beta_1} \left(\frac{f_1}{f_2}\right) (\boldsymbol{\beta})\Big|_{\beta_1 = \widehat{\beta}_1}\right)^2.$$

3.3.2 Longitudinal Networks

It is common for a viral sequence to be only collected once per individual. In our simulation study and application to San Diego County, a link between two individuals is established if the difference of their first sequences are below a threshold. In particular, a link indicates that two individuals at some point in time had a viral sequences below a threshold. Therefore, in our analyses, once a link is established, it remains in the analysis for all subsequent time points, i.e., our study population and connections only grows over time. In conventional analyses, growing samples generally do not pose new challenges. For example, analysis of population statistics (e.g., mean outcome) for a longitudinal study with n_1 subjects at time 1 and n_2 new subjects added at time 2 can be estimated independently for the two samples. However, for our study of network transitivity, this is not possible, as there can be links across the samples. Since we are interested in changes in transitivity over time, i.e., comparing the transitivity at distinct times, such analyses do create new statistical challenges.

For our longitudinal analysis, we identify a discrete series of *T* VGL networks, denoted as $\{G_1, \ldots, G_T\}$. A network G_t consists of all sequences collected up to time *t*. For our application using HIV molecular data from San Diego, we identify a series of networks such that the networks consist of all sequences up to a given year. Let v_i denote the viral genetic sequence of the *i*-th subject sequenced. Let t_i , where $i \in S_n$ and $1 \le t_i \le T$, index the first network G_{t_i} in which sequence v_i appears. As described earlier, if v_i is present in G_{t_i} , v_i will be present in all subsequent networks G_t ($t_i < t \le T$). Let n_t denote the total number of v_i in network G_t . Note that $n_T = n$. Without loss of generality, we index the sequences such that $v_1, \ldots, v_{\sum_{m=1}^t n_m} \in G_t$ for all *t*.

Because HIV sequences mutate rapidly over time, $E[D_{ij}]$ is more likely to be less than $E[D_{ik}]$ if sequences v_i and v_j are sequenced more closely in time than v_i and v_k . For instance, pairs of sequences sequenced within a year may have more similar mutations than pairs whose time of sequencing is separated by more than a year. Thus, G_T consists of T groups from a growing population, whose distribution of sequences is evolving over time.

We use a time-dependent indicator w_{it} to indicate the presence or absence of v_i in G_t , i.e.,

$$w_{it} = I\left(i \in S_{n_t}\right), \quad i \in S_n, \quad 1 \le t \le T.$$

$$\phi_t = E\left[h_2(v_i, v_j, v_k) \mid w_{it} = w_{jt} = w_{kt} = 1\right], \quad (i, j, k) \in \binom{S_{n_t}}{3}, \quad 1 \le t \le T,$$

represent the probability of a two-path among three individuals at time *t*. Note that, unlike the cross-sectional case in Equation (3.3), ϕ_t is defined conditional on the indicators w_{it} , w_{jt} , and w_{kt} . Similar to Equation (3.4) for the cross-sectional case, we can express the conditional expectation of h_1 in terms of transitivity θ_t at time *t* and ϕ_t as:

$$E\left[h_{1}(v_{i}, v_{j}, v_{k}) \mid w_{it} = w_{jt} = w_{kt} = 1\right] = \theta_{t}E\left[h_{2}(v_{i}, v_{j}, v_{k}) \mid w_{it} = w_{jt} = w_{kt} = 1\right] = \theta_{t}\phi_{t},$$

where $(i, j, k) \in {S_{n_t} \choose 3}$ and $1 \le t \le T$. To address the range restrictions for θ_t and ϕ_t , we transform them using the expit function:

$$\theta_t = \frac{\exp\left(\beta_{1t}\right)}{\exp\left(\beta_{1t}\right) + 1}, \quad \phi_t = \frac{\exp\left(\beta_{2t}\right)}{\exp\left(\beta_{2t}\right) + 1}, \quad 1 \le t \le T,$$

where $\beta_{1t}, \beta_{2t} \in \mathbb{R}$ for all *t*.

 β_{1t} and β_{2t} for each *t* can be estimated independently for each network G_t using our method of cross-sectional networks (Equation (3.6)); however, this approach would not incorporate the correlation among the networks G_1, \ldots, G_T , where $G_s \subseteq G_t$ for s < t. Therefore, we adapt our UGEE framework to the evolving longitudinal sample, which is complicated by the nested networks G_t ($1 \le t \le T$).

In order to formulate the FRM necessary to estimate β_{1t} and β_{2t} for each *t*, we define the following:

Let

$$\boldsymbol{\tau} = (\beta_{11}, \dots, \beta_{1T}, \beta_{21}, \dots, \beta_{2T}, \pi_1, \dots, \pi_{T-1})^\top, \quad \pi_t = E[w_{it}] \quad \text{(in particular, } \pi_T = 1\text{)},$$
$$f_{1t}(\boldsymbol{\tau}) = \frac{\exp(\beta_{1t})}{\exp(\beta_{1t}) + 1} \frac{\exp(\beta_{2t})}{\exp(\beta_{2t}) + 1} \pi_t^3, \quad f_{2t}(\boldsymbol{\tau}) = \frac{\exp(\beta_{2t})}{\exp(\beta_{2t}) + 1} \pi_t^3, \quad \text{and} \quad f_{3t}(\boldsymbol{\tau}) = \pi_t.$$

In addition, we use the law of iterated expectations (LIE) [61] to eliminate the need to condition on (w_{it}, w_{jt}, w_{kt}) as random variables, thereby rendering the inference about θ_t and ϕ_t more tractable:

$$E\left[w_{it}w_{jt}w_{kt}h_{1}(v_{i},v_{j},v_{k})\right] = E\left[w_{it}w_{jt}w_{kt}E\left[h_{1}(v_{i},v_{j},v_{k}) \mid w_{it},w_{jt},w_{kt}\right]\right]$$
$$= \theta_{t}\phi_{t}P\left(w_{it} = w_{jt} = w_{kt} = 1\right)$$
$$= f_{1t}(\boldsymbol{\tau}).$$

Similarly, we have the following:

$$E\left[w_{it}w_{jt}w_{kt}h_2(v_i,v_j,v_k)\right] = \phi_t P\left(w_{it} = w_{jt} = w_{kt} = 1\right)$$
$$= f_{2t}(\boldsymbol{\tau}).$$

By considering the contributions of the triad, observed at time t, as their proportion

within the mixture distribution at time *T*, we specify the following FRM:

$$E\left[\mathbf{h}_{ijk}\right] := E\left[\begin{pmatrix} w_{i1}w_{j1}w_{k1}h_{1}(v_{i},v_{j},v_{k})\\ \vdots\\ w_{i(T-1)}w_{j(T-1)}w_{k(T-1)}h_{1}(v_{i},v_{j},v_{k})\\ h_{1}(v_{i},v_{j},v_{k})\\ w_{i1}w_{j1}w_{k1}h_{2}(v_{i},v_{j},v_{k})\\ \vdots\\ w_{i(T-1)}w_{j(T-1)}w_{k(T-1)}h_{2}(v_{i},v_{j},v_{k})\\ h_{2}(v_{i},v_{j},v_{k})\\ \frac{1}{3}(w_{i1}+w_{j1}+w_{k1})\\ \vdots\\ \frac{1}{3}(w_{i(T-1)}+w_{(T-1)}+w_{i(T-1)})\end{pmatrix}\right] = \begin{pmatrix} f_{11}(\tau)\\ \vdots\\ f_{1(T-1)}(\tau)\\ f_{21}(\tau)\\ \vdots\\ f_{2(T-1)}(\tau)\\ f_{2T}(\tau)\\ f_{31}(\tau)\\ \vdots\\ f_{3(T-1)}(\tau) \end{pmatrix}$$
(3.8)

Using Equation (3.8), inference about τ can be readily performed using the following UGEE:

$$\mathbf{U}_{n}(\boldsymbol{\tau},\boldsymbol{\alpha}) = \sum_{(i,j,k)\in\binom{S_{n}}{3}} \mathbf{D}_{ijk}(\boldsymbol{\tau})^{\top} \mathbf{V}_{ijk}(\boldsymbol{\tau},\boldsymbol{\alpha})^{-1} \mathbf{S}_{ijk}(\boldsymbol{\tau}), \qquad (3.9)$$

where

$$\mathbf{S}_{ijk}(\tau) = \mathbf{h}_{ijk} - \mathbf{f}(\tau), \quad \mathbf{D}_{ijk}(\tau) = \frac{\partial}{\partial \tau} \mathbf{f}(\tau), \quad \mathbf{V}_{ijk}(\tau, \alpha) = \mathbf{A}(\tau)^{1/2} \mathbf{R}(\alpha) \mathbf{A}(\tau)^{1/2},$$

$$\begin{pmatrix} f_{11}(\tau) \left(\pi_{1}^{3} - f_{11}(\tau)\right) \\ \vdots \\ f_{1(T-1)}(\tau) \left(\pi_{T-1}^{3} - f_{1(T-1)}(\tau)\right) \\ f_{1T}(\tau) (1 - f_{1T}(\tau)) \\ f_{21}(\tau) \left(\pi_{1}^{3} - f_{21}(\tau)\right) \\ \vdots \\ f_{2(T-1)}(\tau) \left(\pi_{T-1}^{3} - f_{2(T-1)}(\tau)\right) \\ f_{31}(\tau) (1 - f_{31}(\tau)) \\ \vdots \\ f_{3(T-1)}(\tau) (1 - f_{3(T-1)}(\tau)) \end{pmatrix}$$

As in the cross-sectional network, we can choose $\mathbf{R}(\alpha)$ as the independent working correlation to reduce the computational burden for estimating α and apply the Gauss-Newton method to estimate τ . The UGEE estimator $\hat{\tau}$ has similar nice asymptotic properties as the one for the cross-sectional data case, as summarized by the following theorem.

Theorem 3.2. Let $\hat{\tau}$ denote the estimate of the parameter τ obtained by solving the UGEE in Equation (3.9). We assume that $\lim_{n\to\infty} \frac{n}{n_t} < \infty$ for $1 \le t \le T - 1$. Then, under mild regularity conditions and assuming that $\hat{\alpha}$ is \sqrt{n} -consistent, we have:

- 1. $\hat{\tau}$ is consistent.
- 2. $\hat{\tau}$ is asymptotically normal:

$$\sqrt{n}(\widehat{\boldsymbol{\tau}}-\boldsymbol{\tau}) \xrightarrow{d} \mathcal{N}(0,\boldsymbol{\Sigma}_{\boldsymbol{\tau}}),$$

where

$$\Sigma_{\tau} = \mathbf{B}^{-1} \Sigma_{U} \mathbf{B}^{-1},$$

$$\Sigma_{U} = 3^{2} \operatorname{Var}(\mathbf{u}_{i}), \quad \mathbf{u}_{i} = E \left[\mathbf{D}_{ijk}(\tau)^{\top} \mathbf{V}_{ijk}(\tau, \alpha)^{-1} \mathbf{S}_{ijk}(\tau) \mid v_{i}, w_{i1}, \dots, w_{i(T-1)} \right],$$

$$\mathbf{B} = E \left[\mathbf{D}_{ijk}(\tau)^{\top} \mathbf{V}_{ijk}(\tau, \alpha)^{-1} \mathbf{D}_{ijk}(\tau) \right], \quad (i, j, k) \in \binom{S_{n}}{3}.$$

3. $\hat{\tau}$ *is asymptotically efficient.*

Proof. See Appendix B.3 for a proof.

As in the case of cross-sectional network, a consistent estimator of Σ_{τ} can be readily obtained based on the results in Theorem 3.2, assuming independent working correlation. Specifically,

$$\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\tau}} = \widehat{\mathbf{B}}^{-1} \widehat{\boldsymbol{\Sigma}}_U \widehat{\mathbf{B}}^{-1},$$

where

$$\begin{split} \widehat{\boldsymbol{\Sigma}}_{U} &= \frac{9}{n} \sum_{i=1}^{n} \widehat{\mathbf{u}}_{i} \widehat{\mathbf{u}}_{i}^{\top}, \quad \widehat{\mathbf{u}}_{i} = \binom{n-1}{2}^{-1} \sum_{(j,k) \in \binom{S_{n} \setminus \{i\}}{2}} \mathbf{D}_{ijk}(\widehat{\boldsymbol{\tau}})^{\top} \mathbf{V}_{ijk}(\widehat{\boldsymbol{\tau}})^{-1} \mathbf{S}_{ijk}(\widehat{\boldsymbol{\tau}}), \\ \widehat{\mathbf{B}} &= \binom{n}{3}^{-1} \sum_{(i,j,k) \in \binom{S_{n}}{3}} \mathbf{D}_{ijk}(\widehat{\boldsymbol{\tau}})^{\top} \mathbf{V}_{ijk}(\widehat{\boldsymbol{\tau}})^{-1} \mathbf{D}_{ijk}(\widehat{\boldsymbol{\tau}}). \end{split}$$

Similar to the cross-sectional case, we can obtain the asymptotic distribution of $\hat{\theta} := (\hat{\theta}_1, \dots, \hat{\theta}_T)^{\top}$ following the marginal asymptotic distribution of $\hat{\tau}$. A consistent estimate of the asymptotic variance of $\sqrt{n}(\hat{\theta} - \theta)$ can be derived by applying the Delta method:

$$\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}} = \left(\frac{\partial}{\partial \boldsymbol{\beta}_1} \mathbf{g}(\boldsymbol{\tau}) \Big|_{\boldsymbol{\beta}_1 = \widehat{\boldsymbol{\beta}}_1} \right) \cdot \left(\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\tau}} \right)_{1:T,1:T} \cdot \left(\frac{\partial}{\partial \boldsymbol{\beta}_1} \mathbf{g}(\boldsymbol{\tau}) \Big|_{\boldsymbol{\beta}_1 = \widehat{\boldsymbol{\beta}}_1} \right)^{\top},$$

where $\mathbf{M}_{1:T,1:T}$ denotes the $T \times T$ submatrix consisting of the first T rows and first T columns of a matrix $\mathbf{M}, \mathbf{g} := \left(\frac{f_{11}}{f_{21}}, \dots, \frac{f_{1T}}{f_{2T}}\right)^{\top}$, and $\boldsymbol{\beta}_1 := (\beta_{1t}, \dots, \beta_{1T})^{\top}$.

3.4 Simulation Studies

In this section, we illustrate the validity of the proposed approach with simulated data. We investigate the performance of FRM for cross-sectional and longitudinal networks of varying network sizes. The analyses are carried out using code developed on the R software platform [20]. Section 3.5 investigates VGL network data from San Diego County.

3.4.1 Simulation Method

3.4.1.1 Data Generation for Cross-Sectional Analysis

We propose the following approach to generate data to investigate our approach for estimating transitivity for cross-sectional network data. First, we generate a population of size *n*. For ease of presentation, instead of simulating viral sequences, we generate a latent variable z_i for each individual $i \in S_n$. We sample z_i from a generalized extreme value (GEV) distribution:

$$z_i \sim \operatorname{GEV}(\mu, \sigma, \xi), \quad \mu \in \mathbb{R}, \quad \sigma \in \mathbb{R}^{>0}, \quad \xi \in \mathbb{R}, \quad i \in S_n,$$

where μ is the location parameter, σ is the scale parameter, and ξ is the shape parameter. The GEV distribution represents a heavy-tailed distribution that combines the Gumbel, Frechet, and Weibull distributions [62]. The pairwise distance between individuals *i* and *j*, D_{ij} , is set as $|z_i - z_j|$.

In order to simulate networks closely resembling San Diego County VGL network in terms of the degree distribution, we use the following specifications for μ, σ , and ξ .

• μ : As D_{ij} always has a mean of 0 regardless of the value of μ , we can assume $\mu = 0$ in this simulation setting without a loss of generality.

σ and ξ: By tuning the σ and ξ parameters for the GEV distribution, we can achieve transitivity values similar to those calculated from the San Diego molecular epidemiology data. Specifically, we choose σ = 2 and ξ = −0.5 to produce a transitivity similar to that of the real network from 2018. Note that ξ determines the tail behavior of the distribution: for ξ < 0, the distribution has a heavy right tail, and for ξ > 0, it has a heavy left tail. In our setting, we assume that the values of z_i can be any positive numbers, so we choose negative ξ values.

We use the kernel functions $h_1(z_i, z_j, z_k) = I(\max\{D_{ij}, D_{ik}, D_{jk}\} \le c)$ and $h_2(z_i.z_j, z_k) = (I(\max\{D_{ij}, D_{ik}\} \le c) + I(\max\{D_{ij}, D_{jk}\} \le c) + I(\max\{D_{ik}, D_{jk}\} \le c))/3$ to identify whether a triad $\{z_i, z_j, z_k\}$ forms an open triad, closed triad, or neither (i.e., not a two-path).

3.4.1.2 Data Generation for Longitudinal Analysis

For longitudinal network, we analyze only the first viral sequence collected for each individual; therefore, once a link is established, it will remain over time. This implies that the network grows larger over time, with the networks in previous time periods forming an induced subgraph of the current network. We generate a population of size n_t at each time t. Similar to the cross-sectional case, we also generate a latent variable z_{it} for each individual $i \in S_{n_t} \setminus S_{n_{t-1}}$ (assuming S_{n_0} is the empty set) for ease of presentation. We sample z_{it} from a time dependent GEV distribution in chronological order:

$$z_{it} \sim \text{GEV}(\mu_t, \sigma_t, \xi_t), \quad \mu_t \in \mathbb{R}, \quad \sigma_t \in \mathbb{R}^{>0}, \quad \xi_t \in \mathbb{R}, \quad i \in S_{n_t} \setminus S_{n_{t-1}}, \quad 1 \le t \le T.$$

The specifications for μ_t, σ_t , and ξ_t are as follows:

• μ_t : Similar to the cross-sectional analysis, we assume $\mu_t = 0$ for all *t* in our simulation. The parameters σ_t and ξ_t are used to modify transitivity across the longitudinal networks (see below). σ_t and ξ_t: For the second and third time points, we set σ₂ = 3, ξ₂ = −1; σ₃ = 2, ξ₃ = −0.5 to ensure that the resulting transitivity values are close to the transitivity value in 2022. For the first time point, we select parameters σ₁ = 12 and ξ₁ = −3 to generate sufficiently higher transitivity, allowing us to evaluate the performance of our method in detecting differences in transitivity.

The pairwise distance between individuals *i* and *j*, D_{ij} , is set as $|z_{it} - z_{js}|$, where $i \in S_{n_t} \setminus S_{n_{t-1}}$ and $j \in S_{n_s} \setminus S_{n_{s-1}}$, respectively. Then, we establish genetic similarity between individuals in the same way as in the cross-sectional case.

