UNIVERSITY OF CALIFORNIA SAN DIEGO

Novel Machine Learning and Statistical Models for High Dimensional and Observational Study Data: Applications to HIV genetic linkage network, fMRI and Survey Data

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

in

Biostatistics

by

Tuo Lin

Committee in charge:

Professor Xin Tu, Chair
Professor Armin Schwartzman
Professor David Strong
Professor Jingjing Zou

2023

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

University of California San Diego

2023

DEDICATION

To my wife Toni: Thank you for your altruistic love and accompany during my whole Ph.D. life.

To my parents: Thank you for your timely encouragement and support for me to walk through every dark night.

EPIGRAPH

The only moment of possible happiness is the present.
The past gives regrets. And future uncertainties.

Arsène Wenger

Ask and it will be given to you; seek and you will find;
knock and the door will be opended to you.
For everyone who asks receives; the one who seeks finds;
and to the one who knocks, the door will be opened.

Matthew 7:7-8

If I have seen further it is by standing on the shoulders of Giants.

Isaac Newton

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SUPPLEMENTAL FILES

Lin_Supplementary.docx

ACKNOWLEDGEMENTS

This dissertation is not my individual work. Without methodological and spiritual support from my outstanding and open-minded advisors, all skillful faculty members and intelligent peers in the Division of Biostatistics at the University of California, San Diego (UCSD), I can not accomplish the goal of finishing this dissertation. It has been my pleasure to enjoy a wonderful journey of obtaining my doctoral training in this program and working on cutting-edge and thought-provoking research projects. At this moment, I would like to sincerely acknowledge all the remarkable individuals for their unwavering support and belief in me.

First and foremost, I would like to express my most sincere gratitude and appreciation to my advisor and committee chair Professor Xin Tu for his guidance, patience and encouragement. I started to work as a Graduate Student Researcher (GSR) at the UCSD Altman Clinical and Translational Research Institute (ACTRI) under his supervision in my second year of master's in statistics. He led me into the world of biostatistics and taught me how to apply statistical theories to solve biomedical research problems. As a mentor, he also encouraged me millions of times whenever I was feeling down and shared his own stories with me to inspire me to face challenges. It has been my honor to have him as my advisor during my Ph.D. and it is hard to find words to express my gratitude.

I am deeply grateful to Professor Armin Schwartzman for his solid support and genuine advice. He brought me into the brain image research field, which motivate one topic of this dissertation work and more potential work in the future. Thanks to his constant encouragement, I have overcome barriers to involve actively, think critically and ask questions frequently in various research activities. Moreover, he organized many wonderful extracurricular activities, which spiced up the tedious Ph.D. life and helped me build a wide professional network.

I would like to express my deep gratitude to Professor Jingjing Zou for her professional guidance. She generously supported me for more than two years of my GSR work and assisted me in leading a project on COVID-19 surveillance of UCSD's Return to Learn program, which has been included in this dissertation as one major collaborative research work. She has also sparked my interest in functional data analysis (FDA) by working with me on one of our seminal methodological projects and teaching me analytic skills about FDA.

I would like to offer my special thanks to my committee member Professor David Strong for his continued support. I was his student in the *scale development in health behavior* class. He taught me how to write better scientific reports and provided valuable comments on my writing.

I would like to extend my sincere thanks to Professor Victor De Gruttola and Martin Natasha for their unconditional and firm support during my study period and job-hunting procedure. Professor Victor De Gruttola is a renowned biostatistician in network analysis and its application to HIV disease transmission. His wisdom and insightful suggestions have guided me to complete multiple projects on network analysis with him. Professor Martin Natasha is a knowledgeable and well-experienced professor from the Department of Infectious Disease. She guided me to complete the most significant collaborative work in this dissertation.

It has been a rewarding experience working as a GSR at ACTRI for more than three years. I would like to thank Professor Lily Xu and Xinlian Zhang for their professional guidance when I worked at ACTRI.

I would like to express my sincere appreciation to all the faculty members in the Division of Biostatistics at UCSD, especially Professor Loki Natarajan, Florin Vaida and Karen Messer for their broad support and enthusiastic encouragement all along.

I would like to thank Dr. Samuel Davenport, who is a postdoctoral fellow in the division and my friend, for his continued support. I would also like to thank all my peers in the division and all my friends who have encouraged me when I had hard times.

Finally I would like to dedicate this dissertation to my wife Toni and my parents. I could not have accomplished this without their love and support.

Chapter 1, in full, is a reprint of the material as it appears in *Tuo Lin, Tian Chen, Jinyuan Liu, and Xin Tu. (2021). Extending the Mann-Whitney-Wilcoxon Rank Sum Test to Survey Data* *for Comparing Mean Ranks. Statistics in Medicine, 40(7), 1705-1717.* The dissertation author was the primary author of this paper.

Chapter 2, in full, is currently being prepared for submission for publication of the material as it may appear in *Tyler Vu*, Tuo Lin*, Jingjing Zou, Xin Tu and Victor De Gruttola. Doubly Robust Estimation of Network Linkage Probabilities in the Presence of Missing Data*. The dissertation author was the co-primary author of this paper.

Chapter 3, in full, has been submitted for publication of the material as it may appear in *Tuo Lin, Smruthi Karthikeyan, Alysson Satterlund, Robert Schooley, Rob Knight, Victor De Gruttola, Natasha Martin, Jingjing Zou. Optimizing campus-wide COVID-19 test notifications with interpretable wastewater time-series features using machine learning models*. The dissertation author was the primary author of this paper.

Chapter 4, in full, is currently being prepared for submission for publication of the material as it may appear in *Tuo Lin, Tsungchin Wu, Xinlian Zhang, and Xin Tu. Non-parametric Causal Inference for Mann-Whitney-Wilcoxon Rank Sum Test Using Random Forest*. The dissertation author was the primary author of this paper.

Chapter 5, in full, is currently being prepared for submission for publication of the material as it may appear in *Tuo Lin, Armin Schwartzman and Samuel Davenport. Peak p-values for Gaussian random fields on a lattice*. The dissertation author was the primary author of this paper.

VITA

2016	Bachelor of Science, Mathematics (Applied), University of California San Diego Bachelor of Arts, Economics, University of California San Diego
2018	Master of Science, Statistics, University of California San Diego
2015-2017	Teaching Assistant, Department of Mathematics, University of California San Diego
2022-2022	Teaching Assistant, Division of Biostatistics, University of California San Diego
2017-2023	Research Assistant, University of California San Diego
2023	Doctor of Philosophy, Biostatistics, University of California San Diego

PUBLICATIONS

Methodology

- De Gruttola, V., Nakazawa, M., **Lin, T.**, Liu, J., Tu, X., Goyal, R., Little, S. and Mehta, S. (2022). Modeling Homophily in Dynamic Networks with Application to HIV Molecular Surveillance. Under review.

- 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.

- Liu, J., Zhang, X., **Lin, T.**, Chen, R., Zhong, Y., Chen, T., Wu, T., Nguyen, T.T, Lee, E., Jeste, D.V. and Tu, X.M. (2022). A Distance-Based Semiparametric Regression Framework for the Between-Subject Attributes of High-dimensional Data. Under review.

- Zou, J., **Lin, T.**, Di, C., Bellettiere, J., Jankowska, M. M., Hartman, S. J., ... & Natarajan, L. (2022). A Riemann Manifold Model Framework for Longitudinal Changes in Physical Activity Patterns. Annals of Applied Statistics, in press.

- **Lin, T.**, Niu, X., Liu, J., Wu, T., Chen, R., Li, Y., Huang, X., Yang, K., Chen, G., Chen, T., Strong, D.R., Messer, K. and Tu, X.M. (2022). On Outcome and Sampling Weights: An In-depth Look at the Dueling Weights. Under reivew.

- Zhang, J., **Lin, T.**, Yang, K., Wu, T., Chen, R., Chen, T., Suarez-Lopez, J.R. and Tu, X.M. (2022). A Hybrid Parametric and Semi-parametric Regression Model for Informative Missingness in Explanatory Variables Due to Detection Limit. Under review.

- Chen, R., **Lin, T.**, Liu, L., Liu, J., Chen, R, Liu, C., Zou, J., Natarajan, L., Tang, W., and Tu,X.M. (2022). A Double Robust Estimator for Mann Whitney Wilcoxon Rank Sum Test When Applied for Causal Inference in Observational Studies. Under review.

- Vu, T., **Lin, T.**, Novitsky, V., Zou, J., Tu, X.M., De Gruttola, V. (2021). Estimating Viral Genetic Linkage Rates in the Presence of Missing Data. arXiv preprint arXiv:2203.12779.

- **Lin, T.**, Chen, T., Liu, J., & Tu, X. M. (2021). Extending the Mann-Whitney-Wilcoxon Rank Sum Test to Survey Data for Comparing Mean Ranks. Statistics in Medicine, 40(7), 1705-1717. https://doi.org/10.1002/sim.8865.

- Liu, J., Zhang, X., Chen, T., Wu, T., **Lin, T.**, Jiang, L., ... & Tu, X. M. (2021). A semiparametric model for between-subject attributes: Applications to beta-diversity of microbiome data. Biometrics. https://doi.org/10.1111/biom.13487.

Collaboration

- **Lin, T.**, Karthikeyan, S., Satterlund, A., Knight, R., Schooley, R., De Gruttola, V., Martin, N., Zou, J. (2023). Optimizing campus-wide COVID-19 test notification strategy with interpretable wastewater time series features using machine learning models. Under review.

- **Lin, T.**, Zhao, R., Tu, S., Wu, H., Zhang, H., and Tu, X.M. (2023). On modeling relative risks for longitudinal binomial responses: implications from two dueling paradigms. General Psychiatry, 36:e100977. doi: 10.1136/gpsych-2022-100977.

- Bu, Y., Harrington, D.L., Lee, R.R., Shen, Q., Angeles-Quinto, A., Ji, Z., Hansen, H., Hernandez-Lucas, J., Baumgartner, J., Song, T., Nichols, S., Baker, D., Rao, R., Lerman, I., **Lin, T.**, Tu, X.M. and Huang, M. (2023). Magnetoencephalogram-based brain-computer interface for hand-gesture decoding using deep learning. Cerebral Cortex, in press.

- Chronister, B.N.C., Yang, K., Yang, A.R., **Lin, T.**, Tu, X.M., Lopez-Paredes, D., Checkoway, H., Suarez-Torres, J., Gahagan, S., Martinez, D., Ospina, M., Calafat, A.M., Barr, D., Moore, R.C. and Suarez-Lopez, J.R. (2022). Glyphosate, 2,4-D and DEET biomarkers in relation to neurobehavioral performance in Ecuadorian adolescents in the ESPINA cohort. Under review.

- Kim, B. K., Tamaki, N., Imajo, K., Yoneda, M., Sutter, N., Jung, J., **Lin, T.**, Tu, X.M., ... & Loomba, R. (2022). Head to head comparison between MEFIB, MAST, and FAST for detecting stage 2 fibrosis or higher among patients with NAFLD. Journal of Hepatology.

- Grunvald, E., Wei, J., **Lin, T**., Yang, K., Tu, X., Lunde, O., Ross, E., Cheng, J., DeConde, J., Farber, N. (2022). The impact of standardized patients and an interactive lecture on anti-obesity attitudes in third-year medical students: A quasi-experimental study. Under review.

- Dickson, S. D., Thomas, I. C., Bhatia, H. S., Nishimura, M., Mahmud, E., Tu, X. M., **Lin, T.**, Adler, E., Greenberg, B., & Alshawabkeh, L. (2021). Methamphetamine-Associated Heart Failure Hospitalizations Across the United States: Geographic and Social Disparities. Journal of the American Heart Association, 10(16), e018370.

- Odish, M., Yi, C., Tainter, C., Najmaii, S., Ovando, J., Chechel, L., Lipinski, J., Ignatyev, A., Pile, A., Yeong Jang, Y., **Lin, T.**, Tu, X.M., Madani, M., Patel, M., Meier, A., Pollema, T.,& Owens, R. L. (2021). The Implementation and Outcomes of a Nurse-Run Extracorporeal Membrane Oxygenation Program, a Retrospective Single-Center Study. Critical care explorations, 3(6).

- 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).

- Kern, L., Eichberger, L., Wang, H., **Lin, T.**, & Rhee, K. E. (2020). Parental Knowledge and Attitudes About Universal Lipid Screening Among Children Aged 9 to 11 Years. Clinical Pediatrics, 59(4-5), 439-444.

- Richardson, S., **Lin, T.**, Li, Y., Niu, X., Xu, M., Stander, V., & Tu, X. M. (2019). Guidance for use of weights: an analysis of different types of weights and their implications when using SAS PROCs. General Psychiatry, 32(1).

- Proudfoot, J. A., **Lin, T.**, Wang, B., & Tu, X. M. (2018). Tests for paired count outcomes. General psychiatry, 31(1).

- Meier, A., Gross, E. T., Schilling, J. M., Seelige, R., Jung, Y., Santosa, E., Searles, S., **Lin, T.**, Tu, X. M., Patel, H. H., & Bui, J. D. (2018). Isoflurane Impacts Murine Melanoma Growth in a Sex Specific, Immune-Dependent Manner: A Brief Report. Anesthesia and analgesia, 126(6), 1910.

- Galant-Swafford, J., **Lin, T.**, Tu, X., Christiansen, S., & Kim, A. (2018). Methicillin-resistant Staphylococcus Aureus and Clostridium Difficile Infections Among Penicillin-Allergic Patients in a University Hospital. Annals of Allergy, Asthma & Immunology, 121(5), S14-S15.