3.4.1.3 Simulation Assessment

We perform M = 1000 Monte Carlo (MC) simulation replicas for each network size. For each MC replication, the transitivity θ (θ_t for the longitudinal case) and its associated asymptotic variance Σ_{θ} (Σ_{θ_t}) are estimated based on the marginal asymptotic distribution of FRM estimates $\hat{\beta}$ ($\hat{\tau}$). A key focus is on evaluating whether the FRM model provides an accurate estimate of θ (θ_t), which is assessed by testing the following hypotheses at a statistical significance level of 0.05:

Cross-sectional:
$$H_0: \theta = \theta_{(0)}$$
 vs. $H_a: \theta \neq \theta_{(0)}$,
Longitudinal: $H_0: \theta_t = \theta_{(0)t}$ vs. $H_a: \theta_t \neq \theta_{(0)t}$,

where $\theta_{(0)}$ ($\theta_{(0)t}$) denotes the "true transitivity" generated from a given GEV distribution (mixture GEV distribution). To assess the estimation of our variance, we compare the estimated asymptotic variance, $\frac{1}{n}\widehat{\Sigma}_{\theta}$ or $\frac{1}{n}\widehat{\Sigma}_{\theta t}$ (longitudinal), to the empirical variance:

Cross-sectional:
$$\frac{1}{M-1} \sum_{m=1}^{M} \left(\widehat{\theta}^{(m)} - \overline{\widehat{\theta}}\right)^2$$
,
Longitudinal: $\frac{1}{M-1} \sum_{m=1}^{M} \left(\widehat{\theta}^{(m)}_t - \overline{\widehat{\theta}}_t\right)^2$,

where $\overline{\hat{\theta}} = \frac{1}{M} \sum_{m=1}^{M} \widehat{\theta}^{(m)}$ and $\overline{\hat{\theta}}_t = \frac{1}{M} \sum_{m=1}^{M} \widehat{\theta}_t^{(m)}$. We present the results in terms of standard errors rather than variances to facilitate interpretation, as standard errors share the same units as the estimated transitivity.

In the longitudinal case, we are also interested in testing whether θ_t changes over time at a statistical significance level of 0.05 using the following hypothesis:

$$H_0: \theta_1 = \ldots = \theta_T$$
 vs. H_a : There exists t_1 and t_2 such that $\theta_{t_1} \neq \theta_{t_2}$.

If the above omnibus test is rejected at the significance level of 0.05, we can further apply multiple comparisons to test which pairs are different from the others. False discovery rate (FDR) correction is applied to avoid the Type I error inflation issue for multiple comparisons.

Wald statistics are calculated for each of the hypotheses listed above and are compared with the corresponding Chi-square distribution to compute the empirical Type I error and examine the coverage.

3.4.1.4 Network Sizes and Threshold

For cross-sectional networks, results are reported for three network sizes: n = 150, 250, and 500, representing small, medium, and large sizes, respectively. To simulate longitudinal network, we assume that we observe the network at three time points, with the final observation contains either 150, 250, and 500 nodes. The percentage of additional nodes included at each of the three time points correspond to: 40%, 30%, and 30% of the final network size, respectively. We use a sequence similarity cutoff of c = 1.5 in all the simulation cases.

3.4.2 Simulation Results

For the cross-sectional networks, the simulated transitivity was 0.8068 (see Table 3.1). The estimated transitivity values (based on the presented approach) approximate this value closely. The alignment of asymptotic and empirical standard errors indicates reliable estimation under theoretical assumptions and simulation conditions. Type I error rates demonstrate acceptable levels of false positives across different node sizes, underscoring the robustness of the statistical approach used. These findings suggest that the method effectively captures network transitivity at the fixed time point, supporting its applicability in studying real networks where accurate estimation of transitivity is crucial for understanding network properties.

Table 3.1. Estimated transitivity, standard errors, and Type I errors in simulated cross-sectional networks.

Simulated	Number of	Estimated	Asymp.	Emp.	Type I
Transitivity	Nodes (n)	Transitivity	Std. Err.	Std. Err.	Error
	150	0.8054	0.0156	0.0144	0.074
0.8068	250	0.8061	0.0116	0.0111	0.041
	500	0.8066	0.0080	0.0079	0.067

For the longitudinal networks, the estimated transitivity values approximate the simulated transitivity values across varying population sizes and recruitment rates (see Table 3.2). In terms of variance analysis, both asymptotic and empirical standard errors decrease as node sizes increase, indicating robust estimation under various conditions. Larger node sizes reduce the discrepancy between asymptotic and empirical standard errors, especially when nodes are sampled from a more complex mixture distribution. Type I error rates remain close to the significance level as population sizes increase, indicating that the statistical approach maintains reliable performance in controlling false positives. These findings underscore the method's ability to effectively capture and estimate network transitivity over time, highlighting its utility in analyzing longitudinal network data where understanding temporal changes in network properties

is crucial.

Table 3.2.	Estimated	transitivity,	standard	errors,	and	Type 1	I errors	in sin	nulated	longitu	dinal
networks.											

Time (t)	Simulated	Total Recruit-	Number of	Estimated	Asymp.	Emp.	Type I
	Transitivity	ment Rate	Nodes (n_t)	Transitivity	Std. Err.	Std. Err.	Error
			60	0.8784	0.0506	0.0509	0.106
1	0.8853	40%	100	0.8829	0.0379	0.0381	0.077
			200	0.8848	0.0263	0.0266	0.073
			105	0.7819	0.0257	0.0196	0.044
2	0.7820	70%	175	0.7816	0.0161	0.0132	0.057
			350	0.7827	0.0094	0.0080	0.038
			150	0.7829	0.0187	0.0159	0.058
3	0.7816	100%	250	0.7830	0.0132	0.0121	0.058
			500	0.7837	0.0087	0.0082	0.048

Table 3.3 shows the performance of our method in detecting differences in transitivity by assessing the power of the omnibus test detailed in the previous section. For a transitivity difference of approximately 0.1, a sample size of n = 500 achieves an 84.2% power. This indicates that our method is highly effective at detecting significant differences in transitivity with a sufficiently large sample size, demonstrating its robustness and reliability for network analysis.

Table 3.3. Power of the omnibus test for determining whether the transitivity changes over time in simulated longitudinal networks.

Number of	Power for
Total Nodes	Omnibus Test
150	0.200
250	0.486
500	0.842

3.5 HIV Transmission in San Diego County

We apply our proposed approach to the San Diego molecular epidemiology data to infer transitivity and examine changes in the transitivity index over time. The study utilized HIV molecular data from people with HIV in San Diego County between 1985 and 2022. In this dataset, genetic distance between individuals is defined based on sequence similarity, with a cutoff of 1.5% used to establish genetic similarity between individuals [49, 63]. To ensure an adequate sample size for forming both closed and opened triads and to maintain the power of our methodology, we analyze the VGL network starting from 2013, which includes data spanning from 1985 to 2013. We investigate changes in transitivity over time until 2022, covering 10 time points (in years). Nodes with cluster sizes smaller than 3 are excluded from the analysis, as they do not affect the calculation of transitivity, thus reducing the computational burden. After this removal, the total number of nodes in the 2022 network is 1811. Table 3.4 presents the results of applying our method to the longitudinal network data. Our approach accurately estimates transitivity according to the regular definition (Equation (3.1) with finite population). The asymptotic standard errors are small and decrease over time, indicating the robustness of our method. This is further supported by the 95% confidence intervals shown in Figure 3.2. The omnibus test reveals that there are at least two time points where the transitivity values differ significantly at the 0.05 significance level (p-value $\approx 5.32 \times 10^{-76}$). Additionally, the change in transitivity between 2013 and 2022 is also significant (p-value $\approx 1.06 \times 10^{-31}$).

3.6 Discussion

In this paper, we provide a rigorous method for estimating the level of transitivity in both cross-sectional and longitudinal networks. Our method extends the FRM framework to accommodate longitudinal data, specifically data with evolving samples. Simulation studies demonstrate that our approach provides accurate estimates of both point estimates and variances.

When modeling transitivity in longitudinal network data, Exponential Random Graph

Year	Number of	Total Recruit-	Transitivity	Estimated	Asymp.
	Nodes	ment Rate	(regular definition)	Transitivity	Std. Err.
2013	602	33.24%	0.5388	0.5388	0.0206
2014	765	42.24%	0.6643	0.6643	0.0139
2015	914	50.47%	0.7238	0.7238	0.0103
2016	1108	61.18%	0.7881	0.7881	0.0076
2017	1257	69.41%	0.8096	0.8096	0.0064
2018	1388	76.64%	0.8066	0.8066	0.0054
2019	1513	83.55%	0.7950	0.7950	0.0050
2020	1633	90.17%	0.7967	0.7967	0.0047
2021	1733	95.69%	0.7847	0.7847	0.0046
2022	1811	100.00%	0.7807	0.7807	0.0045

Table 3.4. Estimated transitivity and asymptotic standard errors for the San Diego molecular epidemiology data.

Models (ERGMs) have been widely used due to their flexibility in capturing complex dependencies between nodes, including transitivity [64]. However, ERGMs rely on fully parametric assumptions [65], which can be limiting when the true data-generating process does not conform to these specified distributions [66]. In contrast, our method, FRMs, adopts a semiparametric approach, allowing for greater flexibility in capturing the underlying structure of the data without strict adherence to a predetermined parametric form. Alternative network models, such as the congruence class model [67], do not require a parameter assumption and may offer a promising direction.

Nevertheless, the use of FRMs for effective inference requires the assumption that viral genetic sequences are i.i.d. random samples. While this assumption simplifies the modeling process by assuming independent sampling from a superpopulation, it may not fully account for the complex dependencies inherent in infectious disease data. This limitation highlights the importance of carefully considering the underlying population structure to ensure that the parameters of interest (targets or estimands) effectively capture the true dependency relationships within the network while providing reliable variance estimates.



Figure 3.2. Temporal changes in transitivity for the San Diego molecular epidemiology data, with 95% confidence intervals.

We applied our approach to investigate HIV disease dynamics in San Diego County. Our analysis demonstrates the effectiveness of our proposed method in estimating transitivity and detecting significant changes in the transitivity index from 2013 to 2022 within the San Diego HIV cohort. Our approach provides asymptotic standard errors, which are crucial for making robust inferences. This added capability allows us to not only estimate transitivity with precision but also to assess the statistical significance of changes over time. Our results indicate that transitivity is increasing over time, suggesting that the delay between infection and diagnosis is decreasing. However, if only a subset of the HIV-positive population is sequenced, our data may not fully represent the overall transmission dynamics, potentially leading to skewed results.

Although HIV disease dynamics in San Diego County is the primary motivation for developing our statistical approach, it might also be profitably applied in a variety of other contexts. One such application is the investigation of the evolution of scientific paper citation networks over time. Such analyses aim to reveal temporal changes in the extent to which investigators cite primarily those who share their perspectives on contentious issues, as opposed to including citations from those with differing views. It has been widely noted that developments in communication technology appear to intensify partisanship by enabling individuals to virtually inhabit "echo chambers" [68]. However, to our knowledge, the impact of this phenomenon on scientific literature has not received equivalent attention. While the phenomenon of publication bias is well understood, the concept of "citation network bias" is less thoroughly explored, although it may similarly inhibit the dissemination of knowledge to contexts where it could be most beneficial.

Promising extensions of the approach include controlling for covariates in the FRMs used to estimate transitivity. This adjustment would enable us to account for potential confounding variables and isolate the specific effects of interest on transitivity. By incorporating covariates such as demographic characteristics, geographic factors, or specific intervention types, we can mitigate the influence of external factors that might otherwise skew our results. Another potential extension is the incorporation of incomplete *U*-statistics theory [69] when solving the UGEE in Equations (3.7) and (3.9). This approach could alleviate the computational burden associated with large network sizes. For networks containing over a thousand nodes, calculating all combinatorial arrangements of triads becomes impractical due to significant computational demands and limited storage capacity. Incomplete *U*-statistics address this challenge by utilizing a subset of combinatorial terms while preserving essential information through appropriate subsampling. However, a potential drawback is that some connections may be omitted in the incomplete *U*-statistics approach, which could distort the underlying dependency structure of the network. While our current framework provides a robust foundation for estimating transitivity in dynamic networks, these extensions represent important avenues for further refinement. By

accounting for confounding variables and addressing the computational challenges, we can enhance the accuracy and applicability of our methods across diverse network settings.

3.7 Acknowledgements

Chapter 3, in part is currently being prepared for submission for publication of the material. Wu, Tsung-Chin; Nguyen, Kevin; Lin, Tuo; Liu, Jinyuan; De Gruttola, Victor; Little, Susan J.; Tu, Xin; Goyal, Ravi. The dissertation author was the primary investigator and author of this material. This work was supported in part by the National Institute of Allergy and Infectious Diseases (P30 AI036214, R01 AI147441, and R01 MH-132151) and the James B. Pendleton Charitable Trust. This manuscript was also supported by the Health Resources and Services Administration (HRSA) of the U.S. Department of Health and Human Services (HHS) as part of award UT8HA33959. The contents are those of the author(s) and do not necessarily represent the official views of, nor an endorsement, by HRSA, HHS or the U.S. Government Health Resources and Services and Services Administration (HRSA).

Chapter 4 Conclusions and Future Work

This dissertation presents a series of new advanced statistical methods aimed at addressing complex challenges associated with modeling and inference for between-subject attributes in biomedical and public health research. Through three distinct yet interconnected projects, we have made significant contributions to the fields of statistical methodological research and the biomedical and public health domains that have motivated the development of the new statistical methods.

In Chapter 1, we have developed a rigorous approach to derive the asymptotic distribution of the empirical area under the receiver operating characteristic (AUROC) curve when parameters of class probability models are estimated by maximum likelihood estimators (MLEs). Our results demonstrate that the variability introduced by MLE diminishes asymptotically, indicating that the distribution of the empirical AUROC with MLE substituting in place of the true parameters, or plug-in estimator, closely approximates that of the AUROC with the true values of the parameters. The findings show that the impact of the sampling variability from the MLE on the empirical AUROC diminishes to zero as the sample size increases. The established asymptotic results offer a reliable approximation for small to moderate sample sizes to facilitate inference in practice. Future research could extend this work to multi-class AUROC and other types of consistent estimators such as those from regularization techniques to further enhance the robustness of AUROC estimation. Chapter 2 introduces an outlier-robust nonparametric regression approach for modeling between-subject attributes, enhancing the robustness of random forest models against outliers and improving estimation accuracy in the presence of outliers. The key contributions include the development of a rank-based response function for better outlier handling, the extension of random forest regression models to between-subject attributes, and the derivation of asymptotic properties for the model estimators. Future research could focus on integrating this method with other nonparametric techniques and applying it to diverse domains of biomedical and public health research to improve estimation accuracy and inference validity.

In Chapter 3, we address the challenges of estimating transitivity in evolving networks, particularly within the context of longitudinal data. By extending the semiparametric functional response models (FRMs) framework to network settings, our approach effectively addresses the complex population-mixture distribution of the FRM-based transitivity estimator due to evolving samples over time, offering insights into HIV transmission dynamics in San Diego County when applied to the viral genetic linkage data from the County. The results revealed significant temporal variations in transitivity, reflecting changes in the transmission network structure over time. Future research could enhance this method by incorporating covariates to account for potential confounding factors and by addressing computational challenges with incomplete *U*-statistics when applied to large samples. Additionally, application of this approach to other contexts, such as citation network analysis, could provide valuable insights into the dynamics of information dissemination and potential biases in scientific literatures.

Overall, the findings of this dissertation have not only significantly advanced statistical methodological research for complex biomedical and public health data modeling, but also paved the way for future research that can build upon the foundations of the work to explore extensions to tackle new challenges arising from biomedical and public health research.

Appendix A Relevant Existing Theorems

In this appendix, we state the existing theorems we have used in the previous chapters.

A.1 Stochastic Convergence

The following theorems are stated in Van der Vaart [11].

Theorem A.1. (*Continuous mapping*). Let $g : \mathbb{R}^k \to \mathbb{R}^m$ be continuous at every point of a set C such that $P(X \in C) = 1$.

- (i) If $X_n \xrightarrow{d} X$, then $g(X_n) \xrightarrow{d} g(X)$;
- (*ii*) If $X_n \xrightarrow{p} X$, then $g(X_n) \xrightarrow{p} g(X)$;
- (iii) If $X_n \stackrel{a.s.}{\rightarrow} X$, then $g(X_n) \stackrel{a.s.}{\rightarrow} g(X)$.

Proof. See Theorem 2.3 in Van der Vaart [11].

Theorem A.2. Let X_n and X be random vectors. Then

- (i) $X_n \xrightarrow{a.s.} X$ implies $X_n \xrightarrow{p} X$; (ii) $X_n \xrightarrow{p} X$ implies $X_n \xrightarrow{d} X$;
- (iii) $X_n \xrightarrow{p} c$ for a constant c if and only if $X_n \xrightarrow{d} c$.

 \Box

Theorem A.3. (*Slutsky*). Let X_n, X and Y_n be random vectors or variables. If $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{d} c$ for a constant c, then

(i) $X_n + Y_n \xrightarrow{d} X + c;$

(*ii*)
$$Y_n X_n \xrightarrow{d} c X$$
;

(*iii*) $Y_n^{-1}X_n \xrightarrow{d} c^{-1}X$ provided $c \neq 0$.

Proof. See Lemma 2.8 in Van der Vaart [11].

A.2 Two-Sample U-Statistics

The following statements and theorem are stated in Van der Vaart [11].

Suppose the observations consist of two independent samples X_1, \ldots, X_m and Y_1, \ldots, Y_n , i.i.d. within each sample, from possibly different distributions. Let $h(x_1, \ldots, x_r; y_1, \ldots, y_s)$ be a known function that is permutation symmetric in x_1, \ldots, x_r and y_1, \ldots, y_s separately. A two-sample *U*-statistic with kernel *h* has the form

$$U = \frac{1}{\binom{m}{r}\binom{n}{s}} \sum_{\alpha} \sum_{\beta} h(X_{\alpha_1}, \dots, X_{\alpha_r}; Y_{\beta_1}, \dots, Y_{\beta_s}),$$

where α and β range over the collections of all subsets of *r* different elements from $\{1, 2, ..., m\}$ and of *s* different elements from $\{1, 2, ..., n\}$, respectively. Clearly, *U* is an unbiased estimator of the parameter

$$\theta = E\left[h(X_1,\ldots,X_r;Y_1,\ldots,Y_s)\right].$$

If N = m + n is the total number of observations, we assume that, as $m, n \to \infty$,

$$\frac{m}{N} \to \lambda, \quad \frac{n}{N} \to 1 - \lambda, \quad 0 < \lambda < 1.$$

The projection of $U - \theta$ onto the set of all functions of the form $\sum_{i=1}^{m} k_i(X_i) + \sum_{j=1}^{n} l_j(Y_j)$, for arbitrary measurable functions k_i and l_j with $E[k_i^2(X_i)], E[l_j^2(Y_j)] < \infty$, is given by

$$\widehat{U} = \frac{r}{m} \sum_{i=1}^{m} h_{1,0}(X_i) + \frac{s}{n} \sum_{j=1}^{n} h_{0,1}(Y_j),$$

where the functions $h_{1,0}$ and $h_{0,1}$ are defined by

$$h_{1,0}(x) = E [h(x, X_2, \dots, X_r, Y_1, \dots, Y_s)] - \theta,$$

$$h_{0,1}(y) = E [h(X_1, \dots, X_r, y, Y_2, \dots, Y_s)] - \theta.$$

If the kernel is square-integrable, then the sequence \widehat{U} is asymptotically normal by the central limit theorem. The difference between \widehat{U} and $U - \theta$ is asymptotically negligible.