- Zheng, J. Z., Li, Y., **Lin, T.**, Estrada, A., Xiang, L., & Changyong, F. (2017). Sample size calculations for comparing groups with continuous outcomes. Shanghai archives of psychiatry, 29(4), 250.

ABSTRACT OF THE DISSERTATION

Novel Machine Learning and Statistical Models for High Dimensional and Observational Study Data: Applications to HIV genetic linkage network, fMRI and Survey Data

by

Tuo Lin

Doctor of Philosophy in Biostatistics

University of California San Diego, 2023

Professor Xin Tu, Chair

The objective of this dissertation is to develop novel statistical models for modeling different types of high-dimensional data such as large-scale survey data, HIV genetic linkage network data and fMRI data. This dissertation is compromised of five parts. The Mann-Whitney-Wilcoxon rank sum test (MWWRST) is called for when two-sample t-tests fail to provide meaningful results, as they are highly sensitive to outliers. In the first chapter, we develop an approach to extend the MWWRST to survey data to test the null of equal mean rank. Akin to the goal of modeling paired subjects' outcomes, or between-subject outcomes in MWWRST, in the second chapter, we model the probability of HIV genetic linkage by using semiparametric

functional response models (FRM). We apply the proposed method to study the genetic linkage between and within villages in Botswana from the Botswana Combination Prevention Project (BCPP), which is a cluster randomized study to implement interventions to prevent and control HIV transmission in Botswana. Since BCPP is a survey study with nonresponse, we adopt the doubly robust estimator to address the missing data problem.

During the COVID-19 pandemic, at UCSD, daily high-resolution wastewater surveillance at the building level is being used to identify potential undiagnosed infections and trigger notification of residents and responsive testing, but the optimal determinants for notifications are unknown. To fill this gap, we propose a framework for identifying features of a series of wastewater test results that can predict the presence of COVID-19 in residences associated with the test sites by using classification/decision tree models. This collaborative work also motivates us to study the asymptotic properties of an ensemble of multiple classification trees, random forests model, and extend it to model between-subject outcomes in the next chapter.

Finally, my research on high-dimensional data also includes work on functional magnetic resonance imaging (fMRI). To detect peaks and identify the locations of peaks in fMRI data, we develop a Monte Carlo method to compute the height distribution of local maxima of a stationary Gaussian or Gaussian-derived random field that is observed on a regular lattice.

Introduction

The development of this dissertation is motivated by analyzing high-dimensional data and observational data from national surveys, image analysis and network analysis. Technological advances in the internet and digital arena have greatly facilitated data collection and analysis for research and other purposes. In particular, surveys can be conducted to collect data instantaneously to provide important information for timely issues and topics, medical images can be used for incurable diseases such as cancer and Alzheimer's disease diagnosis and prognosis, and network epidemiology can be leveraged to investigate patterns of disease transmission dynamics and the effect of interventions on them. These data bring unprecedented statistical challenges, including the unique and intractable data structure and ultra-high data dimensionality.

Traditional parametric regression models and two-sample t-test require strict data assumptions such as the data need to be independent and identically distributed (iid) and Normally distributed. However one particular problem arises from modern large-scale surveys is increased number of outliers, which violates the Normality assumption, thus yielding uninterpretable and often biased results. Rank-based methods such as Mann-Whitney-Wilcoxon rank sum test (MWWRST) are good alternatives for mean-based methods when the data have outliers. By utilizing the latest development in semiparametric models, we develop a Mann-Whitney form based MWWRST to compare mean ranks between two groups for survey data [61]. Compared with a previous approach from Lumley and Scott [64], which focuses on testing the null of equal distribution, our method has a less restrictive null and works better as an alternative for t-test. We illustrate the proposed approach and show major differences from Lumley and Scott's approach using both NHANES and simulated data.

The iid assumption is critical for traditional statistical methods centered around individual outcomes, which we termed ”within-subject attributes”. This assumption can easily be violated when analyzing ”between-subject attributes”, which is introduced to define functions of paired individual outcomes. For example, outcomes defined by pairwise comparisons between two groups, akin to the MWWRST test statistics, and network connectivity defined for a pair of nodes in a network. To model between-subject attributes, we take advantage of the functional response model (FRM) [62], which is a class of semiparametric models that is free of distribution assumption. A major difficulty of inference about this model is classic asymptotic theories including law of large number (LLN) and central limit theorem (CLT) could not be directly applied since data is not iid. Therefore in this work we illustrate how to utilize U-statistics based weighted generalized estimating equations (UWGEE) to estimate the parameter and justify the consistency and asymptotic normality of the proposed estimator using U-statistics theory [55]. The method has been applied to analyze linkage network data from a large cluster-randomized trial, the Botswana Combination Prevention Project (BCPP), to evaluate the effectiveness of interventions designed to control HIV in Bostwana. Specifically, we develop a doubly robust estimator of network linkage probabilities in the presence of missing data for the BCPP study.

Although the semiparametric methods we have been working on are more robust than the parametric models and can be used to model the non-iid between-subject outcomes, they still require correct specifications of the parametric conditional mean. For longitudinal and observational data, this parametric form is required for the main model of interest, auxiliary components for weights (e.g. logistic regression for IPW methods) and outcome regression (for doubly robust estimators).

A GSR work on leveraging the decision tree method for COVID-19 surveillance immediately piqued my interest in random forests (RF). During the COVID-19 pandemic, wastewater surveillance of the SARS CoV-2 virus has been demonstrated to be effective for population surveillance at the county level down to the building level. At the University of California San Diego (UCSD), daily high-resolution wastewater monitoring at the building level is being used to identify potential undiagnosed infections and trigger notification of residents and responsive testing, but the optimal determinants for notifications are unknown. To fill this gap we propose a framework for identifying features of a series of wastewater test results that can predict the presence of COVID-19 in residences associated with the test sites. Using time series of wastewater results and individual testing results during periods of routine asymptomatic testing among UCSD students from 11/2020-11/2021, we develop hierarchical classification/decision tree models to select the most informative wastewater features (patterns of results) which predict individual infections. We find that the best predictor of positive individual level tests in residence buildings is whether or not the wastewater samples were positive in at least 3 of the past 7 days. We also demonstrate that the tree models outperform the random forest models in modeling the data from our setting. Results of this study have been used to refine campus-wide guidelines and email notification systems to alert residents of potential infections.

After this project, I started developing RF-based nonparametric models. Although RF has been widely applied, asymptotic properties of estimated nonparametric regression relationships have not been carefully studied, until the seminal paper of Wager and Athey (2018) [96]. This serves as a perfect theoretical basis for the extension of RF to between-subject cases because both asymptotic proofs rely heavily on U-statistics theory. By integrating their U-statistics based approach into ours, in this work I extend their RF-based estimators for causal inference for MWWRST. The proposed method can have a wide range of applications in inference about personalized treatment effects by adjusting high dimensional covariates.

Recent development in statistical methods for high-dimensional imaging data facilitates the research on functional magnetic resonance imaging (fMRI), which can measure brain activity by detecting changes associated with blood flow to learn brain structures and functions, guide treatment of brain therapy and understand different types of brain-related disease such as Schizophrenia and Alzheimer's. Many previous studies used random field theory (RFT) for inference in fMRI analysis [21, 80]. Recently, Eklund et al. (2016) investigated the validity of RFT based cluster size and voxelwise inference and found that a number of the traditional assumptions did not hold in practice, including the assumption to view data as a continuous random field [37]. We develop an approach for performing peakwise inference to provide inference without viewing fMRI as a continuous random field. This simulation-based Monte Carlo discrete local maxima (MCDLM) approach works for any stationary Gaussian random field under arbitrary connectivity (i.e., where local maxima are defined with respect to any desired neighborhood). This offers a solution for small values of FWHM when the existing formula derived for continuous domain is not accurate. Additionally, we extended the approaches in Worsley (2005) and Taylor et al.(2007) to compute the height distribution and compared it with our approach as well as the RFT approach in Schwartzman and Telschow (2019) by simulations [101, 87, 80]. The simulation results supported MCDLM as a well-suited approach for inference on peak heights when smoothness levels are low.

Chapter 1

Extending the MWW Rank Sum Test to Survey Data for Comparing Mean Ranks

1.1 Introduction

Survey studies are widely used to provide timely information on important topics of interest in a population of interest, such as demographic distribution, voter opinions and disease prevalence. Analysis of survey study data requires some special attention because of complex sampling designs employed to obtain reliable and efficient population-level inference. Survey methodology allows us to intentionally alter distributions of different subgroups in the survey sample by over- and/or under-sampling some subpopulations so that under-represented groups are well represented to ensure reliable population-level estimates without resorting to extremely large samples. Conventional statistical methods generally yield biased estimates due to such "selection bias".

After Horvitz and Thompson' seminal work [49], many popular statistical methods have been extended to survey data and popular statistical packages such as R, SAS, SPSS and STATA all provide support for such methods. For example, many popular SAS procedures have their survey counterparts to facilitate analysis of survey data, such as PROC SURVEYREG for linear and PROC SURVEYPHREG for Cox regression analysis [85].

More recently, Lumley and Scott (2013) developed an approach to extend the Mann-Whitney-Wilcoxon (MWW) rank sum test to survey data with an accompanying R package for facilitating research [64]. Their approach has filled an important gap in survey methodology, as it tackles a rank-based statistic for which application of inverse probability weighting (IPW) of Horvitz and Thompson is not straightforward. However, their approach tests the null of equal distribution, which, although efficient for comparing two distributions, is limited in practice. For example, consider two normal distributions with a common mean but different variances. Although the two groups are considered no different in most studies, the null tested by Lumley and Scott's approach does not hold true in this example. Moreover, the MWW test is generally called for when two-sample t-tests (with or without equal variance assumed) are problematic to apply, as such mean-based tests are highly sensitive to outliers. In such situations, the MWW test meaningfully compares mean ranks between two groups.

Unlike mean and median, mean rank for a group is calculated based on ranking observations from all groups (two in the case of MWW test) and thus is less intuitive as a measure of the center of one single distribution. Many interpret comparing mean ranks as comparing medians of distributions and apply the MWW for this purpose [71, 35] . Although the two measures are identical for some special distributions (see Section 1.2 for details), they are generally different. Despite such differences, mean rank as a center of a distribution is widely used in practice, such as the MWW test within the current context, Kruskal-Wallis test and rank regression [57, 20].

In this paper, we consider an extension of the MWW to survey data for testing the null of equal mean rank between two groups. After a brief overview of survey sampling and sampling weights in survey studies in Section 1.2, we discuss this null of interest and its relationships to the null of equal distribution and the null of equal median. We then develop an extension of the MWW to testing this null by integrating sampling weights within the context of the MWW test. We examine performance of the proposed approach and illustrate its differences with Lumley and Scott's test of equal distribution using both simulated and real data in Section 1.3. We give our concluding remarks in Section 1.4.

1.2 Mann-Whitney-Wilcoxon Rank Sum Test for Survey Study

1.2.1 Survey Sampling and Sampling Weights

Consider a population Ω_N of finite size N and let y_i denote a continuous outcome of interest. For inference about quantities of interest such as the mean of y_i, we may randomly sample Ω_N to obtain a set S of size n to construct estimates such as the sample mean $\overline{y}_\cdot = \frac{1}{n}\sum_{i \in S} y_i$ of y_i. To facilitate investigation of properties of estimates, S and Ω_N can both be viewed as subsets of a superpopulation $\widetilde{\Omega}$ of infinite size so that their defined sample means are all consistent estimates of $\mu = E\left(y_i\right)$, the mean of y_i defined by the superpopulation $\widetilde{\Omega}$. Under this framework, we can study and compare different estimates defined by the sample S for their asymptotic properties with respect to $\widetilde{\Omega}$ such as consistency of the sample mean $\overline{y}_\cdot = \frac{1}{n}\sum_{i \in S} y_i$ [90, 86]. If sampling fraction $\frac{n}{N}$ is not sufficiently small, non-independence across the y_i's in S may have a non-negligible effect on the asymptotic variance of an estimate and a correction factor $\frac{N-n}{N}$ may be applied [86]. In what follows, we refer to Ω_N as the study population and assume that $\frac{n}{N}$ is sufficiently small so that no correction is necessary for the asymptotic variance.

When the study population is heterogeneous, especially with large variation in subpopulations, simple random sampling by taking a random sample of the population may yield unreliable or even biased estimates due to insufficient representation of under-represented subpopulations unless with an extremely large sample. Survey studies employ complex sampling methods for more efficient inference. For example, under stratified sampling, the population is partitioned into a set of subpopulations, or strata, and simple random sampling is applied within each stratum to obtain a sample of the population. This hierarchical sampling approach ensures sufficient representation of all subpopulations of interest without resorting to an extremely large sample. As the stratified sample no longer represents the study population, sampling weights must be used to construct population-level estimates.