Theorem A.4. If $E\left[h^2(X_1,...,X_r;Y_1,...,Y_s)\right] < \infty$, then the sequence $\sqrt{N}(U-\theta-\widehat{U})$ converges in probability to zero. Consequently, the sequence $\sqrt{N}(U-\theta)$ converges in distribution to the normal law with mean zero and variance $r^2\zeta_{1,0}/\lambda + s^2\zeta_{0,1}/(1-\lambda)$, where, with X_i being i.i.d. variables independent of the i.i.d. variables Y_i ,

$$\zeta_{c,d} = \operatorname{Cov}\left(h(X_1, \dots, X_r; Y_1, \dots, Y_s), h(X_1, \dots, X_c, X'_{c+1}, \dots, X'_r; Y_1, \dots, Y_d, Y'_{d+1}, \dots, Y'_s)\right).$$

Proof. See Theorem 12.6 in Van der Vaart [11].

A.3 Empirical Processes

The following statements, definitions, theorems, and lemma are stated in Van der Vaart [11] and Van der Vaart and Wellner [16].

Let \mathbb{P}_n be the empirical distribution of an i.i.d. sample X_1, \ldots, X_n from a probability distribution P on a measurable space (X, \mathcal{A}) . The empirical distribution \mathbb{P}_n is the discrete uniform measure that puts mass 1/n at each observation if points are measurable. We express the empirical distribution as the linear combination $\mathbb{P}_n = n^{-1} \sum_{i=1}^n \delta_{X_i}$ of the Dirac measures at the observations. Given a measurable function $f : X \to \mathbb{R}$, we write $\mathbb{P}_n f$ for the expectation of funder the empirical distribution, and Pf for the expectation under P. Thus

$$\mathbb{P}_n f = \frac{1}{n} \sum_{i=1}^n f(X_i), \quad Pf = \int f dP.$$

The empirical process evaluated at f is defined as $\mathbb{G}_n f = \sqrt{n} (\mathbb{P}_n f - Pf)$. We mention some definitions before delving into the empirical process theorems.

Definition A.5. A Borel probability measure L is tight if for every $\varepsilon > 0$ there exists a compact set K with $L(K) \ge 1 - \varepsilon$. A Borel measurable map $X : \Omega \to \mathbb{D}$ is called tight if its law $\mathcal{L}(X) = P \circ X^{-1}$ is tight. This is equivalent to there being a countable union of compact sets that has probability I under L or X.

Definition A.6. A Borel probability measure L is separable if there is a separable [70], measurable set with probability 1. A Borel measurable map $X : \Omega \to \mathbb{D}$ is called separable if its law $\mathcal{L}(X) = P \circ X^{-1}$ is separable.

Since a countable union of compact sets in a metric space is separable, separability is slightly weaker than tightness.

A.3.1 Brownian Bridge

To discuss a uniform version of the central limit theorem, let \mathcal{F} be a class of measurable functions such that

$$\sup_{f \in \mathcal{F}} |f(x) - Pf| < \infty, \quad \text{ for every } x.$$

Under this condition the empirical process $\{\mathbb{G}_n f : f \in \mathcal{F}\}\$ can be viewed as a map in $\ell^{\infty}(\mathcal{F}) := \{z : \mathcal{F} \to \mathbb{R} \mid ||z||_{\mathcal{F}} := \sup_{f \in \mathcal{F}} |z(f)| < \infty\}$. Consequently, it makes sense to investigate conditions under which

$$\mathbb{G}_n = \sqrt{n} \left(\mathbb{P}_n - P \right) \xrightarrow{d} \mathbb{G}, \quad \text{in } \ell^{\infty}(\mathcal{F}), \tag{A.1}$$

where the limit \mathbb{G} is a tight Borel measurable element in $\ell^{\infty}(\mathcal{F})$. A class \mathcal{F} for which this is true is called a Donsker class, or *P*-Donsker class to be more complete (with (A.1) known as the uniform central limit theorem).

The nature of the limit process \mathbb{G} follows from consideration of its finite-dimensional distributions. The finite-dimensional distributions $\mathbb{G}_n f$ converge if and only if the functions f are square-integrable. In that case the multivariate central limit theorem yields that for any finite set f_1, \ldots, f_k of functions,

$$(\mathbb{G}_n f_1,\ldots,\mathbb{G}_n f_k) \xrightarrow{d} \mathcal{N}_k(\mathbf{0},\boldsymbol{\Sigma}),$$

where the $k \times k$ -matrix Σ has (i, j)-th element $P(f_i - Pf_i)(f_j - Pf_j)$. Since convergence in $\ell^{\infty}(\mathcal{F})$ implies finite-dimensional convergence, it follows that the limit process { $\mathbb{G}f : f \in \mathcal{F}$ } must be a zero-mean Gaussian process with covariance function

$$E\mathbb{G}f\mathbb{G}g = P(f - Pf)(g - Pg) = Pfg - PfPg,$$

for any $f, g \in \mathcal{F}$, and we call { $\mathbb{G}f : f \in \mathcal{F}$ } the *P*–Brownian bridge. Note that the sample paths of the Brownian bridge are zero at the endpoints $-\infty$ and ∞ . This is a consequence of the fact that the difference of two distributions is zero at these points. A class of measurable functions is called pre-Gaussian if the tight limit process \mathbb{G} in the uniform central limit theorem (A.1) exists. Now, assume that $f(x) = I(x \le t)$, then

$$\mathbb{G}_n f = \sqrt{n} \left(P_n f - P f \right) = \sqrt{n} \left(\mathbb{F}_n(t) - F(t) \right) \xrightarrow{d} \mathcal{N}(0, F(t)(1 - F(t))), \tag{A.2}$$

where $\mathbb{F}_n(t) = \frac{1}{n} \sum_{i=1}^n I(X_i \le t)$ and $F(t) = P(X_1 \le t)$. Note that the right side of (A.2) is because $\operatorname{Var}(f(X_1)) = E\left[(f(X_1) - F(t))^2\right] = E\left[f^2(X_1)\right] - F^2(t) = P(X_1 \le t) - F^2(t) = F(t) - F^2(t)$. Moreover, for any finite set of points t_1, \dots, t_k , let $f_i(x) = I(x \le t_i)$. Then

$$(\mathbb{G}_n f_1, \ldots, \mathbb{G}_n f_k) = \sqrt{n} \left(\mathbb{F}_n(t_1) - F(t_1), \ldots, \mathbb{F}_n(t_k) - F(t_k) \right) \xrightarrow{d} \mathcal{N}_k(\mathbf{0}, \mathbf{\Sigma}),$$

where $\Sigma = (F(t_i \wedge t_j) - F(t_i)F(t_j))_{1 \le i,j \le k}$ by noting that

$$\operatorname{Cov}(f_i(X_1), f_j(X_1)) = E\left[(f_i(X_1) - F(t_i))(f_j(X_1) - F(t_j))\right] = E\left[f_i(X_1)f_j(X_1)\right] - F(t_i)F(t_j)$$

and if $t_i < t_j$,

$$E \left[f_i(X_1) f_j(X_1) \right] = E \left[f_i(X_1) (f_i(X_1) + f_j(X_1) - f_i(X_1)) \right]$$

= $E \left[f_i^2(X_1) \right] + E \left[f_i(X_1) (f_j(X_1) - f_i(X_1)) \right]$
= $F(t_i) + E \left[I(X_1 \le t_i) (I(X_1 \le t_j) - I(X_1 \le t_i)) \right]$
= $F(t_i) + E \left[I(X_1 \le t_i) I(t_i < X_1 \le t_j) \right]$
= $F(t_i).$

Hence, in general, $E[f_i(X_1)f_j(X_1)] = F(t_i \wedge t_j)$, and therefore

$$\operatorname{Cov}\left(f_i(X_1), f_j(X_1)\right) = F(t_i \wedge t_j) - F(t_i)F(t_j).$$

Thus, \mathbb{G}_n converges in distribution to a Brownian bridge \mathbb{G}_F with mean zero and covariance function

$$F(s \wedge t) - F(s)F(t) \tag{A.3}$$

for any $s, t \in \mathbb{R}$.

Theorem A.7. (Donsker). If $X_1, X_2, ...$ are i.i.d. random variables with distribution function F, then the sequence of empirical processes $\sqrt{n}(\mathbb{F}_n - F)$ converges in distribution in the space $D[-\infty, \infty]$ to a tight random element \mathbb{G}_F , whose finite-dimensional distributions are zero-mean normal with covariance function (A.3).

Proof. See Theorem 19.3 and Theorem 19.5 in Van der Vaart [11].

Theorem A.8. If \mathbb{G}_G is a *G*-Brownian bridge process indexed by the half-lines $(-\infty, t]$, then

$$-\int \mathbb{G}_G dF\sim \mathbb{G}_G(F),$$

where the right side of both equality denotes a G-Brownian bridge process indexed by the single function F.

Proof. This theorem is stated as an exercise in Van der Vaart and Wellner [16]. See Section B.1 for a proof. □

A.3.2 Maximum Likelihood Estimators

If X_1, \ldots, X_n are a random sample from a density p_{θ} , then the maximum likelihood estimator $\hat{\theta}_n$ maximizes the function $\theta \mapsto \mathbb{P}_n \log p_{\theta}$. Hence, it is an *M*-estimator with the score function $\dot{\ell}_{\theta} = \frac{\partial}{\partial \theta} \log p_{\theta}$.

Theorem A.9. Suppose that there exists a measurable function $\dot{\ell}$ with $P_{\theta_0}\dot{\ell}^2 < \infty$ such that, for every θ_1 and θ_2 in a neighborhood of $\theta_0 \in \Theta \subset \mathbb{R}^k$,

$$\left|\log p_{\theta_1}(x) - \log p_{\theta_2}(x)\right| \le \dot{\ell}(x) \left\|\theta_1 - \theta_2\right\|.$$

Furthermore, suppose that there exists a measurable vector-valued function $\dot{\ell}_{\theta_0}$ such that, as $\theta \rightarrow \theta_0$,

$$\int \left[\sqrt{p_{\theta}} - \sqrt{p_{\theta_0}} - \frac{1}{2} \left(\theta - \theta_0\right)^{\top} \dot{\ell}_{\theta_0} \sqrt{p_{\theta_0}}\right]^2 d\mu = o\left(\|\theta - \theta_0\|^2\right).$$
(A.4)

If the Fisher information matrix $I_{\theta_0} = P_{\theta_0} \dot{\ell}_{\theta_0} \dot{\ell}_{\theta_0}^{\top}$ is nonsingular and $\hat{\theta}_n$ is consistent, then

$$\sqrt{n}\left(\widehat{\theta}_n - \theta_0\right) = \mathcal{I}_{\theta_0}^{-1} \mathbb{G}_n \dot{\ell}_{\theta_0}\left(X_i\right) + o_p(1).$$

In particular, the sequence $\sqrt{n} \left(\widehat{\theta}_n - \theta_0 \right)$ is asymptotically normal with mean zero and covariance matrix $I_{\theta_0}^{-1}$.

Proof. See Theorem 5.39 in Van der Vaart [11].

Note that $\frac{\partial}{\partial \theta} \sqrt{p_{\theta}} = \frac{1}{2\sqrt{p_{\theta}}} \frac{\partial}{\partial \theta} p_{\theta} = \frac{1}{2} \left(\frac{\partial}{\partial \theta} \log p_{\theta} \right) \sqrt{p_{\theta}}$, so if the map $\theta \mapsto P_{\theta_0} \log p_{\theta}$ is twice continuously differentiable at θ_0 , then the condition (A.4) holds from the Taylor expansion.

A.3.3 Functional Delta Method

A map $\phi : \mathbb{D}_{\phi} \to \mathbb{E}$, defined on a subset \mathbb{D}_{ϕ} of a normed space \mathbb{D} that contains θ , is called Hadamard differentiable at θ if there exists a continuous, linear map $\phi'_{\theta} : \mathbb{D} \to \mathbb{E}$ such that

$$\left\|\frac{\phi\left(\theta+th_{t}\right)-\phi(\theta)}{t}-\phi_{\theta}'(h)\right\|_{\mathbb{E}}\to 0, \quad \text{as } t\downarrow 0,$$

for every $h_t \to h$ such that $\theta + th_t$ is contained in \mathbb{D}_{ϕ} for all small t > 0. The definition of Hadamard differentiable as given requires that $\phi'_{\theta} : \mathbb{D} \to \mathbb{E}$ exists as a map on the whole space \mathbb{D} . If this is not the case, but ϕ'_{θ} exists on a subset \mathbb{D}_0 and the sequences $h_t \to h$ are restricted to converge to limits $h \in \mathbb{D}_0$, then ϕ is called Hadamard differentiable tangentially to this subset. **Theorem A.10.** (Delta method). Let \mathbb{D} and \mathbb{E} be normed linear spaces. Let $\phi : \mathbb{D}_{\phi} \subset \mathbb{D} \to \mathbb{E}$ be Hadamard differentiable at θ tangentially to \mathbb{D}_0 . Let $X_n : \Omega_n \to \mathbb{D}_{\phi}$ be maps with $r_n (X_n - \theta) \xrightarrow{d} X$ for some sequence of numbers $r_n \to \infty$, where X is separable and takes its values in \mathbb{D}_0 . Then $r_n (\phi(X_n) - \phi(\theta)) \xrightarrow{d} \phi'_{\theta}(X)$. If ϕ'_{θ} is defined and continuous on the whole space \mathbb{D} , then the sequence $r_n (\phi(X_n) - \phi(\theta)) = \phi'_{\theta} (r_n (X_n - \theta)) + o_p(1)$.

Proof. See Theorem 3.9.4 in Van der Vaart and Wellner [16].

Given a cadlag function A and a function of bounded variation B on an interval $[a, b] \subset \mathbb{R}$, define

$$\phi(A,B) = \int_{(a,b]} A dB$$

This map is Hadamard differentiable if the domain is restricted to pairs (A, B) such that B is of total variation bounded by some fixed constant.

Lemma A.11. For each fixed M, the map $\phi : D_{\phi} \subseteq D[a,b] \times BV_M[a,b] \to \mathbb{R}$ is Hadamard differentiable at each $(A,B) \in D_{\phi}$ such that $\int |dA| < \infty$. The derivative is given by

$$\phi_{A,B}'(\alpha,\beta) = \int Ad\beta + \int \alpha dB = A\beta \bigg|_a^b - \int \beta_- dA + \int \alpha dB,$$

where β_{-} is denoted as the left-continuous version of a cadlag function β .

Proof. See Lemma 3.9.17 in Van der Vaart and Wellner [16].

A.4 Trees

The following statements, lemmas, and definition are stated in Wager and Athey [31].

Lemma A.12. Let T be a regular, random-split tree and let L(x) denote its leaf containing x. Suppose that $X_1, \ldots, X_s \sim \text{Unif}([0,1]^d)$ independently. Then, for any $0 < \eta < 1$, and for large enough s,

$$P\left(\operatorname{diam}_{j}(L(x)) \ge \left(\frac{s}{2k-1}\right)^{-\frac{0.99(1-\eta)\log\left((1-\alpha)^{-1}\right)}{\log\left(\alpha^{-1}\right)}\frac{\pi}{d}}\right) \le \left(\frac{s}{2k-1}\right)^{-\frac{\eta^{2}}{2}\frac{1}{\log\left(\alpha^{-1}\right)}\frac{\pi}{d}},$$

where $\operatorname{diam}_{j}(L(x))$ denotes the length of the longest such segment that is parallel to the *j*-th axis.

Proof. See Lemma 1 in Wager and Athey [31].

To address the challenges associated with analyzing greedy tree models such as CART trees, Wager and Athey [31] followed the approach of Lin and Jeon [71] and investigated potential nearest neighbor predictors, which is a more general class of predictors performing a nearest-neighbor search over rectangles.

Definition A.13. Consider a set of points $X_1, \ldots, X_s \in \mathbb{R}^d$ and a fixed $x \in \mathbb{R}^d$. A point X_i is a potential nearest neighbor (PNN) of x if the smallest axis-aligned hyperrectangle with vertices x and X_i contains no other points X_j . Extending this notion, a PNN k-set of x is a set of points $\Lambda \subseteq \{X_1, \ldots, X_s\}$ of size $k \leq |L| < 2k - 1$ such that there exists an axis-aligned hyperrectangle L with vertex x containing Λ and no other training points. A training sample X_i is called a k-PNN of x if there exists a PNN k-set of x containing X_i . Finally, a predictor T is a k-PNN predictor over $\{Z\}$ if, given a training set

$$\{Z\} = \{(X_1, Y_1), \dots, (X_s, Y_s)\} \in \{\mathbb{R}^d \times \mathcal{Y}\}^s$$

and a test point $x \in \mathbb{R}^d$, T always outputs the average of the responses Y_i over a k-PNN set of x.

Definition A.13 allows us to describe a wide variety of tree predictors. For example, Lin and Jeon [71] showed that any decision tree that makes axis-aligned splits and has leaves containing between k and 2k - 1 points is a k-PNN predictor. In particular, the CART trees grown up to a leaf size k are k-PNN predictors. The predictions made by k-PNN predictors can always be expressed as

$$T(x;\xi,Z_1,\ldots,Z_s)=\sum_{i=1}^s W_iY_i,$$

where

$$W_i = \begin{cases} \frac{1}{|\{i:X_i \in L(x)\}|} & \text{if } i \in \{i:X_i \in L(x)\}, \\ 0 & \text{otherwise.} \end{cases}$$

If the tree is honest, then W_i is independent of Y_i given X_i for each *i*. A crucial characteristic of *k*-PNN predictors is their ability to indicate whether W_i is likely to be non-zero, even if only Z_i is observed. This feature is essential in showing the incrementality of *k*-PNNs.

Lemma A.14. Suppose that the observations $X_1, X_2, ...$ are i.i.d. on $[0,1]^d$ with a density f that is bounded away from infinity, and let T be any symmetric k-PNN predictor. Then, there is a constant $C_{f,d}$ depending only on f and d such that, as s gets large,

$$s \operatorname{Var}\left(E\left[W_1 \mid Z_1\right]\right) \gtrsim \frac{C_{f,d}}{k \log(s)^d},\tag{A.5}$$

where W_i is the indicator for whether the observation is selected in the subsample. When f is uniform over $[0,1]^d$, the bound holds with $C_{f,d} = 2^{-(d+1)}(d-1)!$.