For example, for inference about the mean μ, if the population is partitioned into H (≥ 2)

strata of size N_h and a sample of size n_h $(\leq N_h)$ is taken from each stratum, sampling weights w_{hi} and an unbiased estimate $\widehat{\mu}$ of μ are constructed as follows:

$$\widehat{\mu} = \frac{1}{w_{\cdot\cdot}} \sum_{h=1}^{H} \sum_{i=1}^{n_h} w_{hi} y_{hi}, \quad f_h = \frac{n_h}{N_h}, \quad w_{hi} = \frac{1}{f_h}, \tag{1.1}$$
$$w_{\cdot\cdot} = \sum_{h=1}^{H} \sum_{i=1}^{n_h} w_{hi} = \sum_{h=1}^{H} N_h = N, \quad n = \sum_{h=1}^{H} n_h, \; 1 \leq h \leq H,$$

where f_h is sampling fraction within each stratum [27]. Although construction of sampling weights depends on specific sampling schemes, estimates of μ are constructed in the same form as in (1.1), regardless of sampling methods used. For example, many large scale survey studies employ multi-stage sampling procedures such as counties, households and individual subjects [86], estimates of μ still have the same form (1.1).

In practice, sampling weights may also account for other issues in survey studies such as non-responses [86, 9]. In the following development, we assume sampling weights are given and focus on inference about quantities of interest with such weights.

1.2.2 Mann-Whitney-Wilcoxon Rank Sum Test

Consider two independent samples and let y_{ki} denote outcome of interest from the ith subject within group k $(1 \leq i \leq n_k, k = 1, 2)$. The Mann-Whitney (U_n) and Wilcoxon (W_n) of the MWW rank sum test statistics are given by [66, 99, 55]:

$$U_n = \frac{1}{n_1} \frac{1}{n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} I\left(y_{1i} \leq y_{2j}\right); \quad W_{nk} = \sum_{i=1}^{n_k} R_{ki}, \tag{1.2}$$

where $I(A)$ is an indicator with $I(A) = 1$ if A is true and 0 otherwise and R_{ki} denotes rank scores based on pooled y_{ki}. The above test statistics are for continuous data. If y_{ki} is discrete, the Mann-Whitney statistic becomes $U_n = \frac{1}{n_1} \frac{1}{n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \left[I\left(y_{1i} < y_{2j}\right) + 1/2 I\left(y_{1i} = y_{2j}\right)\right]$ to account for ties. For simplicity, we focus on the continuous case using the test statistics in (1.2) unless stated otherwise. The same considerations also apply to discrete data [86].

The two forms of the MWW test in (1.2) are equivalent, since

$$W_{nk} = n_1 n_2 U_n + \frac{n_k(n_k+1)}{2}. \tag{1.3}$$

Both statistics can be used to form hypotheses. For example, to test for equal mean rank between two groups, we may test the following null:

$$H_0 : E\left(R_{1i}\right) = E\left(R_{2j}\right) \qquad \text{vs.} \qquad H_a : E\left(R_{1i}\right) \neq E\left(R_{2j}\right). \tag{1.4}$$

which is equivalent to (see Appendix):

$$H_0 : \Delta = \frac{1}{2} \qquad \text{vs.} \qquad H_a : \Delta \neq \frac{1}{2}, \tag{1.5}$$

where $\Delta = E\left[I\left(y_{1i} \leq y_{2j}\right)\right]$ is the mean of stochastic ordering. The U-statistic, U_n, in (1.2) is an unbiased estimator of Δ, and its asymptotic normal distribution is widely used to test hypotheses involving Δ for non-survey study data [55]. Note that since $W_{n1} + W_{n2} = n\left(n+1\right)/2$, only one of the W_{nk}'s can be used as a test statistic in practice ($n = n_1 + n_2$).

The MWW test is also used to test the null of equal distribution:

$$H_0 : F_1\left(y\right) = F_2\left(y\right) \qquad \text{vs.} \qquad H_a : F_1\left(y\right) \neq F_2\left(y\right), \tag{1.6}$$

where $F_k\left(y\right) = \Pr\left(y_{ki} \leq y\right)$ denotes the cumulative distribution function (CDF) of y_{ki} $(k = 1, 2)$. For example, a common H_a in this case is $H_a : F_1\left(y\right) = F_2\left(y + c\right)$ for some constant c, known as location shift. The null (1.6) clearly implies the null (1.4), or (1.5). However, the reverse is generally not true, such as comparing two normals with identical mean but different variances as noted in Section 1.1.

When testing the null of equal distribution in (1.6), we calculate the asymptotic distribution of the statistic (U_n or W_{nk}) under this more restricted null, which is generally different from the asymptotic distribution for testing the null of equal mean rank in (1.4), or (1.5) [55]. If interest lies in testing the null in (1.6), the corresponding asymptotic distribution should be used as it generally leads to a more powerful test than the asymptotic distribution under the null in (1.4), or (1.5). If the MWW test is called for to address limitations of two-sample t-tests, interest generally lies in the null in (1.4), or (1.5).

As noted in Section 1.1, many believe that (1.4), or (1.5), also tests equal median between the two groups [71, 35] . Indeed, for symmetric $F_k(y)$'s, equal median does imply $H_0 : \Delta = \frac{1}{2}$, or $H_0 : E(R_{1i}) = E(R_{2j})$, and vice versa, as asserted by the following theorem (see Appendix for a proof).

Theorem 1. *Two symmetrically distributed y_{ki} have the same median if and only if the null in (1.4), or (1.5), holds true.*

However, for non-symmetric $F_k(y)$'s, there is no clear relationship between the two. For example, if y_{1i} is a χ^2 and y_{2i} is a rotated y_{1i} around its median, then y_{ki} have the same median, but not the same mean rank. On the other hand, for the following distributed y_{ki}:

$$y_{1i} \sim U(0,1), \quad y_{2j} \sim \frac{33+3\sqrt{57}}{64}(y_{1i} - \sqrt{57}/3+2), \tag{1.7}$$

the two groups have the same mean rank, but different medians, 0.5 for y_{1i} and $(\sqrt{57}-3)/(3\sqrt{2}) - \sqrt{57}/3+2$ for y_{2j}. The following example shows that it is also possible for non-symmetric y_{ki}'s to have the same median and same mean rank:

$$F_1(y) = \begin{cases} \frac{y}{6}+\frac{1}{2} & y \leq 0 \\ \frac{1}{2}+\frac{y^2}{2} & y > 0 \end{cases}, \quad F_2(y) = \begin{cases} \frac{y}{2}+\frac{1}{2} & y \leq 0 \\ \frac{1}{2}+y-\frac{y^2}{2} & y > 0 \end{cases}.$$

The class of non-symmetric distributions that have the mean rank and medium is actually quite big. For example, for any symmetrically distributed y_{ki}'s with the same mean rank and medium such as $y_{ki} \sim N(\mu, \sigma_k^2)$ $(\sigma_1^2 = \sigma_2^2)$ and any monotone function $g(\cdot)$ such as $g(\cdot) = \exp(\cdot)$, we can create two non-symmetrically distributed $z_{ki} = g(y_{ki})$ that have the same mean rank and medium. This is because ranks and mediums are invariant under monotone transformation. An implication of this invariance property in practice is that if we can transform non-symmetricaly distributed y_{ki} to symmetrically distributed ones or or nearly so, testing the null of same medium is equivalent to that of equal mean rank or approximately so, and vice versa.

1.2.3 Weighted Mann-Whitney-Wilcoxon Rank Sum Test

Lumley and Scott proposed the following statistic for testing the null of equal distribution in (1.6) [64]:

$$T_n = \frac{1}{w_{1i\cdot}} \sum_{i=1}^{n_1} w_{1i} R_{1i} - \frac{1}{w_{2i\cdot}} \sum_{j=1}^{n_2} w_{2i} R_{2i}, \tag{1.8}$$

where $R_{ki} = F_k^{(e)}(y_i)$ and $F_k^{(e)}(y)$ denotes the empirical CDF of y_{ki} for group k $(k = 1, 2)$. The test statistic in (1.8) compares two weighted mean ranks, which may be viewed as extending the Wilcoxon form to account for sampling weights. Under the null in (1.6), the test statistic T_n has an asymptotic normal distribution with mean zero and thus rejects the null if the weighted mean ranks are significantly different.

Our approach is to extend the Mann-Whitney U-statistic U_n in (1.2) to account for sampling weights. To this end, we consider a weighted Mann-Whitney U-statistic of the form:

$$U_n = \frac{1}{w_{1\cdot}} \frac{1}{w_{2\cdot}} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} w_{1i} w_{2i} I\left(y_{1i} \leq y_{2j}\right), \quad w_{k\cdot} = \sum_{i=1}^{n_k} w_{ki}, \quad k = 1, 2. \tag{1.9}$$

Unlike the Lumley and Scott's test statistic in (1.8), the proposed test statistic in (1.9) is a weighted average of indicators of stochastically ordered pairs $I\left(y_{1i} \leq y_{2j}\right)$, rather than ranks, of observed $\left(y_{1i}, y_{2j}\right)$. Although equivalent in the absence of sampling weights, the two forms of the MWW rank sum test in (1.8) and (1.9) are generally different and lead to difficult conclusions when applied to survey data (see also Section 1.3.1).

As summarized in Theorem 2 below, U_n in (1.9) is also a consistent and asymptotically normal estimate of Δ (see Appendix for a proof).

Theorem 2. *Let*

$$\begin{aligned} h\left(y_{1i}, w_{1i}; y_{2j}, w_{2j}\right) &= w_{1i} w_{2j}\left[I\left(y_{1i} \leq y_{2j}\right) - \Delta\right], \\ h_k\left(y_{ki}, w_{ki}\right) &= E\left[h\left(y_{1i}, w_{1i}; y_{2j}, w_{2j}\right) \mid y_{ki}, w_{ki}\right], \\ \sigma_k^2 &= Var\left(h_k\left(y_{ki}, w_{ki}\right)\right), \quad \sigma_U^2 = \left[E\left(w_{1i}\right) E\left(w_{2j}\right)\right]^{-2}\left(\rho_1^2 \sigma_1^2 + \rho_2^2 \sigma_2^2\right), \\ \lim_{n \to \infty} \frac{n}{n_k} &= \rho_k^2 < \infty, \quad \overline{w}_{k\cdot} = \frac{1}{n_k}\sum_{i=1}^{n_k} w_{ki}, \\ \widehat{h}_k\left(y_{ki}, w_{ki}\right) &= \frac{1}{n_k}\sum_{j=1}^{n_{1k}}\left(w_{1i} w_{2j} I\left(y_{1i} \leq y_{2j}\right) - \widehat{\Delta}\right), \\ \widehat{\sigma}_k^2 &= \frac{1}{n_k - 1}\sum_{i=1}^{n_k}\left[\widehat{h}_k\left(y_{ki}, w_{ki}\right)\right]^2, \quad k = 1, 2. \end{aligned} \tag{1.10}$$

Under mild regularity conditions, we have:

a. *U_n is consistent;*

b. *U_n is asymptotically normal:*

$$\sqrt{n}\left(U_n - \Delta\right) \to_d N\left(0, \sigma_U^2\right); \tag{1.11}$$

c. *A consistent estimate of σ_U^2 is given by:*

$$\widehat{\sigma}_U^2 = \left(\overline{w}_{1\cdot} \overline{w}_{2\cdot}\right)^{-2}\left(\frac{n}{n_k}\widehat{\sigma}_1^2 + \frac{n}{n_k}\widehat{\sigma}_2^2\right), \tag{1.12}$$

where $\to_d$ denotes convergence in distribution.

By using results in Theorem 2, we can readily compute p-values and/or confidence intervals when testing the hypothesis in (1.5), or (1.4). If the null is rejected, confidence intervals also provide information about the extent to which Δ deviates from $1/2$, or difference between the mean ranks. Theorem 2 is also applicable if Δ in (1.5) is different from $1/2$ under H_0, although such nulls are less popular in practice.

1.3 Illustration

We illustrate the proposed approach and compare its performance with Lumley and Scott's approach using simulated as well as real study data. In all examples, we set type I error $\alpha = 0.05$ and Monte Carlo (MC) sample size $M = 2,000$.

1.3.1 Simulation Study

Consider stratified random sampling from a study population of size N consisting of two evenly distributed subpopulations, or strata, of size $N_h = N/2$ $(h = 1,2)$. Let y_{ki} denote an outcome of interest from group k $(= 1,2)$. Under stratified random sampling, we randomly sample m_h subjects within stratum h. Let x_{hi} denote the group indicator within the hth stratum ($x_{hi} + 1 = k$ if group k is sampled) and let y_{khi} denote the sampled outcome within stratum h. Let μ_{kh} denote the population mean of group k in the hth stratum and σ^2 the common variance across both strata and groups.