Proof. See Lemma 4 in Wager and Athey [31].

Note that (A.5) can be interpreted as a lower bound on how much information Z_1 contains about the selection event W_1 . This result allows for the proof of the incrementality of all honest and regular random-split trees. Notice that any symmetric α -regular tree following Definition 2.4 is also a symmetric *k*-PNN predictor.

Appendix B Proofs of the Theorems

In this appendix, we provide proofs for the main theorems stated in the previous chapters.

B.1 Theorem on Empirical AUROC Asymptotics

Theorem A.8 is stated as an exercise in Van der Vaart and Wellner [16] without a proof. We provide a proof for it below.

Proof of Theorem A.8. Note that \mathbb{G}_G has mean zero and covariance function $G(s \wedge t) - G(s)G(t)$ for any $s, t \in \mathbb{R}$. Hence, $-\int \mathbb{G}_G dF$ has mean zero and covariance function

$$\int \int \left(G(s \wedge t) - G(s)G(t) \right) dF(s) dF(t).$$
(B.1)

It can be shown that the covariance function (B.1) has an alternative expression as:

$$\begin{split} &\int_{t\in\mathbb{R}} \left(\int_{s\in(-\infty,t]} \left(G(s) - G(s)G(t) \right) dF(s) + \int_{s\in[t,\infty)} \left(G(t) - G(s)G(t) \right) dF(s) \right) dF(t) \\ &= \int_{t\in\mathbb{R}} \left[\int_{s\in(-\infty,t]} G(s)dF(s) \right] \left(1 - G(t) \right) dF(t) + \int_{t\in\mathbb{R}} \left[\int_{s\in[t,\infty)} \left(1 - G(s) \right) dF(s) \right] G(t)dF(t) \\ &= \int_{t\in\mathbb{R}} \left[\int_{s\in(-\infty,t]} G(s)dF(s) \right] \left(1 - G(t) \right) dF(t) + \int_{t\in\mathbb{R}} G(s)dF(s) - \int_{s\in(-\infty,t]} \left(1 - G(s) \right) dF(s) \right] G(t)dF(t) \\ &= \int_{s\in\mathbb{R}} \left[\int_{t\in[s,\infty)} dF(t) \right] G(s)dF(s) + \int GdF - \left(\int GdF \right)^2 - \int_{t\in\mathbb{R}} \left[\int_{s\in(-\infty,t]} dF(s) \right] G(t)dF(t) \\ &= \int \left(1 - F \right) GdF + \int GdF - \left(\int GdF \right)^2 - \int FGdF \\ &= 2 \int GdF - 2 \int GFdF - \left(\int GdF \right)^2 \\ &= 2 \left(GF \Big|_{-\infty}^{\infty} - \int FdG \right) - \left(GF^2 \Big|_{-\infty}^{\infty} - \int F^2 dG \right) - \left(GF \Big|_{-\infty}^{\infty} - \int FdG \right)^2 \end{split}$$

The integral $-\int \mathbb{G}_G dF$ can be understood as the limit of finite linear combinations of Gaussian processes in the sense of the Riemann–Stieltjes integral, so it is also a Gaussian process. It can be demonstrated that both $\mathbb{G}_G(F)$ and $-\mathbb{G}_G(F)$ satisfy all the conditions mentioned above. To determine the correct process, observe that

$$\int G dF = 1 - \int F dG$$

so the resulting process should have the opposite sign of $-\int \mathbb{G}_G dF$. Thus, $-\int \mathbb{G}_G dF \sim \mathbb{G}_G(F)$.
B.2 Outlier-Robust Random Forest Regression for Between-Subject Attributes

B.2.1 Asymptotic Unbiasedness

Proof of Theorem 2.6. Since a random forest averages independently generated trees, its bias is of the same order as that of a single tree. Thus, it suffices to prove the bound for the tree estimator. Assume that $L(x^{(1)}) \neq L(x^{(2)})$. By the honesty assumption (Definition 2.2), the estimation of θ is independent of the tree construction process. Consequently, $\hat{\theta}_{\text{tree}}$ is an unbiased estimator, and we have:

$$E[T(x^{(1)}, x^{(2)}; Z)] - \theta(x^{(1)}, x^{(2)})$$

= $E[\hat{\theta}_{tree}(x^{(1)}, x^{(2)})] - \theta(x^{(1)}, x^{(2)})$
= $E[E[h(Y_i, Y_j) | (X_i, X_j) \in L(x^{(1)}) \times L(x^{(2)})] - E[h(Y_i, Y_j) | (X_i, X_j) = (x^{(1)}, x^{(2)})]].$ (B.2)

Since $\theta(x^{(1)}, x^{(2)})$ is Lipschitz continuous, it follows that:

$$\begin{aligned} \left| E[h(Y_i, Y_j) \mid (X_i, X_j) \in L(x^{(1)}) \times L(x^{(2)})] - E[h(Y_i, Y_j) \mid (X_i, X_j) = (x^{(1)}, x^{(2)})] \right| \\ \leq KD(L(x^{(1)}), L(x^{(2)})), \end{aligned}$$

where *K* is the Lipschitz constant and $D(L(x^{(1)}), L(x^{(2)})) = \sqrt{\operatorname{diam}(L(x^{(1)}))^2 + \operatorname{diam}(L(x^{(2)}))^2}$. Thus, to bound (B.2), it is necessary to bound the average diameters of $L(x^{(1)})$ and $L(x^{(2)})$ simultaneously.

To achieve this, let $\eta = \sqrt{\log((1-\alpha)^{-1})}$. Since $\alpha \le 0.2$, it follows that $\eta \le 0.48$, ensuring $0.99 \cdot (1-\eta) \ge 0.51$. By applying the Pythagorean theorem, we obtain the following:

$$\left\{ D(L(x^{(1)}), L(x^{(2)})) \ge \sqrt{2d}A(s, 0.51) \right\} \subseteq \bigcup_{j=1}^{d} \left(\left\{ \operatorname{diam}_{j}(L(x^{(1)})) \ge A(s, 0.51) \right\} \cup \left\{ \operatorname{diam}_{j}(L(x^{(2)})) \ge A(s, 0.51) \right\} \right),$$

where $A(s,\zeta) = \left(\frac{s}{2k-1}\right)^{-\zeta \frac{\log((1-\alpha)^{-1})}{\log(\alpha^{-1})} \frac{\pi}{d}}$.

Using Lemma A.12 and Boole's inequality [72], we find that for sufficiently large s,

$$P\left(D(L(x^{(1)}), L(x^{(2)})) \ge \sqrt{2d}A(s, 0.51)\right) \le 2dA(s, 0.5).$$

Thus, for sufficiently large *s*, the bias can be bounded as follows:

$$\begin{split} & \left| E[T(x^{(1)}, x^{(2)}; Z)] - \theta(x^{(1)}, x^{(2)}) \right| \\ \leq & E\left[\left| E[h(Y_i, Y_j) \mid (X_i, X_j) \in L(x^{(1)}) \times L(x^{(2)})] - E[h(Y_i, Y_j) \mid (X_i, X_j) = (x^{(1)}, x^{(2)})] \right| \right] \\ \leq & KE[D(L(x^{(1)}), L(x^{(2)}))] \\ < & K\left(\sqrt{2d}A(s, 0.51) + \sqrt{2d} \cdot 2dA(s, 0.5) \right) \\ \leq & K\sqrt{2d} (1+2d) A(s, 0.5). \end{split}$$

B.2.2 Asymptotic Normality

Given a regression tree *T* and independent training samples Z_1, \ldots, Z_n , the Hájek projection [11] of *T* is defined as

$$\widehat{T}_{\text{Hájek}} = E[T] + \sum_{i=1}^{n} (E[T \mid Z_i] - E[T]).$$
(B.3)

According to classical *U*-statistics theory, the variance of Hájek projection satisfies $Var(\widehat{T}_{Hájek}) \leq Var(T)$. However, for *T*, the following asymptotic condition does not hold:

$$\lim_{n \to \infty} \frac{\operatorname{Var}(\widehat{T}_{\operatorname{Hájek}})}{\operatorname{Var}(T)} = 1 \quad \text{implies that} \quad \lim_{n \to \infty} \frac{E\left[\|\widehat{T}_{\operatorname{Hájek}} - T\|_2^2\right]}{\operatorname{Var}(T)} = 0.$$

This indicates that the classical Hájek projection framework cannot be directly employed to establish the asymptotic normality of T. To address this limitation, we introduce a weaker criterion termed *v*-incrementality for between-subject attributes. This criterion is adapted from the work of Wager and Athey [31].

Definition B.1. The predictor T is said to be v(s)-incremental at $(x^{(1)}, x^{(2)})$ if it satisfies

$$\frac{\operatorname{Var}(\widehat{T}_{H\acute{a}jek}(x^{(1)},x^{(2)};Z_1,\ldots,Z_s))}{\operatorname{Var}(T(x^{(1)},x^{(2)};Z_1,\ldots,Z_s))} \gtrsim \nu(s),$$

where $\widehat{T}_{Hájek}$ represents the Hájek projection of T as defined in (B.3), and the notation $f(s) \ge g(s)$ indicates that

$$\liminf_{s\to\infty}\frac{f(s)}{g(s)}\geq 1.$$

To achieve the asymptotic normality of T, our method involves two main steps. First, we derive lower bounds for the incrementality of regression trees (Theorem B.2). Following this, we demonstrate that by applying subsampling, weakly incremental predictors T can be transformed into 1-incremental ensembles (Lemma B.3), thereby aligning the problem with

classical theoretical frameworks. This approach is based on the framework originally developed by Wager and Athey [31].

Theorem B.2. Assume that the conditions of Lemma A.14 are satisfied and that T is a doublesample tree as described in Procedure 2.1 that meets the criteria of honesty (Definition 2.2), α -regularity (Definition 2.4) with $\alpha \leq 0.2$, and symmetry (Definition 2.5). Further, assume that the conditional moment $\theta(x^{(1)}, x^{(2)})$ is Lipschitz continuous at the test point $(x^{(1)}, x^{(2)})$. Additionally, suppose that the variance $\operatorname{Var}(I(Y_1 \leq Y_2) | X_1 = x^{(1)})$ is positive. Then T is $\nu(s)$ incremental at $(x^{(1)}, x^{(2)})$, where

$$\nu(s) = \frac{C_{f,d}}{16\log(s)^d},$$

and $C_{f,d}$ is the constant from Lemma A.14.

Proof. We first prove the case where the tree is grown using Z_1, \ldots, Z_s without using the responses Y_1, \ldots, Y_s to determine where to place its splits, in contrast to the double-sample trees considered in the theorem. Note that the tree grown in this manner also satisfies honesty.

Assume that $L(x^{(1)}) \neq L(x^{(2)})$. According to Lemma A.14, we have

$$\operatorname{Var}(E[W_1 \mid Z_1]) = \Omega\left(\frac{1}{s\log(s)^d}\right). \tag{B.4}$$

Consider the truncated tree predictor defined by

$$T'(x^{(1)}, x^{(2)}; Z) = T(x^{(1)}, x^{(2)}; Z) \cdot I\left(D(L(x^{(1)}), L(x^{(2)})) \le s^{-\kappa(0.51)}\right),$$

where

$$D(L(x^{(1)}), L(x^{(2)})) = \sqrt{\operatorname{diam}(L(x^{(1)}))^2 + \operatorname{diam}(L(x^{(2)}))^2} \quad \text{and} \quad \kappa(\zeta) = \zeta \frac{\pi}{d} \frac{\log\left((1-\alpha)^{-1}\right)}{\log\left(\alpha^{-1}\right)}.$$

Similarly, define the truncated selection variables as

$$W'_{i} = W_{i} \cdot I\left(D(L(x^{(1)}), L(x^{(2)})) \le s^{-\kappa(0.51)}\right).$$

Applying the ANOVA decomposition (as described by Efron and Stein [73]), we obtain

$$\begin{aligned} \operatorname{Var}\left(E\left[T(x^{(1)}, x^{(2)}; Z) \mid Z_{1}\right] - E\left[T'(x^{(1)}, x^{(2)}; Z) \mid Z_{1}\right]\right) \\ &= \operatorname{Var}\left(E\left[T(x^{(1)}, x^{(2)}; Z)I\left(D(L(x^{(1)}), L(x^{(2)})) > s^{-\kappa(0.51)}\right)\right| Z_{1}\right]\right) \\ &\leq \frac{1}{s}\operatorname{Var}\left(T(x^{(1)}, x^{(2)}; Z)I\left(D(L(x^{(1)}), L(x^{(2)})) > s^{-\kappa(0.51)}\right)\right) \\ &\leq \frac{1}{s|L^{(1)}||L^{(2)}|} \sum_{\substack{i \in L^{(1)} \\ j \in L^{(2)}}} E\left[I(Y_{i} \leq Y_{j})^{2} \mid (X_{i}, X_{j}) \in L(x^{(1)}) \times L(x^{(2)})\right] \times \\ & P\left(D(L(x^{(1)}), L(x^{(2)})) > s^{-\kappa(0.51)}\right) \end{aligned}$$

$$\leq \frac{\sup_{(x^{(1)}, x^{(2)}) \in [0,1]^d \times [0,1]^d} \left\{ P(Y_i \leq Y_j \mid (X_i, X_j) = (x^{(1)}, x^{(2)})) \right\}}{s} \times P\left(D(L(x^{(1)}), L(x^{(2)})) > s^{-\kappa(0.51)} \right)$$

$$\leq \frac{1}{s} P\left(D(L(x^{(1)}), L(x^{(2)})) > s^{-\kappa(0.51)}\right)$$

$$\leq \frac{1}{s} P\left(\bigcup_{j=1}^{d} \left(\left\{\operatorname{diam}_{j}(L(x^{(1)})) > \frac{1}{\sqrt{2d}}s^{-\kappa(0.51)}\right\} \cup \left\{\operatorname{diam}_{j}(L(x^{(2)})) > \frac{1}{\sqrt{2d}}s^{-\kappa(0.51)}\right\}\right)\right)$$

 $\leq \sqrt{2d}s^{-(\kappa(0.5)+1)}$ by Lemma A.12 and Boole's inequality.

Note that since logarithmic growth is slower than polynomial growth, the variance of the difference between $E[T | Z_1]$ and $E[T' | Z_1]$ decays faster than the rate specified in (B.4). Additionally, for random variables *A* and *B*, we have:

$$Var(A - B) = Var(A) + Var(B) - 2 Cov(A, B)$$

$$\geq Var(A) + Var(B) - 2\sqrt{Var(A) Var(B)} \quad \text{by Cauchy-Schwarz inequality [74]}$$

$$= \left(\sqrt{Var(A)} - \sqrt{Var(B)}\right)^{2}.$$

Therefore,

$$\left(\sqrt{\operatorname{Var}(A)} - \sqrt{\operatorname{Var}(B)}\right)^2 \le \operatorname{Var}(A - B) \to 0$$

implies that

$$\operatorname{Var}(A) - \operatorname{Var}(B) \to 0.$$

Thus, as $s \to \infty$,

$$\operatorname{Var}\left(E\left[T(x^{(1)}, x^{(2)}; Z) \mid Z_{1}\right]\right) - \operatorname{Var}\left(E\left[T'(x^{(1)}, x^{(2)}; Z) \mid Z_{1}\right]\right) \to 0.$$
(B.5)

Similarly, by a similar argument, we also have:

$$Var(E[W_1 | Z_1]) - Var(E[W'_1 | Z_1]) \to 0.$$
(B.6)

Since (B.5) holds, we shift our focus from analyzing *T* to analyzing *T'*. Our objective is to establish a lower bound for the variance of the conditional expectation of $T'(x^{(1)}, x^{(2)}; Z)$ given Z_1 . We begin by noting that, for random variables *A*, *B*, and *C*, and applying the law of iterated expectations [61], we obtain:

$$Var(E[A | B, C] - E[A | B]) = E[(E[A | B, C] - E[A | B])^{2}]$$

= $E[E[A | B, C]^{2}] + E[E[A | B]^{2}] - 2E[E[A | B, C]E[A | B]]$
= $E[E[A | B, C]^{2}] - E[E[A | B]^{2}]$
= $Var(E[A | B, C]) - Var(E[A | B]).$

Thus, we have:

$$\operatorname{Var}\left(E\left[T'(x^{(1)}, x^{(2)}; Z) \mid Z_{1}\right]\right) = \operatorname{Var}\left(E\left[T'(x^{(1)}, x^{(2)}; Z) \mid X_{1}\right]\right) + \operatorname{Var}\left(E\left[T'(x^{(1)}, x^{(2)}; Z) \mid Z_{1}\right) - E\left[T'(x^{(1)}, x^{(2)}; Z) \mid X_{1}\right]\right)$$
$$\geq \operatorname{Var}\left(E\left[T'(x^{(1)}, x^{(2)}; Z) \mid Z_{1}\right] - E\left[T'(x^{(1)}, x^{(2)}; Z) \mid X_{1}\right]\right).$$

Therefore, it is sufficient to find a lower bound for the latter term. Next, we observe that, given the honesty condition and the i.i.d. sampling assumption,

$$E[T'(x^{(1)}, x^{(2)}; Z) | Z_1] - E[T'(x^{(1)}, x^{(2)}; Z) | X_1]$$

= $\sum_{j=2}^n E[W'_1(x^{(1)})W'_j(x^{(2)})I(Y_1 \le Y_j) | X_1, Y_1] - \sum_{j=2}^n E[W'_1(x^{(1)})W'_j(x^{(2)})I(Y_1 \le Y_j) | X_1]$
= $E[W'_1(x^{(1)}) | X_1] \left(E\left[\sum_{j=2}^n W'_j(x^{(2)})I(Y_1 \le Y_j) \middle| Y_1 \right] - E\left[\sum_{j=2}^n W'_j(x^{(2)})I(Y_1 \le Y_j) \middle| X_1 \right] \right)$
= $E[W'_1(x^{(1)}) | X_1] (E[I(Y_1 \le Y_2) | Y_1] - E[I(Y_1 \le Y_2) | X_1]).$

Given the honesty condition and Lipschitz continuity assumption, it follows that

$$\operatorname{Var}\left(E\left[T'(x^{(1)}, x^{(2)}; Z) \mid Z_{1}\right] - E\left[T'(x^{(1)}, x^{(2)}; Z) \mid X_{1}\right]\right)$$

=
$$\operatorname{Var}\left(E\left[W'_{1}(x^{(1)}) \mid X_{1}\right]\left(E\left[I(Y_{1} \leq Y_{2}) \mid Y_{1}\right] - \left(E\left[I(Y_{1} \leq Y_{2}) \mid X_{1} = x^{(1)}\right] + O\left(s^{-\kappa(0.5)}\right)\right)\right)\right)$$

=
$$\operatorname{Var}\left(E\left[W'_{1}(x^{(1)}) \mid X_{1}\right]\left(E\left[I(Y_{1} \leq Y_{2}) \mid Y_{1}\right] - E\left[I(Y_{1} \leq Y_{2}) \mid X_{1} = x^{(1)}\right]\right)\right) + O\left(E\left[E\left[W'_{1}(x^{(1)}) \mid X_{1}\right]^{2}\right]s^{-2\kappa(0.5)}\right), \quad (B.7)$$

where it is noteworthy that the error term in Equation (B.7) decays at the order of $s^{-(1+2\kappa(0.5))}$ since $E\left[E[W'_1(x^{(1)}) | X_1]^2\right] = O(1/s)$ by the proof in Lemma A.14. Finally, by utilizing the law of iterated expectations, the principal term in Equation (B.7) can be expressed as

$$\operatorname{Var}\left(E\left[W_{1}'(x^{(1)}) \mid X_{1}\right]\left(E\left[I(Y_{1} \leq Y_{2}) \mid Y_{1}\right] - E\left[I(Y_{1} \leq Y_{2}) \mid X_{1} = x^{(1)}\right]\right)\right)$$

= $E\left[E\left[W_{1}'(x^{(1)}) \mid X_{1}\right]^{2}E\left[\left(E\left[I(Y_{1} \leq Y_{2}) \mid Y_{1}\right] - E\left[I(Y_{1} \leq Y_{2}) \mid X_{1} = x^{(1)}\right]\right)^{2} \mid X_{1}\right]\right] - (B.8)$
 $E\left[E\left[W_{1}'(x^{(1)}) \mid X_{1}\right]E\left[\left(E\left[I(Y_{1} \leq Y_{2}) \mid Y_{1}\right] - E\left[I(Y_{1} \leq Y_{2}) \mid X_{1} = x^{(1)}\right]\right) \mid X_{1}\right]\right]^{2}.$

Given that the first conditional moment of $I(Y_1 \le Y_2)$ given (X_1, X_2) satisfies a Lipschitz condition, we can deduce:

$$E\left[E[W_{1}'(x^{(1)}) \mid X_{1}]^{2}E\left[\left(E\left[I(Y_{1} \leq Y_{2}) \mid Y_{1}\right] - E\left[I(Y_{1} \leq Y_{2}) \mid X_{1} = x^{(1)}\right]\right)^{2} \mid X_{1}\right]\right] - E\left[E[W_{1}'(x^{(1)}) \mid X_{1}]^{2}\right]\operatorname{Var}(I(Y_{1} \leq Y_{2}) \mid X_{1} = x^{(1)}) \to 0.$$

The second term in the expansion (B.8) can be shown to be of order $1/s^2$ since the expected value $E\left[E[W'_1(x^{(1)}) | X_1]\right]$ is O(1/s), as established in the proof in Lemma A.14. Consequently, both this term and the error term in Equation (B.7) are negligible compared to the first term in the expansion (B.8), leading to the conclusion that

$$\operatorname{Var}\left(E\left[T'(x^{(1)}, x^{(2)}; Z) \mid Z_{1}\right]\right) \gtrsim \operatorname{Var}\left(E\left[W'_{1}(x^{(1)}) \mid X_{1}\right]\right) \operatorname{Var}(I(Y_{1} \leq Y_{2}) \mid X_{1} = x^{(1)}).$$

By applying the results from (B.5) and (B.6), we can establish that

$$\operatorname{Var}\left(E\left[T(x^{(1)}, x^{(2)}; Z) \mid Z_1\right]\right) \gtrsim \operatorname{Var}\left(E\left[W_1(x^{(1)}) \mid X_1\right]\right) \operatorname{Var}(I(Y_1 \le Y_2) \mid X_1 = x^{(1)}). \quad (B.9)$$

Building on (B.9) and Lemma A.14, it follows that

$$\operatorname{Var}\left(E\left[T(x^{(1)}, x^{(2)}; Z) \mid Z_{1}\right]\right) \gtrsim \frac{C_{f,d}}{ks \log(s)^{d}} \operatorname{Var}(I(Y_{1} \leq Y_{2}) \mid X_{1} = x^{(1)}).$$

Furthermore, according to Theorem 2.6 and the continuous mapping theorem, we have

$$E[I(Y_1 \le Y_2) \mid X_1 \in L(x^{(1)}; Z)] \xrightarrow{p} E[I(Y_1 \le Y_2) \mid X_1 = x^{(1)}],$$

which, in conjunction with U-statistics theory [11] implies that

$$k \operatorname{Var}(T(x^{(1)}, x^{(2)}); Z)) \leq |\{i : X_i \in L(x^{(1)}; Z)\}| \cdot \operatorname{Var}(T(x^{(1)}, x^{(2)}); Z))$$

= $4 \operatorname{Var}(T(x^{(1)}, x^{(2)}); Z) | X_1)$
 $\xrightarrow{p} 4 \operatorname{Var}(I(Y_1 \leq Y_2) | X_1 = x^{(1)})$

as k remains constant and the range of the feature values within the leaves continues to decrease

in each dimensions of the feature space. Consequently, we can conclude that

$$\frac{\operatorname{Var}(\widehat{T}_{\operatorname{Hájek}}(x^{(1)}, x^{(2)}; Z))}{\operatorname{Var}(T(x^{(1)}, x^{(2)}; Z))} \gtrsim k \frac{s \operatorname{Var}\left(E[T(x^{(1)}, x^{(2)}; Z) \mid Z_1]\right)}{4 \operatorname{Var}(I(Y_1 \le Y_2) \mid X_1 = x^{(1)})} \gtrsim \frac{C_{f,d}}{4 \log(s)^d}.$$
 (B.10)

We now turn to the case of the double-sample tree. To begin, note that for random variables *A* and *B* with Cov(A, B) < 0, we observe that:

 $\operatorname{Var}(A+B) = \operatorname{Var}(A) + \operatorname{Var}(B) + 2\operatorname{Cov}(A,B)$

$$\geq \operatorname{Var}(A) + \operatorname{Var}(B) - 2\sqrt{\operatorname{Var}(A)\operatorname{Var}(B)} \quad \text{by Cauchy-Schwarz inequality}$$
$$\geq \operatorname{Var}(A) + \operatorname{Var}(B) - 2\left(\frac{1}{4}\operatorname{Var}(A) + \operatorname{Var}(B)\right)$$

by the inequality of arithmetic and geometric means

$$=\frac{1}{2}\operatorname{Var}(A)-\operatorname{Var}(B).$$

Then, we have:

$$\operatorname{Var}(\widehat{T}_{\operatorname{Hájek}}) = s \operatorname{Var}(E[T \mid Z_{1}]) = s \operatorname{Var}(E[I(1 \in I)T \mid Z_{1}) + E[I(1 \notin I)T \mid Z_{1}])$$

$$\geq \frac{s}{2} \operatorname{Var}(E[I(1 \in I)T \mid Z_{1}]) - s \operatorname{Var}(E[I(1 \notin I)T \mid Z_{1}])$$

$$\approx \frac{s}{8} \operatorname{Var}(E[T \mid Z_{1}] \mid 1 \in I) - \frac{s}{4} \operatorname{Var}(E[T \mid Z_{1}] \mid 1 \notin I),$$

where the approximation in the final line stems from the fact that the probability $P(1 \in \mathcal{I} | Z_1)$ is approximately 1/2 regardless of the value of Z_1 . Applying the result from (B.10), we obtain

$$\lfloor s/2 \rfloor \operatorname{Var} \left(E\left[T \mid Z_1\right] \mid 1 \in I \right) \gtrsim \frac{C_{f,d}}{4 \log(s)^d} \operatorname{Var}(T).$$

Moreover, by leveraging classical results such as those found in Hoeffding [75], we can assert that

$$\lceil s/2 \rceil \operatorname{Var} \left(E\left[T \mid Z_1\right] \mid 1 \notin I \right) \le \operatorname{Var} \left(E\left[T \mid \{Z_j : j \notin I\}\right] \mid I \right);$$

then, invoking Lemma A.12 and the reasoning used in proving Theorem 2.6, we deduce that

$$\operatorname{Var}(E[T \mid \{Z_j : j \notin I\}] \mid I) = O\left(s^{-\frac{\log\left((1-\alpha)^{-1}\right)}{\log\left(\alpha^{-1}\right)}\frac{\pi}{d}}\right),$$

As a result, the contribution from the condition where $1 \notin I$ can be considered negligibly small compared to the contribution from the condition where $1 \in I$ for sufficiently large *s*. Thus, we have:

$$\operatorname{Var}(\widehat{T}_{\mathrm{Hájek}}) \gtrsim \frac{s}{8} \operatorname{Var}(E[T \mid Z_1] \mid 1 \in \mathcal{I})$$
$$\gtrsim \frac{C_{f,d}}{16 \log(s)^d} \operatorname{Var}(T).$$

г	-		
L			
L		_	

A double-sample tree serves as an honest, symmetric *k*-PNN predictor for the \mathcal{I} -sample, while the data from the \mathcal{J} -sample can be effectively integrated into the auxiliary noise term ξ .

Lemma B.3. Consider $\hat{\theta}_{forest}(x^{(1)}, x^{(2)})$ as the estimate of $\theta(x^{(1)}, x^{(2)})$ provided by a random forest with base learner T as defined in Equation (2.3). Let $\hat{\theta}_{forest, Hájek}$ denote the Hájek projection of $\hat{\theta}_{forest}$ as specified in Equation (B.3). Then, we have

$$E\left[\left(\widehat{\theta}_{forest}(x^{(1)}, x^{(2)}) - \widehat{\theta}_{forest, Hájek}(x^{(1)}, x^{(2)})\right)^2\right] \le \left(\frac{s}{n}\right)^2 \operatorname{Var}\left(T(x^{(1)}, x^{(2)}; \xi, Z_1, \dots, Z_s)\right)$$

for all $2 \le s \le n$, assuming that the variance Var(T) of the base learner is finite.

Proof. Using the ANOVA decomposition approach introduced by Efron and Stein [73] for individual trees *T*, a random forest estimator $\hat{\theta}_{\text{forest}}(x^{(1)}, x^{(2)})$, as specified in Equation (2.3), can be expressed as

$$\begin{aligned} \widehat{\theta}_{\text{forest}}(x^{(1)}, x^{(2)}; Z_1, \dots, Z_n) &= E[T] + \frac{s}{n} \sum_{i=1}^n T_1(Z_i) + \frac{s(s-1)}{n(n-1)} \sum_{(i,j) \in \binom{S_n}{2}} T_2(Z_i, Z_j) + \dots \\ &+ \frac{\prod_{i=0}^{s-1} (s-i)}{\prod_{i=0}^{s-1} (n-i)} \sum_{(i_1, \dots, i_s) \in \binom{S_n}{s}} T_s(Z_{i_1}, \dots, Z_{i_s}). \end{aligned}$$

Based on Equation (B.3), the Hájek projection of $\widehat{\theta}_{\text{forest, Hájek}}$ can be expressed as

$$\widehat{\theta}_{\text{forest, Hájek}}(x^{(1)}, x^{(2)}; Z_1, \dots, Z_n) = E[T] + \frac{s}{n} \sum_{i=1}^n T_1(Z_i).$$
(B.11)

In the case of projections [11], we have

$$E\left[\left(\widehat{\theta}_{\text{forest}}(x^{(1)}, x^{(2)}) - \widehat{\theta}_{\text{forest}, \text{Hájek}}(x^{(1)}, x^{(2)})\right)^2\right]$$

= $\operatorname{Var}\left(\widehat{\theta}_{\text{forest}}(x^{(1)}, x^{(2)}) - \widehat{\theta}_{\text{forest}, \text{Hájek}}(x^{(1)}, x^{(2)})\right).$

Given that the functions $T_k(\cdot)$ are pairwise uncorrelated, it follows that:

$$\begin{split} E\left[\left(\widehat{\theta}_{\text{forest}}(x^{(1)}, x^{(2)}) - \widehat{\theta}_{\text{forest, Hájek}}(x^{(1)}, x^{(2)})\right)^2\right] &= \sum_{k=2}^s \left(\frac{\prod_{i=0}^{k-1}(s-i)}{\prod_{i=0}^{k-1}(n-i)}\right)^2 \binom{n}{k} \operatorname{Var}(T_k(Z_1, \dots, Z_k)) \\ &= \sum_{k=2}^s \frac{\prod_{i=0}^{k-1}(s-i)}{\prod_{i=0}^{k-1}(n-i)} \binom{s}{k} \operatorname{Var}(T_k(Z_1, \dots, Z_k)) \\ &\leq \frac{s(s-1)}{n(n-1)} \sum_{k=2}^s \binom{s}{k} \operatorname{Var}(T_k(Z_1, \dots, Z_k)) \\ &\leq \frac{s(s-1)}{n(n-1)} \sum_{k=1}^n \binom{n}{k} \operatorname{Var}(T_k(Z_1, \dots, Z_k)) \\ &= \frac{s(s-1)}{n(n-1)} \operatorname{Var}(T) \\ &\leq \frac{s^2}{n^2} \operatorname{Var}(T), \end{split}$$

for all $2 \le s \le n$.

The asymptotic normality result in Theorem B.4 follows from combining Theorem B.2 with Lemma B.3.

Theorem B.4. Consider the random forest estimator $\hat{\theta}_{forest,n}(x^{(1)}, x^{(2)})$ trained under the conditions specified in Theorem B.2. Additionally, assume that the subsample size s_n adheres to the following conditions:

 $\lim_{n\to\infty} s_n = \infty \quad and \quad \lim_{n\to\infty} \frac{s_n \log(n)^d}{n} = 0,$

and that for pairs $(i, j) \in \binom{S_n}{2}$,

$$E\left[\left|E[h(Y_i, Y_j) \mid Y_i] - E[h(Y_i, Y_j) \mid X_i = x^{(1)}]\right|^{2+\delta} \middle| X_i = x^{(1)} \right] \le M$$

for some positive constants δ and M, uniformly across all $x^{(1)} \in [0, 1]^d$. Under these assumptions,

there exists a sequence $\sigma_n(x^{(1)}, x^{(2)}) \rightarrow 0$ such that

$$\frac{\widehat{\theta}_{forest,n}(x^{(1)}, x^{(2)}) - E[\widehat{\theta}_{forest,n}(x^{(1)}, x^{(2)})]}{\sigma_n(x^{(1)}, x^{(2)})} \xrightarrow{d} \mathcal{N}(0, 1), \tag{B.12}$$

where $\mathcal{N}(0,1)$ denotes the standard normal distribution.

Proof. By Equation (B.11), the variance of $\hat{\theta}_{\text{forest, Hájek}}$ can be expressed as

$$\sigma_n^2 = \frac{s^2}{n} \operatorname{Var}(T_1(Z_1)) = \frac{s}{n} \operatorname{Var}(\widehat{T}_{\text{Hájek}}) \le \frac{s}{n} \operatorname{Var}(T) \to 0.$$

Furthermore, using Theorem B.2 in conjunction with Lemma B.3, we derive the following:

$$\begin{aligned} \frac{1}{\sigma_n^2} E\left[\left(\widehat{\theta}_{\text{forest}}(x^{(1)}, x^{(2)}) - \widehat{\theta}_{\text{forest, Hájek}}(x^{(1)}, x^{(2)})\right)^2\right] &\leq \left(\frac{s}{n}\right)^2 \frac{\text{Var}(T)}{\sigma_n^2} \\ &= \frac{s}{n} \frac{\text{Var}(T)}{\text{Var}(\widehat{T}_{\text{Hájek}})} \\ &\leq \frac{16s \log(s)^d}{nC_{f,d}} \\ &\to 0. \end{aligned}$$

As a result, by applying Slutsky's theorem, it is sufficient to verify that result (B.12) is satisfied for the Hájek projection of the random forest, $\hat{\theta}_{\text{forest, Hájek}}(x^{(1)}, x^{(2)})$.

Given the definition of σ_n , it reduces to establishing the asymptotic normality of $\hat{\theta}_{\text{forest, Hájek}}(x^{(1)}, x^{(2)})$. This can be accomplished by employing the Lyapunov central limit theorem (CLT) [76]. Specifically, by writing the Hájek projection of the random forest as

$$\widehat{\theta}_{\text{forest, Hájek}}(x^{(1)}, x^{(2)}) - E[\widehat{\theta}_{\text{forest}}(x^{(1)}, x^{(2)})] = \frac{s}{n} \sum_{i=1}^{n} (E[T \mid Z_i] - E[T]),$$

the Lyapunov CLT states that if Lyapunov's condition

$$\lim_{n \to \infty} \sum_{i=1}^{n} \frac{E\left[|E[T \mid Z_i] - E[T]|^{2+\delta}\right]}{\left(\sum_{j=1}^{n} \operatorname{Var}\left(E[T \mid Z_j]\right)\right)^{1+\delta/2}} = 0$$
(B.13)

holds for some $\delta > 0$, then

$$\sum_{i=1}^{n} \frac{E[T \mid Z_i] - E[T]}{\left(\sum_{j=1}^{n} \operatorname{Var}\left(E[T \mid Z_j]\right)\right)^{1/2}} \xrightarrow{d} \mathcal{N}(0, 1).$$

Thus, it suffices to verify Lyapunov's condition (B.13) holds.