The above setting yields the following group mean, group size and sampling weights for sampled subjects:

$$\mu_k = \frac{N_1}{N}\mu_{k1} + \frac{N_2}{N}\mu_{k2}, \quad n_k = \sum_{h=1}^{2}\sum_{i=1}^{m_h} x_{hi}, \quad w_{khi} = \begin{cases} \frac{N_1}{m_1} & \text{if } h = 1 \\ \frac{N_2}{m_2} & \text{if } h = 2 \end{cases}, \quad k = 1,2. \tag{1.13}$$

For our simulations, we set:

$$N_h = \begin{cases} 20000 & \text{if } h=1 \\ 20000 & \text{if } h=2 \end{cases}, \quad m_1 = m, \quad m_2 = 3m, \quad n = 4m, \quad \sigma^2 = 1, \tag{1.14}$$

$$\text{Scenario } I: \mu_{k1} = \begin{cases} -5 & \text{if } k=1 \\ 5 & \text{if } k=2 \end{cases}, \quad \mu_{k2} = \begin{cases} 5 & \text{if } k=1 \\ -5 & \text{if } k=2 \end{cases}, \quad n = 200, 400,$$

$$\text{Scenario } II: \mu_{k1} = \begin{cases} 5 & \text{if } k=1 \\ 2 & \text{if } k=2 \end{cases}, \quad \mu_{k2} = \begin{cases} 10 & \text{if } k=1 \\ 13 & \text{if } k=2 \end{cases}, \quad n = 200, 400,$$

Following (1.13), the *k*th group has the mean:

$$\text{Scenario } I: \mu_k = \frac{1}{2}(\mu_{k1} + \mu_{k2}) = 0, \quad k = 1, 2.$$
$$\text{Scenario } II: \mu_k = \frac{1}{2}(\mu_{k1} + \mu_{k2}) = \frac{15}{2}, \quad k = 1, 2.$$

We simulate x_{hi} and y_{khi} from the following statistical distributions:

$$x_{hi} \sim Bern(0.5), \quad \log(y_{khi}) \mid x_{hi} + 1 = k \sim N\left(\mu_{kh}, \sigma^2\right), \tag{1.15}$$
$$1 \le i \le m_h, \quad k = 1, 2, \quad h = 1, 2,$$

where $Bern(\mu)$ denotes a Bernoulli with mean μ. Under the above setting, each group is an equal mixture of two lognormal distributions. For Scenario *I*, the two groups have the same lognormal mixture and thus identical distribution. For Scenario *II*, the two groups have different lognormal mixture distributions. By Theorem 1 and monotonicity of $\log(\cdot)$, y_{ki} simulated under simple random sampling for both scenarios will have the same mean rank between the groups and $\Delta = 0.5$, thus satisfying the null in (1.4), or (1.5). However, only for Scenario *I* will the two groups have an identical CDF. With stratum 2 sampled 3 times higher than stratum 1, the simulated y_{khi} under stratified random sampling will have different group means (on $\log(y)$

scale):

$$\text{Scenario } I \quad : \quad \mu_k = \frac{m_1}{n}\mu_{k1} + \frac{m_2}{n}\mu_{k2} = \begin{cases} \frac{5}{2} & \text{if } k = 1 \\ -\frac{5}{2} & \text{if } k = 2 \end{cases},$$

$$\text{Scenario } II \quad : \quad \mu_k = \frac{m_1}{n}\mu_{k1} + \frac{m_2}{n}\mu_{k2} = \begin{cases} \frac{35}{4} & \text{if } k = 1 \\ \frac{41}{4} & \text{if } k = 2 \end{cases}.$$

Thus if used directly without sampling weights, the two groups will have a location shift and thus different mean ranks, in which case the (standard) unweighted MWW rank sum test will yield incorrect inference.

Shown in Table 1.1 are estimates of Δ (averaged over MC iterations), along with asymptotic and empirical standard errors, and empirical type I errors for testing H_0 in (1.5), or (1.4), by the proposed approach. For both scenarios and sample sizes, the estimated Δ were quite close (actually identical after being rounded to two decimal points) to the true $\Delta = 0.5$, asymptotic and empirical standard errors were quite close to each other, and empirical type I errors were very close to the nominal value.

Also shown in Table 1.1 are differences between the two group mean ranks (standardized by the total sample size) for both scenarios and sample sizes; the mean rank is unweighted for the unweighted MWW rank sum test and weighted for the Lumley and Scott's MWW test. For the unweighted test, Table 1.1 also showed the estimated Δ's. The estimated Δ's and type I errors were all highly biased for the unweighted across all scenarios and sample sizes. The differences between the unweighted mean ranks were also quite different from 0 in all cases.

For Lumley and Scott's test, differences between the weighted mean ranks were quite close to 0 for Scenario *I* but different from 0 for Scenario *II* for both sample sizes. Likewise, the empirical type I errors were quite close to the nominal $\alpha = 0.05$ for Scenario *I* but were downwardly biased for Scenario *II* for both sample sizes. The findings for Scenario *I* are expected since they are consistent with testing the null of equal distribution. For Scenario *II*, the

Table 1.1. Results from a simulation study to examine performances of proposed weighted (Proposed), along with Lumley and Scott's weighted (L&S) and unweighted MWW rank-sum test for two sample size n = 200 and 400.

	Difference in	Estimates	Variance		
Methods	(weighted) mean rank	of $\Delta = 0.5$	Asymptotic	Empirical	Type I error
Scenario I: Sample size $n = 200$					
Proposed		0.50	5.77×10^{-4}	5.92×10^{-4}	0.053
L&S	-3.5×10^{-3}				0.047
Unweighted	-0.25	0.75	3.21×10^{-4}	1.27×10^{-4}	1
Scenario I: Sample size $n = 400$					
Proposed		0.50	2.81×10^{-4}	2.88×10^{-4}	0.049
L&S	-2.2×10^{-3}				0.049
Unweighted	-0.25	0.75	1.57×10^{-4}	6.29×10^{-4}	1
Scenario II: Sample size $n = 200$					
Proposed		0.50	1.71×10^{-3}	1.70×10^{-3}	0.052
L&S	-0.12				0.039
Unweighted	-0.24	0.74	1.86×10^{-3}	9.69×10^{-4}	1
Scenario II: Sample size $n = 400$					
Proposed		0.50	8.50×10^{-4}	8.70×10^{-4}	0.052
L&S	-0.12				0.042
Unweighted	-0.24	0.74	9.21×10^{-4}	4.87×10^{-4}	1

large differences between the groups' mean ranks results are a bit surprising, since the two groups simulated have the same mean rank under simple random sampling. Thus, unlike its application to non-survey data, the weighted Wilcoxon W_n statistic in Lumley and Scott's extension no longer has a zero mean unless the null of equal distribution holds true. The proposed extension by weighting the Mann-Whitney U_n continues to yield consistent estimates of Δ and provide correct inference for testing the null of equal mean rank.

We also compared power between the two weighted methods by changing the group means within each stratum in (1.14) to:

$$\text{Scenario } I : \mu_{k1} = \begin{cases} -5.5 & \text{if } k = 1 \\ 5.5 & \text{if } k = 2 \end{cases}, \quad \mu_{k2} = \begin{cases} 5 & \text{if } k = 1 \\ -5 & \text{if } k = 2 \end{cases}, \tag{1.16}$$

$$\text{Scenario } II : \mu_{k1} = \begin{cases} -2 & \text{if } k = 1 \\ 2 & \text{if } k = 2 \end{cases}, \quad \mu_{k2} = \begin{cases} 5 & \text{if } k = 1 \\ -5 & \text{if } k = 2 \end{cases}.$$

Table 1.2. Results from a simulation study to compare power between proposed weighted (Proposed) and Lumley and Scott's weighted (L&S) for n = 200.

Methods	Mean Rank Difference		Estimates of Δ		Power	
	I	*II*	*I*	*II*	*I*	*II*
Proposed	0.066	0.09	0.43	0.59	0.79	0.73
L&S	0.066	0.09			0.78	0.89

For Scenario *I*, we again simulated y_{khi} from lognormals, while for Scenario *II*, we simulated y_{khi} from lognormals for stratum 1, but from normals for stratum 2. Thus, the two groups have more different distributions for the second than the first scenario.

Shown in Table 1.2 are Monte Carlo estimates of Δ by the proposed, differences between the weighted mean ranks from the groups, and power estimates by the proposed and Lumley and Scott's tests for $n = 200$. The two approaches yielded similar power estimates for the first, but Lumley and Scott's method provided more power for the second scenario. The large increase in power compared with a small increase in the mean rank difference from the first to the second scenario seems to indicate that Lumley-Scott test is also sensitive to differences in shape parameters between the two groups. There was not much change in power estimates for the proposed approach, actually a slight decrease, due to increased variability of sample mean rank difference.

Note that strictly speaking, y_{khi} should be simulated from a finite population consisting of N subjects. As indicated in Section 1.2.1, given the large difference between n and N, simulating from such finite populations will yield similar results and draw same conclusions. However, simulating y_{khi} from mathematical distributions makes it easier to control simulation parameters and evaluate performance of different estimates.

1.3.2 Real Study

We applied the proposed approach to the NHANES 2013-2014 Questionnaire Data from the National Health and Nutrition Examination Survey (NHANES), a large national health and

Table 1.3. Results from NHANES 2013-2014 to compare serum copper concentration between anemia and non-anemia groups with proposed (Propsed) and Lumley and Scott (L&S) methods, along with two-sample t-test (for unequal variances).

Group	Mean	Std. dev.	Mean rank	p-value		
				t-test	Proposed	L&S
	Original sample					
Anemia	128.0	38.8	0.56	0.05	0.325	0.08
Non-anemia	117.7	29.5	0.49			
	Original sample with simulated data added					
Anemia	113.3	31.9	0.43	0.09	0.11	0.02
Non-anemia	117.7	29.5	0.49			

nutrition examination survey [42], to show difference between the null in (1.5), or (1.4), and the null in (1.6). Lumley and Scott (2013) used NHANES II (1976-1980)[64] to illustrate their approach by comparing their results with those from the unweighted MWW rank-sum test reported by Knovich et al. (2008)[54] in studying serum copper concentrations in people with and without anemia. We used the NHANES 2013-2014 Questionnaire Data, since, unlike the NHANES II (1976-1980), this data set is available for free download from the Center for Disease Control and Prevention website [42].

Out of 10,175 who completed the interview, 9,813 were examined. After excluding all missing data, we obtained 2,266 subjects, 71 anemia and 2,195 non-anemia, with serum copper amount recorded. Shown in Table 1.3 under “Original sample” are weighted sample means, standard deviations and mean ranks of serum copper concentrations for the two groups. The anemia group had a higher (sample) mean, standard deviation and mean rank than the non-anemia. The weighted two-sample t-test (for unequal variance) was significant, while the proposed and Lumley and Scott’s (L&S) test both showed a non-significant difference. Shown in the left plot of Figure 1.1 is the (weighted) empirical CDFs (eCDFs) for the two groups, which seems to indicate a small location shift. The p-value for the proposed is much larger than that for the Lumley and Scott’s, reflecting a small difference (0.077) between the mean ranks.

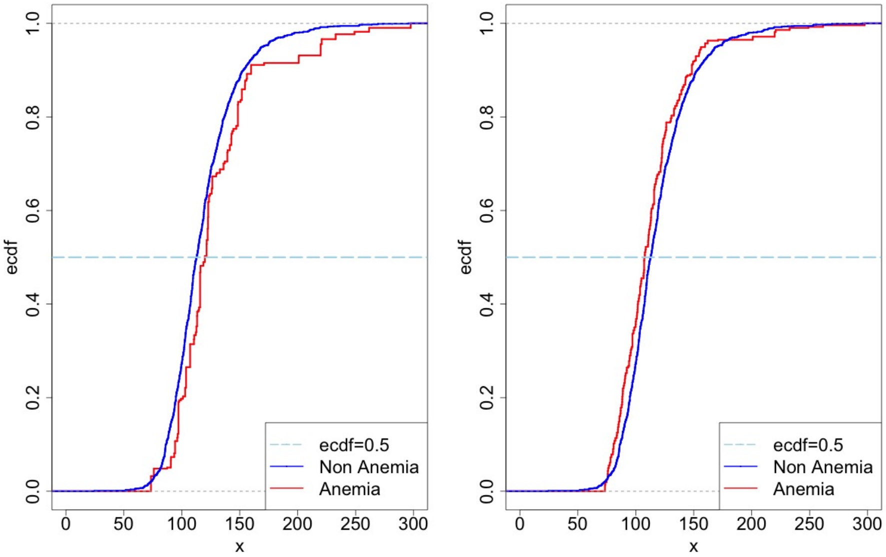


Figure 1.1. Weighted empirical cumulative distribution functions of copper concentration for anemia and non-anemia group for (1) NHANES 2013-2014 Data (left) and (2) NHANES 2013-2014 Data plus 100 simulated values (right).

To further illustrate differences between the two methods, 100 values simulated from N(90, 30) and truncated by the support of the study data were added to the Anemia group of the original sample, with sampling weights based on the mean weight (38457). Shown in Table 1.3 under "Original sample with simulated data added" are the same statistics from this altered study sample and in the right plot of Figure 1.1 is the eCDFs. The decreased difference between the sample means was so significant that it rendered a non-significant difference between the group means, despite increased sample size. Despite reduced difference between the mean ranks (0.063) and the two eCDFs, p-values for both the proposed and Lumley and Scott's became smaller due to increased sample size. Although the proposed remained non-significant, Lumley and Scott's test now indicated a significant difference. With the added observations, the two center measures, mean and mean rank, were no longer significantly different, but their CDFs remained significantly different. Thus, the proposed test is more consistent with testing nulls concerning differences between center measures.

1.4 Discussion

In this paper, we proposed an approach to extend the Mann-Whitney-Wilcoxon (MWW) rank sum test to survey data with sampling weights. Unlike Lumley and Scott's approach, which extends to survey data using the Wilcoxon form, the proposed extension integrates sampling weights into the Mann-Whitney form of the MWW test. Although equivalent in the absence of sampling weights, the two test statistics are generally quite different and should be used for testing their intended null hypotheses. If interest lies in testing the null of identical CDF, Lumley and Scott's approach should be used, as it is generally more efficient. If the MWW is used to address limitations of two-sample t-tests when comparing two groups such as in the presence of outliers, the proposed approach should be used, as it tests a null that is more relevant to what t-tests set out to do. As shown by the simulated and real survey study data, the two extensions of the MWW test generally yield different p-values, rendering them to serve for different purposes.

We also clarified relationships between two center measures, mean rank and medium. Although equal to each other for symmetric distributions, the two center measures have no clearly relationship for non-symmetric distributions. The original and proposed extension of the MWW test can only test the null of equal mean rank, not equal medium. However, if two non-symmetric distributions can be transformed to symmetric ones, then testing the null of medium is the same as testing the null of equal mean rank, thanks to rank invariance under transformation. Thus, in practice, if two groups have approximately symmetric distributions after transformation, they will have similar mediums if they have same mean ranks and vice versa.

In addition to sampling weights, missing data due to non-response or other reasons is another common problem. For example, in our analysis of the NHANES 2013-2014 Questionnaire Data, we obtained only 2,266 subjects out of 9,813 who were examined. Missing data may well be informative and results based on the subsample with complete data may not be generalized to the study population sampled in this survey. Future work is needed to address such missing data.

Chapter 1, in full, is a reprint of the material as it appears in *Tuo Lin, Tian Chen, Jinyuan Liu, and Xin Tu. (2021). Extending the Mann-Whitney-Wilcoxon Rank Sum Test to Survey Data for Comparing Mean Ranks. Statistics in Medicine, 40(7), 1705-1717.* The dissertation author was the primary author of this paper.

Chapter 2

Doubly Robust Estimation of Network Linkage Probabilities in the Presence of Missing Data.

2.1 Introduction

Molecular epidemiology is increasingly used to investigate patterns of HIV transmission, epidemic dynamics; in addition, both the CDC and NIH have proposed that such analyses be used to guide resources intended to end the AIDS epidemic [39]. An important feature of such analyses is investigation of HIV genetic linkage; such linkage can be based on the genetic distance between genetic sequences taken from pairs of individuals from whom HIV transmission may have occurred. Such analyses can reveal which viral strains are propagating within and between communities, the characteristics of people infected with such strains, and the effects of interventions designed to control HIV on the rates at which viral genetic clusters grow. However, in the presence of potentially informatively missing data, observed viral genetic linkage networks do not represent the true underlying networks in populations under study, rendering inferences based on sampled networks unreliable [31]. Specifically, estimates of probabilities of linkage, defined as two individuals selected at random from their respective groups being linked, that ignore the impact of missing data (henceforth referred to as unadjusted estimators) will be biased.

Carnegie et.al. provided consistent estimates of probabilities of linkage under the assumption that viral genetic sequences were missing at random (MAR) given group membership [17] . However, they did not demonstrate asymptotic normality for this estimator. It follows in our previous work, under the assumption that viral sequences were missing completely at random, we developed an unbiased estimator through a subsampling approach and demonstrated consistency and asymptotic normality using a U-Statistics framework [95]. However, in our previous work, demonstrating consistency required strict conditions. Specifically the network generating process of the complete network had to be known and the degree distribution for the complete network would be approximately the same as that of the sampled network (which seemed to be feasible when the sampling proportion was at least 0.40). In this paper, we propose a more flexible approach that allows data to be MAR given auxiliary variables that associated with missing and yields a consistent and asymptotically normal estimator without the strict conditions required in our previous work.

We consider linkage to occur between two individuals if the pairwise genetic distance between their viral genetic sequences is less than some threshold. Obtaining asymptotic properties for estimators of probabilities of linkage, which are informative regarding linkage rates, is challenging, because indicators of linkage across pairs of individuals are between-, rather than, within-subject attributes in conventional statistical analyses. Thus, standard asymptotic methods such as the central limit theorem and law of large numbers cannot be directly applied to these estimators [61, 62]. In this paper, we develop estimators for probabilities of linkage under the assumption that unobserved viral genetic sequences are missing at random (MAR) and derive asymptotic properties for these estimators.

The choice of the threshold indicating linkage is an important scientific question in the analysis of viral genetic data. In general it may be best to investigate the sensitivity of findings, but the methods developed here apply regardless of the threshold value.

We apply the proposed methods to analyze HIV sequences from the Botswana Combination Prevention Project (BCPP), which has motivated the development of the proposed approach, but we note that our methods apply in any setting wherein nodes are sampled from networks. We demonstrate that the methods can be applied to networks more generally.

2.2 Botswana Combination Prevention Project (BCPP)

The BCPP was a large cluster-randomized trial of a combination HIV prevention intervention compared to standard of care in 30 villages in Botswana. In this section we review the sampling design of the BCPP along with the layers of missingness in this study.

2.2.1 Study Introduction

At baseline, 20% of the households in each community in Botswana were targeted for participation in a baseline household survey, which collected demographic and household data among those household members willing to participate. For those unwilling to participate, such demographic and house data were generally provided by heads of households. All participants were tested for HIV infection and virus from blood samples were sequenced for all HIV+ participants; the remaining participants who were HIV- form the incidence cohort. For the next two years the HIV incidence cohort was annually tested for HIV; once again, all virus from those participants who became HIV+ was sequenced. At the end of the BCPP, six communities were selected to participate in a survey of all households, denoted the End of Survey Study (ESS). Because of the inclusiveness of this survey, we illustrate our methods using data from ESS villages.

As our research question focuses on viral linkage without regard to timing of infection— in other words on a static VGL network– we do not consider time as variable in our models. Dynamic VGL models have been described but require information about time of infection, which is generally not available in our study population.

2.2.2 Missing Data

In BCPP, we have three layers of missingness in the observed viral genetic linkage (VGL) network data. First, HIV status is unknown for non-participating household members. Note that unlike common survey studies, demographic data for non-participating household members are also observed (obtained through head of household), provided that the head of the household participated in the BCPP. Second, genetic sequences are unavailable for those who were not tested; hence they are available for only a subsample of those who tested positive. Third, we do not have any observed data on households that did not participate in the ESS. In this paper we will only be addressing the first two layers of missingness and hence. In other words, We assume that our population of interest to consist of only individuals from ESS-participating households.

A common approach for addressing non-response in survey studies is to model this missingness probability, or propensity score, using all observed participants' information such as demographic and HIV status in the current study and then use the inverse of the propensity as propensity score weights, in addition to weights due to multi-stage sampling frames if applicable, to construct consistent population-level estimators under the missing at random (MAR) mechanism [86]. Because the second layer of missingness causes all genetic links to be missing for those who were HIV+ but never tested, this usual approach cannot be applied to address non-response for BCPP. We propose instead to prior estimates of HIV prevalence in Botswana to address this missing not at random (MNAR) mechanisms in the current BCPP. We consider this analysis in two steps: 1) to address the missing HIV status of non-participants, and 2) to address the missingness of links among HIV+ nonparticipants and between them and others who might have been linked to them.

2.3 Notation and Setting

Consider a population of individuals, Ω_N, of finite size N. As in the literature [27, 90], we regard Ω_N as a sample from a superpopulation Ω_∞, i.e., $\Omega_N \subseteq \Omega_\infty$. Let $\mathbf{y}$ denote a $m \times 1$ random vector denoting the viral sequence of a HIV+ individual. This random vector is defined with respect to some probability space with Ω_∞ as its sample space.

Let $S_m = \{\mathbf{y}_i; 1 \leq i \leq m\}$ denote a random sample of $\mathbf{y}_i$ (since the BCPP is a randomized trial). We consider a pair of $\mathbf{y}_i$ and $\mathbf{y}_j$ to be similar, or linked, if the distance, $D_{ij} = d\left(\mathbf{y}_i, \mathbf{y}_j\right)$, between $\mathbf{y}_i$ and $\mathbf{y}_j$ is less than some given threshold, $\delta\ (> 0)$, where $d(\cdot,\cdot)$ denotes a similarity, or distance, metric from $R^m \times R^m$ to R. The random variable, $D_{ij} = d\left(\mathbf{y}_i, \mathbf{y}_j\right)$, as well as its distribution is well defined. We are interested in the mean linkage:

$$\rho = E\left[I\left(D_{ij} \leq \delta\right)\right], \qquad (2.1)$$

where $E(\cdot)$ is defined with respect to the distribution of D_{ij}. Again, as in the literature, we assume that N is sufficiently large such that inference about ρ for the finite population Ω_N can be based on the superpopulation Ω_∞.

Within the current study context, this sample S_m can be partitioned into four subsamples: (1) HIV+ responders, $S_{n_1}^{r+}$; (2) HIV+ non-responders, $S_{n_2}^{nr+}$; (3) HIV- responders, $S_{n_3}^{r-}$; and (4) HIV-non-responders, $S_{n_4}^{nr-}$, where n_k denotes the sample size of each subsample with $m = \sum_{k=1}^{4} n_k$. Our goal is to make inference about ρ for the HIV+ subjects, $S_n^+ = S_{n_1}^{r+} \cup S_{n_2}^{nr+}$, where $n = n_1 + n_2$. However, we have viral sequences $\mathbf{y}_i$ only for the HIV+ responder subsample $S_{n_1}^{r+}$. In BCPP, non-response may not arise from the missing completely at random (MCAR) mechanism; the probability of non-participation is likely to depend on HIV status and demographic variables (A low participation rate among young males was observed in the BCPP as a whole). Thus inference based on observed $\mathbf{y}_i$ in $S_{n_1}^{r+}$ is likely to be invalid for the HIV+ population.

In Section 2.3.1 and 2.3.2, we assume that the missing-response probability is known for $S_{n_2}^{nr+}$ and derive three different estimators for consistent estimation of ρ. In Section 2.3.3, we discuss how to construct an appropriate $S_{n_2}^{nr+}$ by leveraging some external information and use this constructed $S_{n_2}^{nr+}$ to estimate ρ.

2.3.1 Inverse Probability Weighting Estimator

Let $S_n^+ = S_{n_1}^{r+} \cup S_{n_2}^{nr+} = \{\mathbf{y}_i; 1 \le i \le n\}$, let $r_i = 1$ if $\mathbf{y}_i \in S_{n_1}^{r+}$ and $r_i = 0$ otherwise. Let $r_{ij} = r_i r_j$ and y_{ij} denote the linkage indicator between $\mathbf{y}_i$ and $\mathbf{y}_j$ as:

$$y_{ij} = I\left(D_{ij} \le \delta\right) = I\left(d\left(\mathbf{y}_i, \mathbf{y}_j\right) \le \delta\right), \quad \mathbf{y}_i, \mathbf{y}_j \in S_n, \quad (i,j) \in C_2^n, \tag{2.2}$$

where C_2^n denotes the set of $\binom{n}{2}$ combinations of two distinct elements (i,j) from the integer set $\{1, \ldots, n\}$. We can readily estimate ρ in (2.1) by the sample mean:

$$\widehat{\rho} = \left(\sum_{(i,j)\in C_2^n} r_{ij}\right)^{-1} \sum_{(i,j)\in C_2^n} r_{ij} y_{ij}. \tag{2.3}$$

If non-response is independent of HIV status, i.e., $r_{ij} \perp y_{ij}$, the above estimator $\widehat{\rho}$ is unbiased. However, asymptotic properties about $\widehat{\rho}$ cannot be obtained by conventional statistical methods such as laws of large numbers and central limit theorem, because of correlated summands $r_{ij} y_{ij}$. By leveraging the theory of U-statistics, $\widehat{\rho}$ can be shown to be consistent (see Appendix A).

As in other survey studies, non-response in BCPP is likely dependent on HIV status and other demographic variables; a low participation rate among young males was observed in the BCPP as a whole. To accommodate such selection bias, we assume that non-response is independent of $\mathbf{y}_i$ given a vector of covariates $\mathbf{z}_i$, i.e., $r_{ij} \perp y_{ij} \mid \mathbf{z}_i, \mathbf{z}_j$, and denote the response probability by:

$$\pi_{ij} = E\left(r_{ij} \mid \mathbf{z}_i, \mathbf{z}_j\right).$$

Under this missing at random (MAR) assumption, we can estimate ρ consistently by the inverse probability weighted (IPW) estimator:

$$\widehat{\rho}_n^{IPW} = \binom{n}{2}^{-1} \sum_{(i,j)\in C_2^n} \frac{r_{ij}}{\pi_{ij}} y_{ij}. \tag{2.4}$$

As with IPW estimators based on i.i.d. summands, this $\widehat{\rho}_n^{IPW}$ is readily shown to be consistent, if π_{ij} is known (see Appendix A).

Since

$$\pi_{ij} = E\left(r_{ij} \mid \mathbf{z}_i, \mathbf{z}_j\right) = E\left(r_i \mid \mathbf{z}_i\right) E\left(r_j \mid \mathbf{z}_j\right),$$

we can model π_{ij} above through modeling $E\left(r_i \mid \mathbf{z}_i\right)$. We can use any member of the generalized linear models for binary responses such as the logistic regression to model $\pi\left(\mathbf{z}_i; \gamma\right)$. In this study, we use the logistic regression:

$$E\left(r_i \mid \mathbf{z}_i\right) = \pi\left(\mathbf{z}_i; \gamma\right) = \operatorname{expit}\left(\gamma_0 + \gamma_1^\top \mathbf{z}_i\right), \tag{2.5}$$

where $\gamma = \left(\gamma_0, \gamma_1\right)^\top$ and $\operatorname{expit}(\cdot) = \operatorname{logit}^{-1}(\cdot)$ with $\operatorname{logit}(\cdot)$ denote the logit link. By estimating γ and substituting an estimator $\widehat{\gamma}$ in place of γ, the revised estimator by substituting $\pi\left(\mathbf{z}_i; \widehat{\gamma}\right)$ in place of $\pi\left(\mathbf{z}_i; \gamma\right)$ is also consistent. We discuss asymptotic properties of this revised estimator after introducing two other estimators.