Assuming that $L(x^{(1)}) \neq L(x^{(2)})$, we can express *T* as:

$$T = \sum_{(i,j) \in \binom{S_n}{2}} W_i(x^{(1)}) W_j(x^{(2)}) I(Y_i \le Y_j),$$

following the notations defined in Section A.4. Based on the honesty condition, the law of iterated expectations, and the assumption of i.i.d. sampling, we have:

$$E\left[\sum_{\substack{(i,j)\in \binom{S_n\setminus\{1\}}{2}}} W_i(x^{(1)})W_j(x^{(2)})I(Y_i \le Y_j) \middle| Z_1\right]$$

= $E\left[E\left[\sum_{\substack{(i,j)\in \binom{S_n\setminus\{1\}}{2}}} W_i(x^{(1)})W_j(x^{(2)})I(Y_i \le Y_j) \middle| X_i, Z_1\right] \middle| Z_1\right]$
= $E\left[\sum_{\substack{(i,j)\in \binom{S_n\setminus\{1\}}{2}}} W_i(x^{(1)})E[W_j(x^{(2)})I(Y_i \le Y_j) \mid X_i] \middle| Z_1\right].$

Thus,

$$E[T | Z_{1}] - E[T]$$

$$= E\left[W_{1}(x^{(1)})\left(\sum_{j=2}^{n} W_{j}(x^{(2)})I(Y_{1} \leq Y_{j}) - \sum_{j=2}^{n} E[W_{j}(x^{(2)})I(Y_{1} \leq Y_{j}) | X_{1}]\right) | Z_{1}\right] + E\left[\sum_{(i,j)\in\binom{S_{n}}{2}} W_{i}(x^{(1)})E[W_{j}(x^{(2)})I(Y_{i} \leq Y_{j}) | X_{i}] | Z_{1}\right] - E[T]. \quad (B.14)$$

Note that the two right-hand-side terms in Equation (B.14) are both mean-zero. By invoking Jensen's inequality [77], we know that for any random variables A and B,

$$\left|\frac{A+B}{2}\right|^{2+\delta} \leq \frac{|A|^{2+\delta}+|B|^{2+\delta}}{2}.$$

Applying this inequality in our context, we obtain the following bound:

$$2^{-(1+\delta)}E\left[|E\left[T \mid Z_{1}\right] - E\left[T\right]|^{2+\delta}\right] \le E\left[\left|E\left[W_{1}(x^{(1)})\left(\sum_{j=2}^{n}W_{j}(x^{(2)})I(Y_{1} \leq Y_{j}) - \sum_{j=2}^{n}E\left[W_{j}(x^{(2)})I(Y_{1} \leq Y_{j}) \mid X_{1}\right]\right)\right|Z_{1}\right]\right|^{2+\delta}\right] + E\left[\left|E\left[\sum_{(i,j)\in\binom{S_{n}}{2}}W_{i}(x^{(1)})E\left[W_{j}(x^{(2)})I(Y_{i} \leq Y_{j}) \mid X_{i}\right]\right|Z_{1}\right] - E\left[T\right]\right|^{2+\delta}\right].$$
 (B.15)

Furthermore, due to the honesty condition, where $E[W_1(x^{(1)}) | Z_1] = E[W_1(x^{(1)}) | X_1]$, and under the uniform $(2+\delta)$ -moment bounds on the conditional distribution of $I(Y_i \le Y_j)$ given X_i , we deduce:

$$E\left[\left|E\left[W_{1}(x^{(1)})\left(\sum_{j=2}^{n}W_{j}(x^{(2)})I(Y_{1} \leq Y_{j}) - \sum_{j=2}^{n}E\left[W_{j}(x^{(2)})I(Y_{1} \leq Y_{j}) \mid X_{1}\right]\right)\right|Z_{1}\right]\right|^{2+\delta}\right]$$

$$=E\left[E\left[W_{1}(x^{(1)}) \mid X_{1}\right]^{2+\delta}E\left[\left|E\left[I(Y_{1} \leq Y_{2}) \mid Y_{1}\right] - E\left[I(Y_{1} \leq Y_{2}) \mid X_{1}\right]\right|^{2+\delta}|X_{1}\right]\right]$$

$$\leq ME\left[E\left[W_{1}(x^{(1)}) \mid X_{1}\right]^{2+\delta}\right]$$

$$\leq ME\left[E\left[W_{1}(x^{(1)}) \mid X_{1}\right]^{2}\right] \quad \text{since } W_{1}(x^{(1)}) \leq 1.$$
(B.16)

Meanwhile, recognizing that $E[I(Y_i \le Y_j) | X_i] \le 1$ for all $(i, j) \in \binom{S_n}{2}$, and leveraging the i.i.d. sampling assumption along with the law of iterated expectations, we can deduce the following variance bound:

$$\begin{aligned} \operatorname{Var} & \left(E\left[\sum_{(i,j) \in \binom{S_n}{2}} W_i(x^{(1)}) E\left[W_j(x^{(2)}) I(Y_i \le Y_j) \mid X_i \right] \mid Z_1 \right] \right) \\ &= \operatorname{Var} \left(E\left[W_1(x^{(1)}) \mid X_1 \right] E\left[I(Y_1 \le Y_n) \mid X_1 \right] + \sum_{i=2}^{n-1} E\left[W_i(x^{(1)}) E\left[I(Y_i \le Y_n) \mid X_i \right] \right] \right) \\ &= \operatorname{Var} \left(E\left[W_1(x^{(1)}) \mid X_1 \right] E\left[I(Y_1 \le Y_n) \mid X_1 \right] \right) + (n-2) \operatorname{Var} \left(E\left[W_2(x^{(1)}) E\left[I(Y_2 \le Y_n) \mid X_2 \right] \right] \right) \\ &\leq \operatorname{Var} \left(E\left[W_1(x^{(1)}) \mid X_1 \right] \right) \\ &\leq E\left[E\left[W_1(x^{(1)}) \mid X_1 \right]^2 \right]. \end{aligned}$$

Subsequently, we can establish the following bound:

$$E\left[\left|E\left[\sum_{(i,j)\in\binom{S_{n}}{2}}W_{i}(x^{(1)})E\left[W_{j}(x^{(2)})I(Y_{i} \leq Y_{j}) \mid X_{i}\right] \mid Z_{1}\right] - E\left[T\right]\right|^{2+\delta}\right]$$

$$\leq 2^{\delta} \operatorname{Var}\left(E\left[\sum_{(i,j)\in\binom{S_{n}}{2}}W_{i}(x^{(1)})E\left[W_{j}(x^{(2)})I(Y_{i} \leq Y_{j}) \mid X_{i}\right] \mid Z_{1}\right]\right)$$

$$\leq 2^{\delta} E\left[E\left[W_{1}(x^{(1)}) \mid X_{1}\right]^{2}\right].$$
(B.17)

Thus, by combining inequalities (B.15), (B.16), and (B.17), Lyapunov's condition (B.13) reduces to

$$\lim_{n \to \infty} \frac{nE\left[E[W_1(x^{(1)}) \mid X_1]^2\right]}{\left(n\operatorname{Var}\left(E\left[T \mid Z_1\right]\right)\right)^{1+\delta/2}} = 0.$$
(B.18)

From (B.9), we have

$$\operatorname{Var}\left(E\left[T \mid Z_{1}\right]\right) = \Omega\left(E\left[E\left[W_{1}(x^{(1)}) \mid X_{1}\right]^{2}\right]\operatorname{Var}\left(I(Y_{1} \leq Y_{2}) \mid X_{i} = x^{(1)}\right)\right).$$

Given that $Var(I(Y_i \le Y_j) | X_i = x^{(1)}) > 0$ by assumption, we can derive the following using Lemma A.14:

$$\frac{nE\left[E\left[W_{1}(x^{(1)})\mid X_{1}\right]^{2}\right]}{\left(n\operatorname{Var}\left(E\left[T\mid Z_{1}\right]\right)\right)^{1+\delta/2}} \lesssim \left(nE\left[E\left[W_{1}(x^{(1)})\mid X_{1}\right]^{2}\right]\right)^{-\delta/2} \left(\operatorname{Var}(I(Y_{1} \le Y_{2})\mid X_{i} = x^{(1)})\right)^{-(1+\delta/2)} \\ \lesssim \left(\frac{nC_{f,d}}{2ks\log(s)^{d}}\right)^{-\delta/2} \left(\operatorname{Var}(I(Y_{1} \le Y_{2})\mid X_{i} = x^{(1)})\right)^{-(1+\delta/2)} \\ \to 0.$$

Note that the factor of 2 arises from the application of a double-sample tree in the analysis. Thus,

It is important to highlight that honesty is required to achieve pointwise centered asymptotic normality results, as demonstrated in Theorem B.4. Wager and Athey [31] conducted a simulation study showing that honest forest estimators remain unbiased for any sample size n, while classical forest estimators exhibit bias. This bias arises because CART trees often isolate outliers from the main data, resulting in an over-representation of outliers at the corners of the feature space. As n increases, classical forests tend to push outliers further into these corners, intensifying the bias. In contrast, honest trees avoid this issue since they do not know the location of the I-sample outliers when making splits with only the \mathcal{J} -sample.

Proof of Theorem 2.7. Given the results from Theorem B.4, the remaining task is to demonstrate that

$$\frac{E[\widehat{\theta}_{\text{forest},n}(x^{(1)}, x^{(2)})] - \theta(x^{(1)}, x^{(2)})}{\sigma_n(x^{(1)}, x^{(2)})} \to 0$$

in order to establish (2.6); the rest will follow by applying Slutsky's theorem. According to Theorem 2.6, it holds that

$$\left| E[\widehat{\theta}_{\text{forest},n}(x^{(1)}, x^{(2)})] - \theta(x^{(1)}, x^{(2)}) \right| = O\left(s^{-\frac{1}{2}\frac{\log\left((1-\alpha)^{-1}\right)}{\log\left(\alpha^{-1}\right)}\frac{\pi}{d}}\right) = O\left(n^{-\frac{\gamma}{2}\frac{\log\left((1-\alpha)^{-1}\right)}{\log\left(\alpha^{-1}\right)}\frac{\pi}{d}}\right).$$

Moreover, Theorem B.2 and the proof of Theorem B.4 imply that

$$\sigma_n^2(x^{(1)}, x^{(2)}) \gtrsim \frac{s}{n} \frac{C_{f,d}}{16\log(s)^d} \operatorname{Var}(T).$$

Given that T is honest, we have

$$\operatorname{Var}(T) \gtrsim \frac{4\operatorname{Var}(I(Y_1 \le Y_2) \mid X_1 = x^{(1)})}{\left|\left\{i : X_i \in L(x^{(1)})\right\}\right|} \ge \frac{4\operatorname{Var}(I(Y_1 \le Y_2) \mid X_1 = x^{(1)})}{2k - 1}$$

which leads to

$$\sigma_n^2(x^{(1)}, x^{(2)}) \gtrsim \frac{s}{n} \frac{C_{f,d}}{4\log(s)^d} \frac{\operatorname{Var}(I(Y_1 \le Y_2) \mid X_1 = x^{(1)})}{2k - 1} = \Omega\left(n^{\gamma - 1 - \varepsilon}\right)$$

for any $\varepsilon > 0$. Consequently, we have

$$\frac{E[\widehat{\theta}_{\text{forest},n}(x^{(1)}, x^{(2)})] - \theta(x^{(1)}, x^{(2)})}{\sigma_n(x^{(1)}, x^{(2)})} = O\left(n^{\frac{1}{2}\left(1 + \varepsilon - \gamma\left(1 + \frac{\log\left((1 - \alpha)^{-1}\right)}{\log\left(\alpha^{-1}\right)} \frac{\pi}{d}\right)\right)}\right).$$

The bound on the right-hand side converges to 0 for sufficiently small $\varepsilon > 0$, provided that

$$\gamma > \left(1 + \frac{\log\left((1-\alpha)^{-1}\right)}{\log\left(\alpha^{-1}\right)} \frac{\pi}{d}\right)^{-1} = 1 - \left(1 + \frac{\log\left(\alpha^{-1}\right)}{\log\left((1-\alpha)^{-1}\right)} \frac{d}{\pi}\right)^{-1} =: \gamma_{\min}.$$

B.3 UGEE-based FRMs for Longitudinal Networks

In this section, we provide a proof of Theorem 3.2, as Theorem 3.1 is a special case of Theorem 3.2.

Let $\mathbf{x}_i = \{v_i, \mathbf{w}_i\}$, where $\mathbf{w}_i = (w_{i1}, \dots, w_{i(T-1)})^{\top}$, and let $\mathbf{x}_i = \{\mathbf{x}_{i_1}, \mathbf{x}_{i_2}, \mathbf{x}_{i_3}\}$, with $1 \le i \le n$ and $\mathbf{i} = (i_1, i_2, i_3) \in {S_n \choose 3}$. We assume that τ is a *p*-dimensional vector. Then, $\mathbf{U}_n(\tau) = (U_{n,1}(\tau), \dots, U_{n,p}(\tau))^{\top}$ is also a *p*-dimensional random vector. Let $\mathbf{U}_{ni}(\tau)$ denote the summands of $\mathbf{U}_n(\tau)$. Without loss of generality, we consider the normalized quantity ${n \choose 3}^{-1} \sum_{i} \mathbf{U}_{ni}(\tau)$ and continue to denote this normalized quantity as $\mathbf{U}_n(\tau)$ for notational brevity.

B.3.1 Consistency of $\hat{\tau}$

We first assume that α is known. Consider a neighborhood $N(\tau)$ of τ . We assume that $U_n(\tau)$ has continuous second-order partial derivatives, and that there exists an integrable function $g(\mathbf{x_i})$ with $\operatorname{Var}(g(\mathbf{x_i})) < \infty$ such that

$$\left|\frac{\partial^2}{(\partial\lambda_k)^{\gamma}(\partial\lambda_\ell)^{2-\gamma}}U_{n\mathbf{i},j}(\boldsymbol{\lambda})\right| \le g(\mathbf{x}_{\mathbf{i}}) \quad \text{for all} \quad \boldsymbol{\lambda} \in N(\boldsymbol{\tau}), \ \gamma \in \{0,1,2\}, \ 1 \le k, \ell, j \le p, \quad (B.19)$$

where $U_{n\mathbf{i},j}$ denotes the *j*-th component of $\mathbf{U}_{n\mathbf{i}}(\tau)$. Under the regularity conditions (B.19), the UGEE estimator $\hat{\tau}$ is consistent and asymptotically normal.

Under the regularity conditions (B.19), $\operatorname{Var}(U_{n\mathbf{i},j}(\lambda)) < \infty$ for all $\lambda \in N(\tau)$ and $1 \le j \le p$. According to the theory of multivariate *U*-statistics [12], $\mathbf{U}_n(\tau)$ is consistent and asymptotically normal, i.e.,

$$\mathbf{U}_{n}(\boldsymbol{\tau}) \xrightarrow{p} \mathbf{0}, \quad \sqrt{n} \mathbf{U}_{n}(\boldsymbol{\tau}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{U}),$$
 (B.20)

where Σ_U is the $p \times p$ asymptotic variance of $\mathbf{U}_n(\boldsymbol{\tau})$.

Let

$$g_n = {\binom{n}{3}}^{-1} \sum_{\mathbf{i}} g(\mathbf{x}_{\mathbf{i}}), \quad \mathbf{B}(\boldsymbol{\tau}) = E\left[\frac{\partial \mathbf{U}_{n\mathbf{i}}(\boldsymbol{\tau})}{\partial \lambda^{\top}}\right].$$

Again, by the theory of multivariate *U*-statistics, we have:

$$g_n = E\left[g\left(\mathbf{x_i}\right)\right] + o_p(1), \quad \frac{\partial \mathbf{U}_n(\boldsymbol{\tau})}{\partial \boldsymbol{\lambda}^{\top}} = \mathbf{B}(\boldsymbol{\tau}) + \mathbf{o}_p(1), \tag{B.21}$$

where $o_p(1)$ and $o_p(1)$ denote stochastic scalars and matrices, respectively [12]. By the mean-

value theorem for vector-valued functions [78] and the regularity conditions (B.19), we have:

$$\mathbf{U}_{n}(\boldsymbol{\lambda}) = \mathbf{U}_{n}(\boldsymbol{\tau}) + \frac{\partial}{\partial \boldsymbol{\lambda}^{\top}} \mathbf{U}_{n}(\boldsymbol{\tau})(\boldsymbol{\lambda} - \boldsymbol{\tau}) + \mathbf{R}_{n}^{(2)}(\boldsymbol{\lambda} - \boldsymbol{\tau}, \boldsymbol{\xi})$$

$$= \mathbf{U}_{n}(\boldsymbol{\tau}) + \mathbf{B}(\boldsymbol{\tau})(\boldsymbol{\lambda} - \boldsymbol{\tau}) + \mathbf{R}_{n}^{(2)}(\boldsymbol{\lambda} - \boldsymbol{\tau}, \boldsymbol{\xi}) + \mathbf{o}_{p}(1),$$
(B.22)

where

$$\begin{split} \mathbf{R}_{n}^{(2)}(\boldsymbol{\lambda}-\boldsymbol{\tau},\boldsymbol{\xi}) &= \left(R_{n,1}^{(2)}\left(\boldsymbol{\lambda}-\boldsymbol{\tau},\boldsymbol{\xi}_{1}\right),\dots,R_{n,p}^{(2)}\left(\boldsymbol{\lambda}-\boldsymbol{\tau},\boldsymbol{\xi}_{p}\right)\right)^{\top},\\ R_{n,j}^{(2)}\left(\boldsymbol{\lambda}-\boldsymbol{\tau},\boldsymbol{\xi}_{j}\right) &= \frac{1}{2}\sum_{\ell=1}^{p}\sum_{k=1}^{p}\sum_{\gamma=0}^{2}\binom{2}{\gamma}\frac{\partial^{2}}{(\partial\lambda_{k})^{\gamma}(\partial\lambda_{\ell})^{2-\gamma}}U_{n,j}\left(\boldsymbol{\xi}_{j}\right)\left(\lambda_{k}-\boldsymbol{\tau}_{k}\right)^{\gamma}\left(\lambda_{\ell}-\boldsymbol{\tau}_{\ell}\right)^{2-\gamma},\\ \boldsymbol{\xi} &= \left\{\boldsymbol{\xi}_{j}: 1 \leq j \leq p\right\}, \quad \boldsymbol{\xi}_{j} \in N(\boldsymbol{\tau}), \quad 1 \leq j \leq p. \end{split}$$

By the regularity conditions (B.19) and Equation (B.21), we can express $R_{n,j}^{(2)} \left(\lambda - \tau, \xi_j \right)$ as:

$$\begin{split} R_{n,j}^{(2)}\left(\boldsymbol{\lambda}-\boldsymbol{\tau},\boldsymbol{\xi}_{j}\right) &= \frac{1}{2}\sum_{\ell=1}^{p}\sum_{k=1}^{p}\sum_{\gamma=0}^{p}\binom{2}{\gamma}\boldsymbol{\zeta}_{k,\ell,\gamma,j}\left(E\left[g\left(\mathbf{x_{i}}\right)\right]+o_{p}(1)\right)\left(\boldsymbol{\lambda}_{k}-\boldsymbol{\tau}_{k}\right)^{\gamma}\left(\boldsymbol{\lambda}_{\ell}-\boldsymbol{\tau}_{\ell}\right)^{2-\gamma}\right.\\ &= \frac{1}{2}\sum_{\ell=1}^{p}\sum_{k=1}^{p}\sum_{\gamma=0}^{p}\binom{2}{\gamma}\boldsymbol{\zeta}_{k,\ell,\gamma,j}E\left[g\left(\mathbf{x_{i}}\right)\right]\left(\boldsymbol{\lambda}_{k}-\boldsymbol{\tau}_{k}\right)^{\gamma}\left(\boldsymbol{\lambda}_{\ell}-\boldsymbol{\tau}_{\ell}\right)^{2-\gamma}+o_{p}(1)\right.\\ &= H_{j}\left(\boldsymbol{\lambda}-\boldsymbol{\tau},\boldsymbol{\zeta}_{j}\right)+o_{p}(1),\\ &\left|\boldsymbol{\zeta}_{k,\ell,\gamma,j}\right| < 1, \quad \gamma \in \{0,1,2\}, \quad 1 \le k, \ell, j \le p, \end{split}$$

where $\boldsymbol{\zeta} = \{\boldsymbol{\zeta}_j; 1 \le j \le p\}$ and $\boldsymbol{\zeta}_j = \{\boldsymbol{\zeta}_{k,\ell,\gamma,j}; 1 \le k, \ell \le p\}$. Further, for $1 \le j \le p$, we have:

$$\left| H_{j}\left((\boldsymbol{\lambda} - \boldsymbol{\tau}), \boldsymbol{\zeta}_{j} \right) \right| \leq \frac{1}{2} \sum_{l=1}^{p} \sum_{k=1}^{p} \sum_{\gamma=0}^{2} {2 \choose \gamma} \left| \boldsymbol{\zeta}_{k,\ell,\gamma,j} \right| \left| E\left[g\left(\mathbf{x}_{\mathbf{i}} \right) \right] (\boldsymbol{\lambda}_{k} - \boldsymbol{\tau}_{k})^{\gamma} (\boldsymbol{\lambda}_{\ell} - \boldsymbol{\tau}_{\ell})^{2-\gamma} \right|$$

$$= \frac{1}{2} \sum_{\ell=1}^{p} \sum_{k=1}^{p} \sum_{\gamma=0}^{2} {2 \choose \gamma} \left| E\left[g\left(\mathbf{x}_{\mathbf{i}} \right) \right] \right| \left| (\boldsymbol{\lambda}_{k} - \boldsymbol{\tau}_{k})^{\gamma} (\boldsymbol{\lambda}_{\ell} - \boldsymbol{\tau}_{\ell})^{2-\gamma} \right|.$$
(B.23)