2.3.2 Doubly Robust Estimator

The IPW estimator $\widehat{\rho}_n^{IPW}$ in (2.4) of Section 5.3.1 only uses the observed subsample $S_{n_1}^{r+}$ for HIV+ responders. Alternatively, we may impute the missing $\mathbf{y}_i$ given $\mathbf{z}_i$ for $\mathbf{y}_i \in S_{n_2}^{nr+}$ and use the imputed $\mathbf{y}_i$ to create $I\left(D_{ij} \leq \delta\right)$ for all $\mathbf{y}_i \in S_n^+$. Given the high dimension of $\mathbf{y}_i$, this can lead to quite complex models for the association of $\mathbf{y}_i$ with $\mathbf{z}_i$. Siince we are only interested in the linkage $\rho = E\left(y_{ij}\right)$, we can impute y_{ij} directly given $\mathbf{z}_{ij} = \left(\mathbf{z}_i, \mathbf{z}_j\right)$ and then estimate ρ by averaging the observed and imputed y_{ij}.

We first posit an outcome regression model for y_{ij} given $\mathbf{z}_{ij} = \left(\mathbf{z}_i, \mathbf{z}_j\right)$:

$$E\left(y_{ij} \mid \mathbf{z}_i, \mathbf{z}_j\right) = g_{ij} = g\left(\mathbf{z}_i, \mathbf{z}_j; \beta\right), \quad i \in C_2^n. \tag{2.6}$$

We can use logistic or other GLM for binary responses to model $g\left(\mathbf{z}_i, \mathbf{z}_j; \beta\right)$. To ensure symmetric with respect to (i,j) as y_{ij}, we need $g\left(\mathbf{z}_i,\mathbf{z}_j;\beta\right)=g\left(\mathbf{z}_j,\mathbf{z}_i;\beta\right)$. For example, if using logistic regression, we may model the symmetric $g\left(\mathbf{z}_i,\mathbf{z}_j;\beta\right)$ as:

$$g\left(\mathbf{z}_i,\mathbf{z}_j;\beta\right)=\operatorname{expit}\left(\beta_0+\beta_1^{\top}\left(\mathbf{z}_i+\mathbf{z}_j\right)\right).$$

The linkage ρ can be expressed in terms of (2.6) as:

$$\rho=E\left(y_{ij}\right)=E\left\{E\left(y_{ij}\mid\mathbf{z}_i,\mathbf{z}_j\right)\right\}=E\left(g\left(\mathbf{z}_i,\mathbf{z}_j;\beta\right)\right).$$

Thus, we can define another estimator based on the mean score imputed (MSI) $g\left(\mathbf{z}_i,\mathbf{z}_j;\beta\right)$ as:

$$\widehat{\rho}_n^{MSI}=\binom{n}{2}^{-1}\sum_{\mathbf{i}\in C_2^n}\left\{r_{ij}y_{ij}+\left(1-r_{ij}\right)g_{ij}\right\}. \tag{2.7}$$

Again, as with mean score imputation estimators for i.i.d. summands (ref), this $\widehat{\rho}_n^{MSI}$ is readily shown to be consistent, if $g\left(\mathbf{z}_i,\mathbf{z}_j;\beta\right)$ is correctly specified and β is known (see Appendix A).

By combining $\widehat{\rho}_n^{IPW}$ and $\widehat{\rho}_n^{MSI}$, we can yet construct a third estimator. For $\widehat{\rho}_n^{IPW}$ ($\widehat{\rho}_n^{MSI}$) to be consistent, the response probability model π_{ij} in (2.5) (outcome regression model $g\left(\mathbf{z}_i,\mathbf{z}_j;\beta\right)$ in (2.5)) must be correctly specified. By combining $\widehat{\rho}_n^{IPW}$ and $\widehat{\rho}_n^{MSI}$, we can derive a doubly robust estimator when only one of the two models is correctly specified. Within the current setting, let

$$\widehat{\rho}_n^{DR}=\binom{n}{2}^{-1}\sum_{\mathbf{i}\in C_2^n}\left\{\frac{r_{ij}}{\pi_{ij}}y_{ij}+\left(1-\frac{r_{ij}}{\pi_{ij}}\right)g_{ij}\right\} \tag{2.8}$$

It is readily shown that $\widehat{\rho}_n^{DR}$ is consistent if either $\pi\left(\mathbf{z}_i,\mathbf{z}_j;\gamma\right)$ or $g\left(\mathbf{z}_i,\mathbf{z}_j;\beta\right)$ is correctly specified, and γ and β are known (see Appendix A).

To use any of the three estimators above, we must estimate γ for the IPW, β for the MSI or both (γ,β) for the DR estimator. This requires that we know the covaraites $\mathbf{z}_i$ for all subjects in $S_{n_2}^{nr+}$. Although we have $\mathbf{z}_i$ for then entire sample S_n, we do not know which $\mathbf{z}_i$ belong to $S_{n_2}^{nr+}$. We do not even know the size n_2 of $S_{n_2}^{nr+}$. Thus before discussing inference about (γ,β) and any of the estimators, we first discuss how to construct $S_{n_2}^{nr+}$.

2.3.3 Addressing Missing HIV+ Nonparticipants

Because HIV status is missing for all non-participants, $S_{n_2}^{nr+} \cup S_{n_4}^{nr-}$, the BCPP sample S_n alone does not provide sufficient information to identify $S_{n_2}^{nr+}$. To address this missing not at random (MNAR) mechanism [86], we leverage estimates of HIV prevalence for the study population obtained from a national household survey (BAIS) [65]. We assume that within $S_{n_1}^{r+} \cup S_{n_2}^{nr+}$, the HIV+ non-responders $S_{n_2}^{nr+}$ is missing at random (MAR) given the covariates $\mathbf{z}_i$. Under this assumption, S_n^+ is a random subsample of the HIV+ subpopulation of the study population and thus a random subsample of the sample S_n.

Let p denote the HIV prevalence of the study population. Then, $mp = n$ and $n_2 = mp - n_1$. We can take a random subsample of size n_2 from the complement, $S_n \setminus S_{n_1}^{r+}$, of $S_{n_1}^{r+}$, regardless of their HIV status and sampled or not and use the sampled subjects' covariates $\mathbf{z}_i$ in this subsample as $S_{n_2}^{nr+}$ to estimate (γ, β). We may also take multiple such subsamples, $S_{n_{2j}}^{nr+}$, estimate (γ, β) based on each subsample $\left\{S_{n_1}^{r+} \cup S_{n_{2j}}^{nr+}\right\}$, and combine estimates across all $\left\{S_{n_1}^{r+} \cup S_{n_{2j}}^{nr+}\right\}$ $(1 \leq j \leq J)$.

2.3.4 Joint Inference of Linkage and Model Parameters

One approach to inference about ρ is to estimte γ and β separately, substitute the estimators in place of γ and β, and derive asymptotic normality of the resulting linkage estimator of ρ by accounting for sampling variability of the estimated γ and β. An alternative is to leverage the semiparametric functional response models (FRM) for joint inferences about all parameters $\theta = \left(\rho, \gamma^\top, \beta^\top\right)^\top$ (ref).

To this end, consider the following FRM.

$$
\begin{aligned}
&E\left(\mathbf{f}_{ij} \mid \mathbf{z}_i, \mathbf{z}_j\right) = \mathbf{h}_{ij}\left(\mathbf{z}_i, \mathbf{z}_j; \theta\right), \quad \mathbf{f}_{ij} = \left(f_{ij1}, f_{ij2}, f_{ij3}\right)^{\top}, \quad \mathbf{h}_{ij} = \left(h_{ij1}, h_{ij2}, h_{ij3}\right)^{\top}, \\
&f_{ij1} = r_i + r_j, \quad f_{ij2} = y_{ij}, \quad f_{ij3} = \frac{r_{ij}}{\pi_{ij}} y_{ij} + \left(1 - \frac{r_{ij}}{\pi_{ij}}\right) g_{ij}, \\
&h_{ij1} = \pi\left(\mathbf{z}_i; \gamma\right) + \pi\left(\mathbf{z}_j; \gamma\right), \quad h_{ij2} = g\left(\mathbf{z}_i, \mathbf{z}_j; \beta\right), \quad h_{ij3} = \rho, \\
&\pi\left(\mathbf{z}_i; \gamma\right) = \operatorname{expit}\left(\gamma_0 + \gamma_1^{\top} \mathbf{z}_i\right), \quad g\left(\mathbf{z}_i, \mathbf{z}_j; \beta\right) = \operatorname{expit}\left(\beta_0 + \beta_1^{\top}\left(\mathbf{z}_i + \mathbf{z}_j\right)\right).
\end{aligned}
$$

In the model above, the response $\mathbf{f}_{ij}$ is indexed by a pair of subjects. It is a member of a class of functional response models (FRM). This class of models is useful, because of its ability to model relationships of interest that involve interactions between subjects [38, 106]. Let

$$
\begin{aligned}
&S_{ij} = \mathbf{f}_{ij} - \mathbf{h}_{ij}, \quad D_{ij} = \frac{\partial}{\partial \theta} \mathbf{h}_{ij}(\theta), \\
&\mathbf{V}_{ij} = Var\left(\mathbf{f}_{ij} \mid \mathbf{z}_i, \mathbf{z}_j\right) = \begin{pmatrix} V_{ij1} & 0 & 0 \\ 0 & V_{ij2} & 0 \\ 0 & 0 & V_{ij3} \end{pmatrix}^{\frac{1}{2}} R(\alpha) \begin{pmatrix} V_{ij1} & 0 & 0 \\ 0 & V_{ij2} & 0 \\ 0 & 0 & V_{ij3} \end{pmatrix}^{\frac{1}{2}}, \\
&V_{ij1} = Var\left(f_{ij1} \mid \mathbf{z}_i, \mathbf{z}_j\right), \quad V_{ij2} = Var\left(f_{ij2} \mid \mathbf{z}_i, \mathbf{z}_j\right), \quad V_{ij2} = Var\left(f_{ij3} \mid \mathbf{z}_i, \mathbf{z}_j\right),
\end{aligned}
$$

The variances $V_{\mathbf{i}k}$ are given in Appendix B. For inference, consider a class of U-Statistics based weighted generalized estimating equations (UWGEE):

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

Theorem 3. *Let*

$$\mathbf{v}_{ri} = E\left(\mathbf{U}_{n,ij} \mid \mathbf{y}_{ri}, r_{ri}, \mathbf{z}_{ri}\right), \quad \mathbf{v}_{sj} = E\left(\mathbf{U}_{n,ij} \mid \mathbf{y}_{sj}, r_{sj}, \mathbf{z}_{sj}\right), \quad B = E\left(D_{ij} V_{ij}^{-1} D_{ij}^{\top}\right), \tag{2.9}$$
$$\Sigma_r = Var\left(\mathbf{v}_{ri}\right), \quad \Sigma_s = Var\left(\mathbf{v}_{sj}\right), \quad \rho_r^2 = \lim_{n\to\infty} \frac{n}{n_r} < \infty, \quad \rho_s^2 = \lim_{n\to\infty} \frac{n}{n_s} < \infty,$$
$$n = n_r + n_s, \quad \Sigma_U = \rho_r^2 \Sigma_r + \rho_s^2 \Sigma_s, \quad \Sigma_\theta = B^{-1} \Sigma_U B^{-\top}.$$

Then, under mild regularity conditions, we have

1. *$\widehat{\theta}$ is consistent.*

2. *If $\sqrt{n}(\widehat{\theta} - \theta) \to_d N(\mathbf{0}, \Sigma_\theta)$.*

See proof of Theorem 2 in Appendix B. To estimate Σ_θ, we first estimate B by:

$$\widehat{B} = \frac{1}{n_r} \frac{1}{n_s} \sum_{i=1}^{n_r} \sum_{j=1}^{n_s} \widehat{D}_{ij} \widehat{V}_{ij}^{-1} \widehat{D}_{ij}^{\top},$$

where $\widehat{B}$ denotes B with θ substituted by $\widehat{\theta}$. We then estimate Σ_r and Σ_s by:

$$\widehat{\Sigma}_r = \frac{1}{n_r} \sum_{i=1}^{n_r} \widehat{\mathbf{v}}_{ri} \widehat{\mathbf{v}}_{ri}^{\top}, \quad \widehat{\mathbf{v}}_{ri} = \frac{1}{n_s} \sum_{j=1}^{n_s} \widehat{\mathbf{U}}_{n,ij},$$
$$\widehat{\Sigma}_s = \frac{1}{n_s} \sum_{j=1}^{n_s} \widehat{\mathbf{v}}_{sj} \widehat{\mathbf{v}}_{sj}^{\top}, \quad \widehat{\mathbf{v}}_{sj} = \frac{1}{n_r} \sum_{i=1}^{n_r} \widehat{\mathbf{U}}_{n,ij},$$

where $\widehat{\mathbf{U}}_{n,ij}$ denotes $\mathbf{U}_{n,ij}$ with θ substituted by $\widehat{\theta}$. A consistent estimator of Σ_θ is given by:

$$\widehat{\Sigma}_\theta = \widehat{B}^{-1} \widehat{\Sigma}_U \widehat{B}^{-\top} = \widehat{B}^{-1} \left(\frac{n}{n_r} \widehat{\Sigma}_r + \frac{n}{n_s} \widehat{\Sigma}_s \right) \widehat{B}^{-\top}.$$

As in the case of GEE, a working correlation structure $R(\alpha)$ for $\mathbf{f}_{ij}$ parameterized by some vector α may be assumed and incorporated into V_{ij} to improve efficiency of estimates of θ. In this more general case, $V_{ij}(\alpha)$ depends on α as well. Like GEE, the UWGEE estimator $\widehat{\theta}$ by solving the equations above is asymptotically normal by Theorem 1. For notational brevity, we only consider working independence structure below unless otherwise stated.

2.4 Application

2.4.1 Simulation Study

We apply both IPW estimator proposed in 2.3.1 and DR estimator in 2.3.2 to simulated data to examine and compare their performances. For simplicity, in this simulation study we will not generate the individual level outcome but linkage outcome directly. Consider two villages: village r and s. Let $N_r = 200$ and $N_s = 200$ so we can access the performance of the FRM for relative small sample sizes.

Let $\Pr\left(D_{ij}^{rs} \leq \delta\right)$ be the linkage rate for two samples from two villages r and s, respectively ($1 \leq i \leq n_r$, $1 \leq j \leq n_s$). We generate data from the following setup for network data:

$$\Pr(D_{ij}^{rs} \leq \delta; \beta) = \frac{\exp\left(\beta_0 + \beta_1 z_{ri} + \beta_2 z_{sj}\right)}{1 + \exp\left(\beta_0 + \beta_1 z_{ri} + \beta_2 z_{sj}\right)} = p_{\mathbf{i}}, \quad \mathbf{i} = (i, j)$$

$$f_{\mathbf{i}} \mid D_{ij}^{rs} \sim \text{Bernoulli}(p_{\mathbf{i}}), \quad z_{ri} \sim N(\mu_{ri}, \sigma_{ri}^2), \quad z_{sj} \sim N(\mu_{sj}, \sigma_{sj}^2).$$

We let the parameters $\beta_0 = \beta_1 = \beta_2 = 1$, $\mu_{ri} = 0.8$, $\mu_{sj} = 0.2$, $\sigma_{ri} = \sigma_{sj} = 1$ such that the true network linkage rate between two subjects from two villages is 0.816. Next we generate data for missing mechanism:

$$\pi_{ri} = E(r_i \mid z_{ri}) = \frac{\exp(\gamma_0 + \gamma_1 z_{ri})}{1 + \exp(\gamma_0 + \gamma_1 z_{ri})}, \quad r_i \mid z_{ri} \sim \text{Bernoulli}(\pi_{ri}),$$
$$\pi_{sj} = E(r_j \mid z_{sj}) = \frac{\exp(\gamma_0' + \gamma_1' z_{sj})}{1 + \exp(\gamma_0' + \gamma_1' z_{sj})}, \quad r_j \mid z_{sj} \sim \text{Bernoulli}(\pi_{sj}),$$

where $\eta_0 = 1$, $\eta_1 = -1$. If $r_i = r_j = 1$, then $f_{\mathbf{i}}$ is observed, otherwise missing. For the simulation study, we employ Monte Carlo (MC) simulations and set the MC sample size to $M = 1000$.

Table 2.1. Estimates and standard error of Doubly Robust estimator and IPW estimator for simulated network data.

Doubly Robust Estimates			
	True value	Estimates	Standard error
β_0	1	1.001	0.016
β_1	1	1.000	0.017
β_2	1	1.000	0.017
η_0	1	0.996	0.146
η_1	-1	-0.997	0.144
γ	0.816	0.816	0.012
IPW Estimates			
η_0	1	1.004	0.146
η_1	-1	-1.012	0.145
γ	0.816	0.817	0.043

Table 2.1 shows both the DR and IPW estimates and standard errors for simulated network data from above. In our simualtion we correctly specify the propensity score model and the outcome regression model, thus leading to consistent estimates from both estimators. Moreover, by comparing the standard errors of the linkage rate estimates between the two approaches (highlighted with red), we conclude that DR estimator has higher efficiency than IPW estimator, since a smaller standard error is observed. This is no coincidence. The doubly robust estimator for within-subject outcomes from a semiparametric model has been shown to improve the semiparametric efficiency when both models are modeled correctly [93]. More recently, the semiparametric efficiency theory has been extended to study the FRM efficiency to facilitate the research on between-subject outcomes such as genetic linkage network in this paper [63].

2.4.2 BCPP Study

In this section, we perform analysis for the BCPP study introduced in Section 2.2. As we described in Section 2.2, at the end of BCPP study, all households were targeted for a survey in 6 participating villages (Gumare, Maunatala, Mmankgodi, Mmathethe, Ramokgonami, Shakawe), known as ESS villages. We obtain the HIV viral sequences for the participants but

Table 2.2. The number and participation rate of HIV+ individuals in each community that participated in the BCPP.

	Total	Respondent	Non-respondent	Response Rate
Gumare	1125	313	812	27.8%
Maunatala	694	347	347	50.0%
Mmankgodi	1058	265	793	25.0%
Mmathethe	763	314	449	41.2%
Ramokgonami	732	331	401	45.2%
Shakawe	1003	455	548	45.4%

not nonparticipants. The demographic data are provided by the head of household for the whole population. Table 2.2 shows summary statistics of the number and the proportions of HIV+ in individuals that participated in the BCPP. We observe large variability in response rate across different villages; for 4 of the 6 villages, the proportions were over 40% but for the other 2, they were below 30%.

We apply the proposed doubly robust estimator to estimate the mean linkage of the ESS villages. As stated in Section 2.3, two individuals are considered linked if the pairwise genetic distance between their viral genetic sequences is less than some given threshold. Following Novistky et. al. [70], we use a threshold of $c = 0.07$ to define genetic linkage. Figure 2.1 provides a heat map of the intensity of linkage rates after applying the proposed method to address the missing data, within and across the ESS villages. The analysis provides evidence of a larger within- than between-village linkage. In addition, we perform the Wald test to test the difference between the within- and the between-village linkage. We select three pairs of villages (Gumare-Maunatala, Mmankgodi-Mmathethe, Ramokgonami-Shakawe) and the test results (p-value) are shown in Table 2.3. The test results indicate significant difference between the within- and between-village linkage.

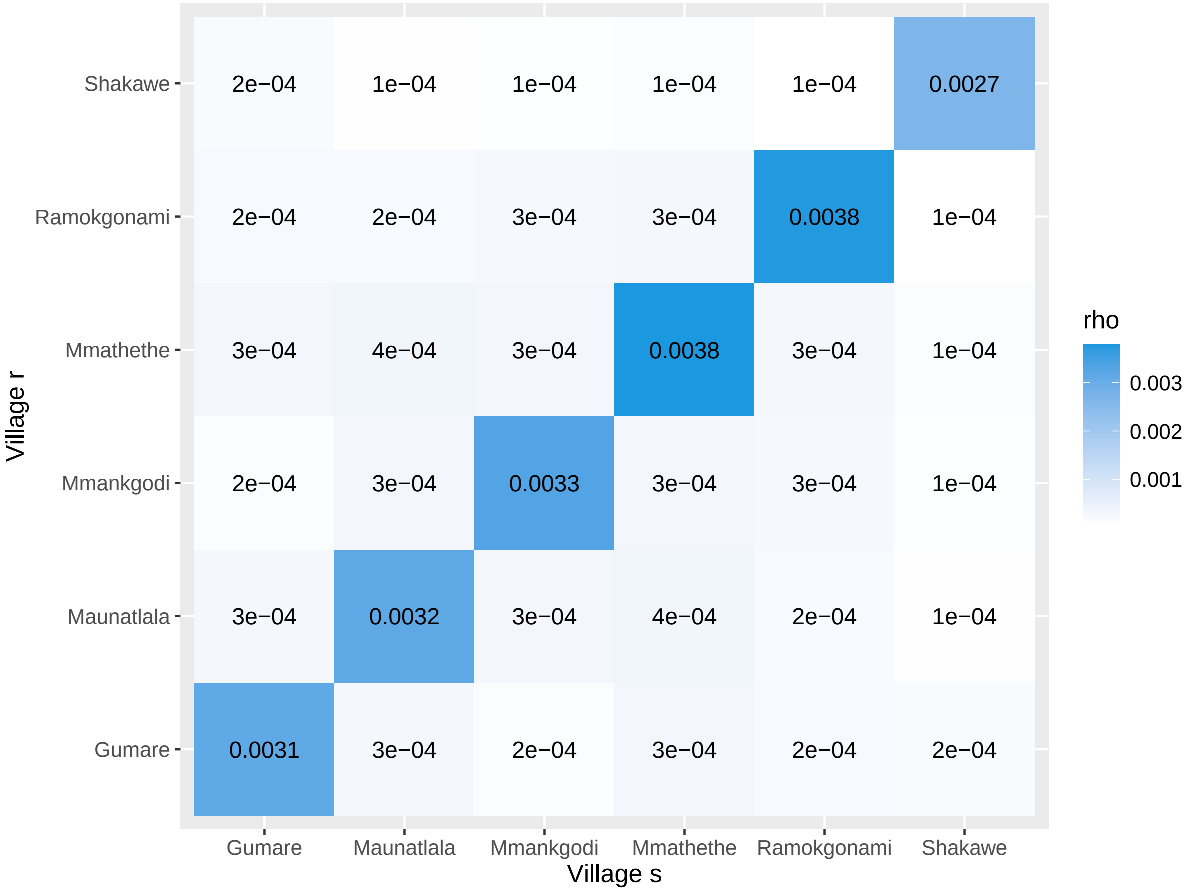


Figure 2.1. Doubly robust estimates for mean linkage between and within the villages from ESS of BCPP. The linkage estimates are indicated by the values and colors of cells, the darker the color, the larger the linkage rate.

Table 2.3. Wald test for difference between the the within- and the between-village linkage

	Gumare-Maunatala vs. Gumare	Mmankgodi-Mmathethe vs. Mmankgodi	Ramokgonami-Shakawe vs. Ramokgonami
p-value	2.46×10^{-22}	6.31×10^{-25}	3.32×10^{-62}

2.5 Discussion

Viral genetic linkage analysis play an important role in molecular epidemiology in it s ability to reveal features of transmission patterns within and across communities such analyses may prove useful in control of COVID-19 and other outbreaks. While methods have been proposed for viral genetic linkage analyses in the presence of sampling bias, this paper is the first to ground such methods in a statistical framework uniquely positioned to address between-subject, rather than within-subject attributes as as the primary focus of analyses. Through the use of FRM and UWGEE, we were able to show consistency and asymptotic normality of our estimators under the assumption that non-responses are MAR, thereby permitting unbiased point and interval estimates, as demonstrated by our simulation results.

Our illustrative example made use of data from an HIV prevention study in Botswana—the BCPP. We demonstrated that VGL linkage across communities is common—which implies that a treatment-as-prevention intervention applied at the village level will likely have effects on HIV incidence that are attenuated compared to effects that would occur if all relationships took place within villages. Furthermore such estimates would also be attenuated compared to another estimand of interest—the counterfactual expected difference in incidence between a setting in which the intervention was implemented in all villages and a setting in which it was in none. Hence these VGL analyses are useful in both design and interpretation of cluster randomized trials for control of endemic diseases or disease outbreaks.

In many real studies, we can estimate missing response probabilities under the MAR assumption. In this case, the FRM with inference based on a class of UWGEE will provide valid inference about linkage among network nodes. In the BCPP study, data are missing on people within households who were enumerated but who did not provide blood samples (used to assess HIV status as well as to obtain sequences) —leading to data that are MNAR. the By utilizing population level estimates and multiple imputation, we addressed this statistical challenge. The idea is similar to raking–an approach used in survey research to improve estimation of sampling weights by utilizing aggregated population-level estimates. Our methods would apply networks of all types, for which sampling of nodes is not complete but for which there exist sufficient covariate information to help identify the MAR mechanism. This paper also illustrates how to address a type of MNAR mechanism in survey research by taking advantage of general information regarding the population survey.

Chapter 2, in full, is currently being prepared for submission for publication of the material as it may appear in *Tyler Vu*, Tuo Lin*, Jingjing Zou, Xin Tu and Victor De Gruttola. Doubly Robust Estimation of Network Linkage Probabilities in the Presence of Missing Data*. The dissertation author was the co-primary author of this paper.

Chapter 3

Optimizing Campus-wide COVID-19 Test Notifications with Interpretable Wastewater Time-series Features using Machine Learning Models

3.1 Introduction

The ongoing spread of SARS CoV-2 creates an urgent need for rapid detection of the SARS CoV-2 virus that aids in development of effective decision making to contain its transmission in communities– particularly those with high density congregate living such as university campuses [91, 8]. Campus-wide monitoring systems capable of rapid detection of new infections remain an important public health priority [59, 98, 72, 67].

Wastewater surveillance has been demonstrated to be a cost-effective approach to monitoring viral spread, by virtue of its ability to 1) detect individual infections at early stages in some settings, 2) identify variants of concern, and 3) provide a less biased assessment of population infection dynamics–particularly in settings where infections are underreported to health departments [53, 52, 51, 58, 74, 4, 68, 104, 32, 47, 46].

As part of the "Return to Learn" (RTL) program of the University of California, San Diego (UCSD), a campus-wide GIS (geographic information systems)-enabled wastewater surveillance system has been implemented for the detection of SARS CoV-2 since Fall 2020

[76, 15]. Currently, the program has 131 samplers collecting daily from $>$340 buildings (both residential and non-residential). A previous study at UCSD from 2020 showed that the wastewater surveillance system was highly sensitive in detecting individual infections (85% of the buildings where a residential student was diagnosed with SARs-COV-2 had a positive wastewater signal prior to individual identification). Additionally, notification of building residents that their building had a positive signal doubled testing rates among residents, even during a period of routine asymptomatic testing [52]. Information on wastewater results is provided on the UCSD public daily dashboard, and targeted email notifications are sent to those living or working in buildings with concerning signals.