Let

$$\mathbf{H}(\boldsymbol{\lambda}-\boldsymbol{\tau},\boldsymbol{\zeta}) = \left(H_1\left(\boldsymbol{\lambda}-\boldsymbol{\tau},\boldsymbol{\zeta}_1\right),\ldots,H_p\left(\boldsymbol{\lambda}-\boldsymbol{\tau},\boldsymbol{\zeta}_p\right)\right)^{\top}.$$

Then we can express $\mathbf{R}_n^{(2)}(\lambda - \tau, \boldsymbol{\xi})$ as:

$$\mathbf{R}_{n}^{(2)}(\boldsymbol{\lambda}-\boldsymbol{\tau},\boldsymbol{\xi}) = \mathbf{H}(\boldsymbol{\lambda}-\boldsymbol{\tau},\boldsymbol{\zeta}) + \mathbf{o}_{p}(1).$$
(B.24)

By Equations (B.20), (B.22) and (B.24), we have:

$$\mathbf{U}_{n}(\boldsymbol{\lambda}) = \mathbf{B}(\boldsymbol{\tau})(\boldsymbol{\lambda} - \boldsymbol{\tau}) + \mathbf{H}(\boldsymbol{\lambda} - \boldsymbol{\tau}, \boldsymbol{\zeta}) + \mathbf{o}_{p}(1).$$
(B.25)

Multiplying both sides of Equation (B.25) by $\mathbf{F}(\tau) = \mathbf{B}^{-1}(\tau)$, we obtain:

$$\mathbf{W}_{n}(\boldsymbol{\lambda}) = \mathbf{F}(\boldsymbol{\tau})\mathbf{U}_{n}(\boldsymbol{\lambda}) = \boldsymbol{\lambda} - \boldsymbol{\tau} + \mathbf{F}(\boldsymbol{\tau})\mathbf{H}(\boldsymbol{\lambda} - \boldsymbol{\tau}, \boldsymbol{\zeta}) + \mathbf{o}_{p}(1)$$

$$= \boldsymbol{\lambda} - \boldsymbol{\tau} + \mathbf{G}(\boldsymbol{\lambda}, \boldsymbol{\tau}, \boldsymbol{\zeta}) + \mathbf{z}_{n},$$
(B.26)

where $\mathbf{G}(\lambda, \tau, \zeta) = \mathbf{F}(\tau)\mathbf{H}(\lambda - \tau, \zeta)$ and $\mathbf{z}_n = \mathbf{o}_p(1)$. Below we construct a consistent estimator $\hat{\tau}$ of τ based on $\mathbf{W}_n(\lambda) = \mathbf{0}$, which implies that $\hat{\tau}$ is also a consistent estimator of τ based on $\mathbf{U}_n(\lambda) = \mathbf{0}$.

$$\lambda_{1m} = \tau - \frac{1}{m}, \quad \lambda_{2m} = \tau + \frac{1}{m},$$

$$\Omega_{n,\frac{1}{m}} = \left\{ \omega; \max \left| z_{nj}(\omega) \right| < \frac{1}{4m}, 1 \le j \le p \right\},$$

$$C = \max_{1 \le i, j \le p} \left| F_{ij}(\tau) \right|, \quad m \ge 1,$$

(B.27)

where $\mathbf{1}_p$ denotes a $p \times 1$ column vector of 1's, z_{nj} is the *j*-th component of \mathbf{z}_n , and $F_{ij}(\tau)$ is the (i, j)-th element of $\mathbf{F}(\tau)$. For λ_{sm} (s = 1, 2), by inequality (B.23), we have for each component $G_j(\lambda_{sm}, \tau, \zeta_j)$ of $\mathbf{G}(\lambda_{sm}, \tau, \zeta)$:

$$\begin{aligned} \left| G_{j} \left(\lambda_{sm}, \boldsymbol{\tau}, \boldsymbol{\zeta}_{j} \right) \right| &\leq C \sum_{j=1}^{p} \left| H_{j} \left(\left(\pm \frac{1}{m} \right) \mathbf{1}_{p}, \boldsymbol{\zeta}_{j} \right) \right| \\ &\leq \frac{C}{2} \sum_{j=1}^{p} \sum_{\ell=1}^{p} \sum_{k=1}^{p} \sum_{\gamma=0}^{2} \binom{2}{\gamma} \left| E \left[g \left(\mathbf{x}_{i} \right) \right] \right| \left(\frac{1}{m} \right)^{2} \\ &= D \left(\frac{1}{m} \right)^{2}, \end{aligned}$$

where

$$D = \frac{C}{2} \sum_{j=1}^{p} \sum_{\ell=1}^{p} \sum_{k=1}^{p} \sum_{\gamma=0}^{2} {\binom{2}{\gamma}} |E[g(\mathbf{x}_{i})]|.$$

Thus, for sufficiently large m (e.g., m > 4D), we have

$$\left|G_{j}\left(\boldsymbol{\lambda}_{sm},\boldsymbol{\tau},\boldsymbol{\zeta}_{j}\right)\right| < \frac{1}{4m}, \quad s = 1, 2, \quad 1 \le j \le p, \quad \boldsymbol{\lambda}_{sm} \in N(\boldsymbol{\tau}).$$
(B.28)

Since $\mathbf{z}_n = \mathbf{o}_p(1)$, for such *m*, we can find $N_{\frac{1}{m}}$ such that

$$P\left(\Omega_{n,\frac{1}{m}}\right) \ge 1 - \frac{1}{m}, \quad \text{for all } n \ge N_{\frac{1}{m}}.$$
 (B.29)

Let

For $\omega \in \Omega_{n,\frac{1}{m}}$, it follows from Equations (B.26), (B.27) and (B.28) that for all $1 \le j \le p$:

$$W_{n,j}(\lambda_{1m}) = -\frac{1}{m} + G_j(\lambda_{1m}, \tau, \zeta_j) + z_{nj} < -\frac{1}{m} + \frac{1}{4m} + \frac{1}{4m} = -\frac{1}{2m} < 0,$$

$$W_{n,j}(\lambda_{2m}) = \frac{1}{m} + G_j(\lambda_{2m}, \tau, \zeta_j) + z_{nj} > \frac{1}{m} - \frac{1}{4m} - \frac{1}{4m} = \frac{1}{2m} > 0.$$

Since $\mathbf{W}_n(\lambda)$ is continuous in λ , there exists $\boldsymbol{\eta}_m \in N(\boldsymbol{\tau})$ such that $\mathbf{W}_n(\boldsymbol{\eta}_m) = \mathbf{0}$ by the intermediate value theorem.

For
$$n \ge N_{\frac{1}{m}}$$
, let

$$\widehat{\boldsymbol{\tau}}^{(m)} = \inf \left\{ \left\| \boldsymbol{\eta}_m \right\| : \left\{ \boldsymbol{\eta}_m \in N(\boldsymbol{\tau}) \right\} \cap \left\{ \mathbf{w}_n \left(\boldsymbol{\eta}_m \right) = \mathbf{0} \right\} \right\}.$$

Then, $\{\widehat{\tau}^{(m)}: m \ge 1\}$ is a sequence of random vectors. Further, by inequality (B.29), we have:

$$P\left(\left\|\widehat{\boldsymbol{\tau}}^{(m)} - \boldsymbol{\tau}\right\| \ge \frac{1}{m}\right) \le P\left(\Omega_{n,\frac{1}{m}}^c \text{ for } n \ge N_{\frac{1}{m}}\right) \le \frac{1}{m} \to 0.$$
(B.30)

Thus, $\widehat{\tau} := \lim_{m \to \infty} \widehat{\tau}^{(m)}$ is a consistent estimate of τ .

We now consider the case where α , a *r*-dimensional vector, is unknown and is estimated by a \sqrt{n} -consistent estimator, $\hat{\alpha}$, i.e., $\sqrt{n}(\hat{\alpha} - \alpha) = \mathbf{O}_p(1)$, where $\mathbf{O}_p(1)$ denotes a vector of the same dimension as α with all its components stochastically bounded [12]. We consider λ_1 to be the *p*-dimensional vector corresponding to τ and λ_2 to be the *r*-dimensional vector corresponding α . We assume that the regularity conditions (B.19) hold with respect to the augmented vector $(\lambda_1^{\top}, \lambda_2^{\top})^{\top}$. Similar to Equations (B.22) and (B.25) for the case of known α , we have:

$$\mathbf{U}_{n}(\lambda_{1},\lambda_{2}) = \mathbf{U}_{n}(\tau,\alpha) + \frac{\partial}{\partial\lambda_{1}^{\top}} \mathbf{U}_{n}(\tau,\alpha) (\lambda_{1}-\tau) + \frac{\partial}{\partial\lambda_{2}^{\top}} \mathbf{U}_{n}(\tau,\alpha) (\lambda_{2}-\alpha) + \mathbf{H}(\lambda_{1}-\tau,\lambda_{2}-\alpha,\zeta)$$
$$= \mathbf{B}(\tau,\alpha) (\lambda_{1}-\tau) + \frac{\partial}{\partial\lambda_{2}^{\top}} \mathbf{U}_{n}(\tau,\alpha) (\lambda_{2}-\alpha) + \mathbf{H}(\lambda_{1}-\tau,\lambda_{2}-\alpha,\zeta) + \mathbf{o}_{p}(1),$$
(B.31)

where $\mathbf{B}(\tau, \alpha)$ and $\mathbf{H}(\lambda_1 - \tau, \lambda_2 - \alpha, \zeta)$ in this case are given by:

$$\mathbf{B}(\boldsymbol{\tau},\boldsymbol{\alpha}) = E\left[\frac{\partial \mathbf{U}_{n\mathbf{i}}(\boldsymbol{\tau},\boldsymbol{\alpha})}{\partial \lambda_{1}^{\mathsf{T}}}\right],$$

$$\mathbf{H}(\lambda_{1}-\boldsymbol{\tau},\lambda_{2}-\boldsymbol{\alpha},\boldsymbol{\zeta}) = \mathbf{H}_{1}(\lambda_{1}-\boldsymbol{\tau},\boldsymbol{\zeta}) + \mathbf{H}_{2}(\lambda_{1}-\boldsymbol{\tau},\lambda_{2}-\boldsymbol{\alpha},\boldsymbol{\zeta}) + \mathbf{H}_{3}(\lambda_{2}-\boldsymbol{\alpha},\boldsymbol{\zeta}),$$

$$H_{1,j}\left(\lambda_{1}-\boldsymbol{\tau},\boldsymbol{\zeta}_{j}\right) = \frac{1}{2}\sum_{\ell=1}^{p}\sum_{k=1}^{p}\sum_{\gamma=0}^{2}\binom{2}{\gamma}\zeta_{k,\ell,\gamma,j}\left(E\left[g\left(\mathbf{x}_{\mathbf{i}}\right)\right]+o_{p}(1)\right)(\lambda_{1k}-\boldsymbol{\tau}_{k})^{\gamma}(\lambda_{1\ell}-\boldsymbol{\tau}_{\ell})^{2-\gamma},$$

$$H_{2,j}\left(\lambda_{1}-\boldsymbol{\tau},\lambda_{2}-\boldsymbol{\alpha},\boldsymbol{\zeta}_{j}\right) = \sum_{\ell=1}^{r}\sum_{k=1}^{p}\sum_{\gamma=0}^{2}\binom{2}{\gamma}\zeta_{k,\ell,\gamma,j}\left(E\left[g\left(\mathbf{x}_{\mathbf{i}}\right)\right]+o_{p}(1)\right)(\lambda_{1k}-\boldsymbol{\tau}_{k})^{\gamma}(\lambda_{2\ell}-\boldsymbol{\alpha}_{\ell})^{2-\gamma},$$

$$H_{3,j}\left(\lambda_{2}-\boldsymbol{\alpha},\boldsymbol{\zeta}_{j}\right) = \frac{1}{2}\sum_{\ell=1}^{r}\sum_{k=1}^{r}\sum_{\gamma=0}^{2}\binom{2}{\gamma}\zeta_{k,\ell,\gamma,j}\left(E\left[g\left(\mathbf{x}_{\mathbf{i}}\right)\right]+o_{p}(1)\right)(\lambda_{2k}-\boldsymbol{\alpha}_{k})^{\gamma}(\lambda_{2\ell}-\boldsymbol{\alpha}_{\ell})^{2-\gamma}.$$
(B.32)

By substituting $\widehat{\alpha}$ for λ_2 , we have:

$$H_{2,j}\left(\lambda_1 - \tau, \widehat{\alpha} - \alpha, \zeta_j\right) = \sum_{k=1}^p o_p(1) \left(\lambda_{1k} - \tau_k\right), \quad H_{3,j}\left(\widehat{\alpha} - \alpha, \zeta_j\right) = o_p(1).$$
(B.33)

Thus, by Equations (B.32) and (B.33), $\mathbf{H}(\lambda_1 - \tau, \lambda_2 - \alpha, \zeta)$ can be expressed as:

$$\mathbf{H}(\lambda_1 - \boldsymbol{\tau}, \widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}, \boldsymbol{\zeta}) = \mathbf{H}_1(\lambda_1 - \boldsymbol{\tau}, \boldsymbol{\zeta}) + \mathbf{o}_p(1).$$
(B.34)

By the definition of $\mathbf{U}_{n\mathbf{i}}(\tau, \alpha)$, we have:

$$\frac{\partial}{\partial \lambda_2^{\top}} \mathbf{U}_n(\boldsymbol{\tau}, \boldsymbol{\alpha}) = {\binom{n}{3}}^{-1} \sum_{\mathbf{i}} \frac{\partial}{\partial \lambda_2^{\top}} \mathbf{U}_{n\mathbf{i}}(\boldsymbol{\tau}, \boldsymbol{\alpha}) = {\binom{n}{3}}^{-1} \sum_{\mathbf{i}} \frac{\partial}{\partial \lambda_2^{\top}} \left(\mathbf{D}_{\mathbf{i}}(\boldsymbol{\tau}) \mathbf{V}_{\mathbf{i}}^{-1}(\boldsymbol{\tau}, \boldsymbol{\alpha}) \mathbf{S}_{\mathbf{i}}(\boldsymbol{\tau}) \right).$$

Note that by the law of iterated expectations, we have:

$$E\left[\frac{\partial}{\partial\lambda_{2}^{\top}}\left(\mathbf{D}_{\mathbf{i}}(\tau)\mathbf{V}_{\mathbf{i}}(\tau,\alpha)^{-1}\mathbf{S}_{\mathbf{i}}(\tau)\right)\right] = E\left[E\left[\left.\frac{\partial}{\partial\lambda_{2}^{\top}}\left(\mathbf{D}_{\mathbf{i}}(\tau)\mathbf{V}_{\mathbf{i}}(\tau,\alpha)^{-1}\mathbf{S}_{\mathbf{i}}(\tau)\right)\middle|\mathbf{x}_{\mathbf{i}}\right]\right]\right]$$
$$= E\left[E\left[\left.\frac{\partial}{\partial\lambda_{2}^{\top}}\left(\mathbf{D}_{\mathbf{i}}(\tau)\mathbf{V}_{\mathbf{i}}^{-1}(\tau,\alpha)\right)\mathbf{S}_{\mathbf{i}}(\tau)\middle|\mathbf{x}_{\mathbf{i}}\right]\right]\right]$$
$$= E\left[\frac{\partial}{\partial\lambda_{2}^{\top}}\left(\mathbf{D}_{\mathbf{i}}(\tau)\mathbf{V}_{\mathbf{i}}^{-1}(\tau,\alpha)\right)E\left[\mathbf{S}_{\mathbf{i}}(\tau)\middle|\mathbf{x}_{\mathbf{i}}\right]\right]$$
$$= \mathbf{0}.$$

Therefore, by the theory of multivariate U-statistics,

$$\frac{\partial}{\partial \lambda_2^{\top}} \mathbf{U}_n(\boldsymbol{\tau}, \boldsymbol{\alpha}) = \mathbf{o}_p(1). \tag{B.35}$$

It follows from Equations (B.31), (B.34) and (B.35) that

$$\mathbf{U}_n(\boldsymbol{\lambda}_1, \widehat{\boldsymbol{\alpha}}) = \mathbf{B}(\boldsymbol{\tau}, \boldsymbol{\alpha}) (\boldsymbol{\lambda}_1 - \boldsymbol{\tau}) + \mathbf{H}_1 (\boldsymbol{\lambda}_1 - \boldsymbol{\tau}, \boldsymbol{\zeta}) + \mathbf{o}_p(1).$$

The term $\mathbf{H}_1(\lambda_1 - \tau, \zeta)$ has the same properties as $\mathbf{H}(\lambda - \tau, \zeta)$ in Equation (B.25). Thus, the consistency of $\hat{\tau}$ follows from the same argument applied in the steps between Equations (B.26) and (B.30) for the case of known α .

B.3.2 Asymptotic Normality of $\hat{\tau}$

Multiplying both sides of Equation (B.31) by \sqrt{n} and using Equation (B.35), we obtain:

$$\sqrt{n}\mathbf{U}_{n}(\lambda_{1},\lambda_{2}) = \sqrt{n}\mathbf{U}_{n}(\tau,\alpha) + \sqrt{n}\frac{\partial}{\partial\lambda_{1}^{\top}}\mathbf{U}_{n}(\tau,\alpha)(\lambda_{1}-\tau) + \sqrt{n}\frac{\partial}{\partial\lambda_{2}^{\top}}\mathbf{U}_{n}(\tau,\alpha)(\lambda_{2}-\alpha) + \sqrt{n}\mathbf{H}(\lambda_{1}-\tau,\lambda_{2}-\alpha,\zeta)$$

$$= \sqrt{n}\mathbf{U}_{n}(\tau,\alpha) + \mathbf{B}(\tau,\alpha)\sqrt{n}(\lambda_{1}-\tau) + \sqrt{n}\mathbf{H}(\lambda_{1}-\tau,\lambda_{2}-\alpha,\zeta)$$

$$+ \mathbf{o}_{p}(1)\sqrt{n}(\lambda_{2}-\alpha) + \mathbf{o}_{p}(1).$$
(B.36)

By substituting $\hat{\tau}$ for λ_1 and $\hat{\alpha}$ for λ_2 , along with the fact that $\mathbf{U}_n(\hat{\tau}, \hat{\alpha}) = \mathbf{0}$ and that both $\hat{\tau}$ and $\hat{\alpha}$ are \sqrt{n} -consistent, Equation (B.36) can be expressed as follows after some rearrangement:

$$\sqrt{n}(\widehat{\boldsymbol{\tau}} - \boldsymbol{\tau}) = -\mathbf{B}^{-1}(\boldsymbol{\tau}, \boldsymbol{\alpha})\sqrt{n}\mathbf{U}_n(\boldsymbol{\tau}, \boldsymbol{\alpha}) + \mathbf{o}_p(1).$$
(B.37)

Since $\sqrt{n}\mathbf{U}_n(\tau, \alpha)$ has the same asymptotic distribution as its projection $\sqrt{n}\widehat{\mathbf{U}}_n(\tau, \alpha)$:

$$\sqrt{n}\widehat{\mathbf{U}}_{n}(\boldsymbol{\tau},\boldsymbol{\alpha}) = \frac{\sqrt{n}}{n}\sum_{i=1}^{n}\widehat{\mathbf{U}}_{n,i}(\boldsymbol{\tau},\boldsymbol{\alpha}) = \frac{\sqrt{n}}{n}\sum_{i=1}^{n}3E\left[\mathbf{U}_{n,\mathbf{i}}(\boldsymbol{\tau},\boldsymbol{\alpha})\mid\mathbf{x}_{i}\right],$$

we can rewrite Equation (B.37) as:

$$\sqrt{n}(\widehat{\boldsymbol{\tau}}-\boldsymbol{\tau}) = -\mathbf{B}(\boldsymbol{\tau},\boldsymbol{\alpha})^{-1}\sqrt{n}\widehat{\mathbf{U}}_n(\boldsymbol{\tau},\boldsymbol{\alpha}) + \mathbf{o}_p(1).$$

As $\sqrt{n}\widehat{\mathbf{U}}_n(\tau, \alpha)$ converges to a normal distribution, specifically $\mathcal{N}(\mathbf{0}, 9\Sigma(\tau, \alpha))$, it follows from Slutsky's theorem that

$$\sqrt{n}\,(\widehat{\boldsymbol{\tau}}-\boldsymbol{\tau})\stackrel{d}{\to}\mathcal{N}\left(\boldsymbol{0},9\mathbf{B}^{-1}(\boldsymbol{\tau},\boldsymbol{\alpha})\boldsymbol{\Sigma}(\boldsymbol{\tau},\boldsymbol{\alpha})\mathbf{B}^{-1}(\boldsymbol{\tau},\boldsymbol{\alpha})\right).$$

B.3.3 Asymptotic Efficiency of $\hat{\tau}$

Under the regularity conditions (B.19), $\hat{\tau}$ is an asymptotically linear (AL) estimator of τ , and it can be expressed as follows:

$$\sqrt{n}(\widehat{\boldsymbol{\tau}}-\boldsymbol{\tau}) = \sqrt{n} {\binom{n}{3}}^{-1} \sum_{\mathbf{i}} \boldsymbol{\varphi}(\mathbf{x}_{\mathbf{i}};\boldsymbol{\tau}) + \mathbf{o}_{p}(1),$$

where $\varphi(\mathbf{x}_i; \tau)$ is a measurable influence function with mean zero and a finite, nonsingular covariance matrix $E[\varphi\varphi^{\top}]$. The asymptotic variance of $\hat{\tau}$ is given by:

$$\Sigma_{\tau} = 9 \operatorname{Var} \left(E \left[\varphi(\mathbf{x}_{i}; \tau) \mid \mathbf{x}_{i} \right] \right).$$

As in the case of conventional semiparametric models for within-subject attributes, we restrict our consideration of asymptotic efficiency to regular and asymptotically linear (RAL) estimators [21]. Under the regularity conditions (B.19), $\hat{\tau}$ is a RAL estimator [60].

For two RAL estimators of τ with influence functions $\varphi(\mathbf{x}_i; \tau)$ and $\psi(\mathbf{x}_i; \tau), \varphi(\mathbf{x}_i; \tau)$ is more efficient than $\psi(\mathbf{x}_i; \tau)$ if

$$\operatorname{Var}\left(E\left[\psi\left(\mathbf{x}_{\mathbf{i}};\boldsymbol{\tau}\right) \mid \mathbf{x}_{i}\right]\right) - \operatorname{Var}\left(E\left[\varphi\left(\mathbf{x}_{\mathbf{i}};\boldsymbol{\tau}\right) \mid \mathbf{x}_{i}\right]\right)$$

is positive semi-definite. By considering parametric submodels, akin to conventional semiparametric models for within-subject attributes, we can similarly define an efficiency-bound UGEE estimator, $\hat{\tau}$, and show that the $\hat{\tau}$ obtained by solving the UGEE achieves the efficiency bound if the working variance, V_i , equals the true variance [60]. Note that Liu, T. Lin, et al. [60] consider FRMs defined by two subjects, but the theory can be extended to FRMs defined by three subjects.

 \Box

Bibliography

- Liu, J., Zhang, X., Chen, T., Wu, T.-C., Lin, T., Jiang, L., Lang, S., Liu, L., Natarajan, L., Tu, J. X., Kosciolek, T., Morton, J., Nguyen, T. T., Schnabl, B., Knight, R., Feng, C., Zhong, Y., & Tu, X. M. (2022). A semiparametric model for between-subject attributes: Applications to beta-diversity of microbiome data. *Biometrics*, 78(3), 950–962.
- [2] Vu, T., Lin, T., Zou, J., Novitsky, V., Tu, X., & De Gruttola, V. (2022). Estimating viral genetic linkage rates in the presence of missing data. *arXiv preprint arXiv:2203.12779*.
- [3] Chen, T., Kowalski, J., Chen, R., Wu, P., Zhang, H., Feng, C., & Tu, X. M. (2016). Rankpreserving regression: A more robust rank regression model against outliers. *Statistics in Medicine*, 35(19), 3333–3346.
- [4] Keeling, M. J., & Eames, K. T. (2005). Networks and epidemic models. *Journal of the Royal Society Interface*, 2(4), 295–307.
- [5] Kamarudin, A. N., Cox, T., & Kolamunnage-Dona, R. (2017). Time-dependent ROC curve analysis in medical research: Current methods and applications. *BMC Medical Research Methodology*, 17(1), 1–19.
- [6] Gao, B., Wu, T.-C., Lang, S., Jiang, L., Duan, Y., Fouts, D. E., Zhang, X., Tu, X.-M., & Schnabl, B. (2022). Machine learning applied to omics datasets predicts mortality in patients with alcoholic hepatitis. *Metabolites*, 12(1), 41.
- [7] Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8), 861–874.
- [8] Pepe, M. S. (2003). *The statistical evaluation of medical tests for classification and prediction*. Oxford University Press, USA.
- [9] Pepe, M. S., Janes, H., Longton, G., Leisenring, W., & Newcomb, P. (2004). Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. *American Journal of Epidemiology*, 159(9), 882–890.

- [10] Pepe, M. S., Cai, T., & Longton, G. (2006). Combining predictors for classification using the area under the receiver operating characteristic curve. *Biometrics*, 62(1), 221–229.
- [11] Van der Vaart, A. W. (2000). Asymptotic statistics (Vol. 3). Cambridge University Press.
- [12] Kowalski, J., & Tu, X. M. (2008). *Modern applied U-statistics*. John Wiley & Sons.
- [13] Nelder, J. A., & Wedderburn, R. W. (1972). Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, 135(3), 370–384.
- [14] Ma, S., & Huang, J. (2007). Combining multiple markers for classification using ROC. *Biometrics*, 63(3), 751–757.
- [15] Heller, G., Seshan, V. E., Moskowitz, C. S., & Gönen, M. (2017). Inference for the difference in the area under the ROC curve derived from nested binary regression models. *Biostatistics*, 18(2), 260–274.
- [16] Van der Vaart, A. W., & Wellner, J. A. (1996). *Weak convergence and empirical processes: With applications to statistics*. Springer.
- [17] Ghoudi, K., & Rémillard, B. (1998). Empirical processes based on pseudo-observations. In *Asymptotic methods in probability and statistics* (pp. 171–197). Elsevier.
- [18] Mann, H. B., & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, *18*(1), 50–60.
- [19] Van Der Vaart, A. W., & Wellner, J. A. (2007). Empirical processes indexed by estimated functions. *Lecture Notes-Monograph Series*, *55*, 234–252.
- [20] R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/
- [21] Tsiatis, A. A. (2006). Semiparametric theory and missing data. Springer.
- [22] Harris, S., & Harris, D. (2015). *Digital design and computer architecture*. Morgan Kaufmann.
- [23] Hand, D. J., & Till, R. J. (2001). A simple generalisation of the area under the ROC curve for multiple class classification problems. *Machine Learning*, *45*, 171–186.
- [24] Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*(1), 55–67.

- [25] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1), 267–288.
- [26] Breiman, L. (2001). Random forests. *Machine Learning*, 45, 5–32.
- [27] Hengl, T., Nussbaum, M., Wright, M. N., Heuvelink, G. B., & Gräler, B. (2018). Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. *PeerJ*, *6*, e5518.
- [28] Brence, M., & Brown, D. E. (2006). Improving the robust random forest regression algorithm. *Systems and Information Engineering Technical Papers*, 1–18.
- [29] Brence, J. R. (2004). Analysis of robust measures for random forest regression. University of Virginia.
- [30] Li, A. H., & Martin, A. (2017). Forest-type regression with general losses and robust forest. In *Proceedings of the International Conference on Machine Learning* (pp. 2091–2100). PMLR.
- [31] Wager, S., & Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, *113*(523), 1228–1242.
- [32] Lin, T. (2023). Novel machine learning and statistical models for high dimensional and observational study data: Applications to HIV genetic linkage network, fMRI and survey data. University of California, San Diego.
- [33] Breiman, L. (2017). *Classification and regression trees*. Routledge.
- [34] Mentch, L., & Hooker, G. (2016). Quantifying uncertainty in random forests via confidence intervals and hypothesis tests. *Journal of Machine Learning Research*, *17*(26), 1–41.
- [35] Biau, G., Devroye, L., & Lugosi, G. (2008). Consistency of random forests and other averaging classifiers. *Journal of Machine Learning Research*, 9(9), 2015–2033.
- [36] Scornet, E., Biau, G., & Vert, J.-P. (2015). Consistency of random forests. *The Annals of Statistics*, *43*(4), 1716–1741.
- [37] Wager, S., & Walther, G. (2015). Adaptive concentration of regression trees, with application to random forests. *arXiv preprint arXiv:1503.06388*.
- [38] Meinshausen, N., & Ridgeway, G. (2006). Quantile regression forests. *Journal of Machine Learning Research*, 7(6), 983–999.

- [39] Athey, S., Tibshirani, J., & Wager, S. (2019). Generalized random forests. *The Annals of Statistics*, 47(2), 1148–1178.
- [40] Efron, B. (1992). Bootstrap methods: Another look at the jackknife. In *Breakthroughs in statistics: Methodology and distribution* (pp. 569–593). Springer.
- [41] Floudas, C. A., & Pardalos, P. M. (2008). *Encyclopedia of optimization*. Springer Science & Business Media.
- [42] Kreindler, G. E., & Young, H. P. (2014). Rapid innovation diffusion in social networks. *Proceedings of the National Academy of Sciences*, *111*(supplement_3), 10881–10888.
- [43] Fauci, A. S. (2001). Infectious diseases: Considerations for the 21st century. *Clinical Infectious Diseases*, *32*(5), 675–685.
- [44] Fowler, J. H. (2006). Connecting the congress: A study of cosponsorship networks. *Political Analysis*, *14*(4), 456–487.
- [45] Newman, M. (2018). *Networks*. Oxford University Press.
- [46] Goodreau, S. M., Kitts, J. A., & Morris, M. (2009). Birds of a feather, or friend of a friend? Using exponential random graph models to investigate adolescent social networks. *Demography*, 46(1), 103–125.
- [47] Christakis, N. A., & Fowler, J. H. (2009). Social network visualization in epidemiology. *Norsk Epidemiologi*, 19(1), 5.
- [48] Shiino, T., Otani, M., Kikuchi, T., Yoshimura, K., & Sugiura, W. (2023). Signs of late HIV diagnosis and outbreaks in transmission networks in Japan. In Special Issue: Abstracts from the 2023 Conference on Retroviruses and Opportunistic Infections. Topics in Antiviral Medicine, 31(2), 938. https://www.croiconference.org/abstract/signs-of-late-hiv-diagnosis-and-outbreaks-in-transmission-networks-in-japan/
- [49] Little, S. J., Kosakovsky Pond, S. L., Anderson, C. M., Young, J. A., Wertheim, J. O., Mehta, S. R., May, S., & Smith, D. M. (2014). Using HIV networks to inform real time prevention interventions. *PLOS ONE*, 9(6), e98443.
- [50] Song, R., Hall, H. I., Green, T. A., Szwarcwald, C. L., & Pantazis, N. (2017). Using CD4 data to estimate HIV incidence, prevalence, and percent of undiagnosed infections in the United States. *JAIDS Journal of Acquired Immune Deficiency Syndromes*, 74(1), 3–9.
- [51] Tang, M. E., Goyal, R., Anderson, C. M., Mehta, S. R., & Little, S. J. (2023). Assessing the reliability of the CD4 depletion model in the presence of Ending the HIV Epidemic initiatives. *AIDS*, 37(10), 1617–1624.

- [52] Gopalappa, C., Farnham, P. G., Chen, Y.-H., & Sansom, S. L. (2017). Progression and transmission of HIV/AIDS (PATH 2.0) a new, agent-based model to estimate HIV transmissions in the United States. *Medical Decision Making*, 37(2), 224–233.
- [53] Katz, I. T., & Maughan-Brown, B. (2017). Improved life expectancy of people living with HIV: Who is left behind? *The Lancet HIV*, *4*(8), e324–e326.
- [54] Liang, K.-Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1), 13–22.
- [55] Breslow, N. E., & Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88(421), 9–25.
- [56] Wasserman, S., & Faust, K. (1994). *Social network analysis: Methods and applications*. Cambridge University Press.
- [57] Schweinberger, M., Krivitsky, P. N., Butts, C. T., & Stewart, J. R. (2020). Exponentialfamily models of random graphs. *Statistical Science*, *35*(4), 627–662.
- [58] Carey, V., Zeger, S. L., & Diggle, P. (1993). Modelling multivariate binary data with alternating logistic regressions. *Biometrika*, 80(3), 517–526.
- [59] Tang, W., He, H., & Tu, X. M. (2023). *Applied categorical and count data analysis*. CRC Press.
- [60] Liu, J., Lin, T., Chen, T., Zhang, X., & Tu, X. M. (2022). On semiparametric efficiency of an emerging class of regression models for between-subject attributes. *arXiv preprint arXiv:2205.08036*.
- [61] Weiss, N. A., Holmes, P. T., & Hardy, M. (2006). *A course in probability*. Boston, MA, USA: Pearson Addison Wesley.
- [62] Jenkinson, A. F. (1955). The frequency distribution of the annual maximum (or minimum) values of meteorological elements. *Quarterly Journal of the Royal Meteorological Society*, 81(348), 158–171.
- [63] Wertheim, J. O., Kosakovsky Pond, S. L., Forgione, L. A., Mehta, S. R., Murrell, B., Shah, S., Smith, D. M., Scheffler, K., & Torian, L. V. (2017). Social and genetic networks of HIV-1 transmission in New York City. *PLoS Pathogens*, 13(1), e1006000.
- [64] Snijders, T. A., Pattison, P. E., Robins, G. L., & Handcock, M. S. (2006). New specifications for exponential random graph models. *Sociological Methodology*, 36(1), 99– 153.

- [65] Robins, G., Pattison, P., Kalish, Y., & Lusher, D. (2007). An introduction to exponential random graph (p*) models for social networks. *Social Networks*, 29(2), 173–191.
- [66] Goyal, R., De Gruttola, V., & Onnela, J.-P. (2023). Framework for converting mechanistic network models to probabilistic models. *Journal of Complex Networks*, 11(5), cnad034.
- [67] Goyal, R., Blitzstein, J., & De Gruttola, V. (2014). Sampling networks from their posterior predictive distribution. *Network Science*, *2*(1), 107–131.
- [68] Dubois, E., & Blank, G. (2018). The echo chamber is overstated: The moderating effect of political interest and diverse media. *Information, Communication & Society*, 21(5), 729–745.
- [69] Blom, G. (1976). Some properties of incomplete U-statistics. *Biometrika*, 63(3), 573–580.
- [70] Dudley, R. M. (2018). *Real analysis and probability*. Chapman; Hall/CRC.
- [71] Lin, & Jeon, Y. (2006). Random forests and adaptive nearest neighbors. *Journal of the American Statistical Association*, *101*(474), 578–590.
- [72] Boole, G. (1847). *The mathematical analysis of logic*. CreateSpace Independent Publishing Platform.
- [73] Efron, B., & Stein, C. (1981). The jackknife estimate of variance. *The Annals of Statistics*, 9(3), 586–596.
- [74] Mukhopadhyay, N. (2020). Probability and statistical inference. CRC Press.
- [75] Hoeffding, W. (1992). A class of statistics with asymptotically normal distribution. In *Breakthroughs in statistics: Foundations and basic theory* (pp. 308–334). Springer.
- [76] Billingsley, P. (2017). Probability and measure. John Wiley & Sons.
- [77] Jensen, J. L. W. V. (1906). Sur les fonctions convexes et les inégalités entre les valeurs moyennes. *Acta Mathematica*, *30*(1), 175–193.
- [78] Feng, C., Wang, H., Chen, T., & Tu, X. M. (2014). On exact forms of Taylor's theorem for vector-valued functions. *Biometrika*, *101*(4), 1003.