A key question challenging programs using wastewater for early detection is when targeted notifications, including email notifications, should be issued to populations at risk in order to increase testing or enhance other mitigation efforts to contain potential transmissions. Crucial to answering this question is quantitative assessment of the relationship between the risk of individual COVID-19 infections and the wastewater test results from associated samplers. There is a recognized need for real-time analysis of the wastewater results to inform decision making.

Results from correlative studies have demonstrated a significant relationship between the viral load in wastewater and individual COVID-19 PCR-based test results. J. Vallejo et al. (2020) [94] used a linear model for the relationship between COVID test cases and viral load detected in the wastewater in A Coruña, Spain. I. Bar-Or et al. (2020) [7] also applied a linear model and concluded that the concentration of the virus RNA in the Bnei Brak sewage correlates with the number of COVID-19 positive individuals in the city. S. Agrawal et al. (2021) [3] found a significant correlation between COVID-19 incidence and viral load observed in wastewater in the Frankfurt metropolitan area. X. Li et al. (2021) [60] performed a meta-analysis for multi-national wastewater data and compared three different models, multiple linear regression, artificial neural network, and adaptive neuro fuzzy inference system for predicting COVID-19 community prevalence (# of infections per 100,000 people) based on wastewater-based quantities including the SARS-CoV-2 RNA concentration.

Several studies utilized not only wastewater results from single time points but also longitudinal time series of wastewater data. N. Krivoňáková et al. (2021) [56] found a high correlation between the number of viral particles in wastewater and the number of individual cases tested 2 weeks later in data from Bratislava. Y. Cao et al. (2021) [16] analyzed the time series of wastewater results using the vector autoregression model to model the weekly variations on the SARS-CoV-2 wastewater concentrations and COVID-19 cases in the Borough of Indiana, PA. Ai et al. (2022) [5] compared different time-series and non-time-series machine learning and deep learning methods including linear model, gradient-boosting decision tree, feed-forward deep neural networks, Facebook Prophet and long short-term memory for the predictive performance of COVID-19 cases in central Ohio. Their results indicated that time-series models outperformed non-time-series models. Other studies [41, 50, 1] have also compared advanced neural networks to predict COVID-19 cases. However, few existing studies focused on extracting interpretable predicting features from time series wastewater results and using them to predict individual test results, which is crucial for facilitating transparent and informed community-level decision making as well as evaluations of the reliability and robustness of the decisions. Comparing to black-box type models, models that can identify the importance of features are particularly advantageous because they provide decision makers with a clear understanding of the factors that contribute to the model's predictions, allowing for more targeted interventions and informed decision making.

In this study, we propose a new framework for feature extraction of longitudinal wastewater test results and for predicting individual COVID-19 infections with the features. As we discuss below, wastewater testing is one example of pooled testing [45, 28, 34]. What is different in our setting is that in standard pooling, investigators can control and standardize how many samples are pooled and how much sample from each person is contributed. In our setting, these factors are impacted by the design of wastewater systems and depend on processes that experimenters do not control. But some principles remain the same; and our analyses are examples of evaluation of diagnostic tests—in our case wastewater tests–based on their properties:

sensitivity, specificity, positive and negative predictive values. Wastewater test results are used to predict the outcome at the level of sets of residence buildings that are associated with manholes in which samplers have been installed. The outcome we seek to predict is whether or not at least one person is infected in the set of buildings associated with a given sampler. We use machine learning to make use of longitudinal time series of wastewater tests to develop optimal rules for notification based on the test properties.

Specifically, we develop hierarchical classification/decision tree models to select important features from the longitudinal series of tests that should trigger notification—that is, that makes it likely that at least one resident is positive. Our analyses of the data on wastewater tests and infections among residents at UCSD derive from information collected in the period from Nov. 2020 to Nov. 2021, covering approximately a whole academic year. Results indicate that by leveraging single-day, long-term and short-term features extracted from the time series of wastewater results, the classification tree model can predict the presence of a positive resident with high sensitivity and satisfactory specificity. Important wastewater features are identified in a hierarchical manner; the most important feature is having a positive wastewater test in at least 3 out of 7 past days. If fewer than 3 out of 7 past days have positive wastewater test results, then the next most important feature is whether 1 out of 5 past days have positive wastewater tests. When applying the model to a set-apart testing set, the prediction accuracy is 72.3%. We also compare the performance of the proposed model to that of random forest models as a benchmark; results indicate the proposed model can predict outcomes with equal or better accuracy while maintaining a high level of interpretability.

Findings derived from the proposed approach have been used to evaluate and refine the current notification system at UCSD. This system sends out timely email notifications to alert residents to a positive wastewater sample associated with their residence buildings and recommend individual COVID-19 tests to contain transmissions at early stages [52]. As a result of this study, in 2021 UCSD modified the email notification system to notify after 3 days of a positive signal. However, during the omicron surge the email notifications were issued after 2 positive days due to the short viral kinetics, indicating the need for ongoing analysis as the virus and epidemiology change.

Our study addresses the urgent need for real-time analysis of data from wastewater surveillance systems and predictive models using wastewater features to predict COVID-19 infections. Results of our study facilitate informed decision making for community-level recommendations and policies intended to contain and prevent transmissions of COVID-19. The approach proposed here provides accurate prediction of individual COVID-19 infection and interpretable feature engineering, and can be readily implemented and applied to other similar systems.

3.2 Model and Methods

3.2.1 Pre-processing of Wastewater Test Results.

As part of the UCSD return-to-learn program, a total number of 140 commercial autosamplers have been deployed in manholes across the UCSD campus, covering teaching, administrative, and residence buildings, including four isolation buildings for students who test positive for COVID-19. In this study, we focus on the data from the 73 manholes covering the 239 residence buildings and their ∼9,700 residents. Figure 1 shows the structure of manholes associated with residence buildings. Twenty-four-hour composite wastewater samples are collected daily from the manholes and analyzed in the laboratory for viral concentration. SARS-CoV-2 signatures are screened via real-time quantitative PCR (RT-qPCR) for the N1, N2, and the E genes [53]. Results are integrated with the campus GIS database to traceback from the manholes to associated upstream residence buildings and identify potential sources of any positive SARS-CoV-2 signals.

As mentioned above, wastewater tests are used to estimate the sensitivity and specificity of different rules for predicting that at least one person will test positive among residents in a set of buildings associated with a given sampler. This requires tracing the source of positive

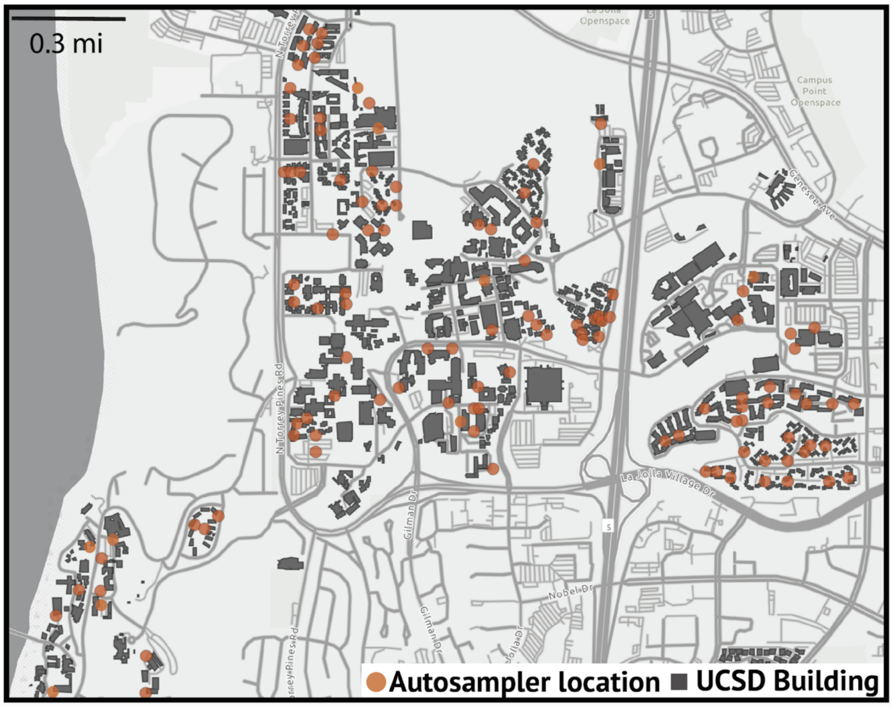


Figure 3.1. Locations of autosamplers installed in manholes (orange circles) connected to UCSD buildings (grey blocks).

signals back to buildings in a way that accounts for the upstream/downstream structure of the sewer network: only the buildings that can contribute to the wastewater are matched to a given manhole. Shown in Figure 3.1 is the structure of manholes connecting to residence buildings [52]. However, the set of buildings associated with a sampler can depend on the results of the wastewater tests. For example, if wastewater from sampler B tests positive but that from an upstream sampler A tests negative, only the buildings contributing wastewater into the sewer between samplers A and B are considered relevant for analysis of signals in sampler B. By contrast, if both samplers are positive, then all buildings associated with either A or B are included in the analysis. The spatially enabled sewer network and subsequent trace of samplers to buildings were stored in and performed by ArcGIS Pro 2.7 (Esri). More details about the sewer network and tracing of samplers can be found in [52] and the interactive web interface at https://returntolearn.ucsd.edu/dashboard/index.html.

Our analysis focuses on the time period of 11/23/20 - 11/13/21, which covers the majority of the academic year 2020-2021 and the first quarter of year 2021-2022. A total of 23,282 wastewater daily samples were collected during this period, and a cutoff of the quantification cycle [Cq] values 39 [53, 51] was used to categorize these samples as positive (<39) vs. negative (>=39). Among the samples, 3,488 were positive and 19,794 were negative.

3.2.2 Ascertainment of Individual Tests Results of COVID-19

During the COVID-19 pandemic, UCSD student residents were required to take individual COVID-19 tests weekly (reduced to bi-weekly after Spring 2021). In addition, in an effort to alert individuals of potential infections in their buildings and encourage them to be tested in one of the on-campus diagnostic testing sites or self-administered test-kit vending machines, targeted email notifications were sent to residents of associated source buildings when positive wastewater SARS-CoV-2 signals were detected in manholes. Notices were also sent to the UCSD campus when a potentially positive building contained a common access area open to the public [52]. Tests are sent to UC San Diego Health labs for processing and the results are saved in an electronic health record (EHR) system [76, 77]. Results of individual tests are available within one day of testing.

Daily individual diagnostic COVID-19 test results of residents in each building are aggregated and merged with the daily wastewater results from manholes associated with the buildings. After excluding all the missing observations, there are a total of 8,853 daily wastewater test records in the merged data, of which 1,212 are positive and 7,641 are negative. The corresponding COVID-19 individual diagnostic test results among students resident in campus housing indicate 170 are positive and 8,683 are negative.

Of the 170 COVID-19 individual diagnostic positive test results among students residing in campus housing, only 54.7% have a tested-positive wastewater sample from the associated manhole on the same day of the individual test, indicating using daily wastewater test results alone cannot achieve satisfactory prediction of individual infections of COVID-19 in associated buildings. Potential reasons for the observed discrepancy between individual tests and wastewater results include delays in being tested or getting results among those who had become infected. For example, among infected residents, there could be a delay in the manifestation of symptoms or absence of symptoms; for those reasons or others, the individual tests may not take place until a few days after the actual onset of the infection. There can also be false negative wastewater test results arising from low viral concentration, even if one or more residents in associated buildings have become infected. In addition, there is a possibility of false positives in the wastewater results. To understand the implications of the wastewater samples and to optimize the utility of the wastewater surveillance system in detecting individual infections, a definition of the outcome of individual infections that accounts for potential lags between the wastewater and individual test results is needed.

Here we propose a 3-day time window approach to define the outcome of individual infections. Using the date of wastewater test as an anchor point, for each manhole we examine individual diagnostic test results of residents in associated buildings in the 3-day window including the date of wastewater test and the day before and after the wastewater test. This outcome is defined as positive for an individual-level test if there exists at least one positive individual COVID-19 test result among residents in associated buildings in this time window. The proposed time window addresses the time lag between the wastewater and individual tests by including positive individual tests in intervals of one day before to one day after the detection of a positive wastewater test. A sensitivity analysis using a longer window of 6 days has also been conducted and its results are described in the Appendix; this choice of window leads to a similar model as does the analysis with a 3-day window.

3.2.3 Model for Predicting Individual COVID-19 Infections Using Wastewater Results

To detect individual COVID-19 infections, we use multiple interpretable features extracted from wastewater time series data, which includes both single-day test results and short-term/long-term trends. The proposed features provide a comprehensive characterization of different aspects of the wastewater test results. The list of features includes single-day wastewater results up to five days before the day in question, short-term wastewater trends including whether at least 1 out of the past 3 days, 1 out of the past 4 days, 3 out of the past 4 days, 1 (or 2, 3) out of the past 5 days contains positive wastewater signals, and long-term wastewater trend including whether at least 2 (or 3) out of the past 7 days contains positive signals, and whether wastewater results in all of past 3 consecutive days are positive.

We adopt a machine learning approach–classification trees– [14, 82, 48, 84], to predict individual COVID-19 infections defined using the 3-day window with the above features extracted from wastewater signals. The classification tree derives from a hierarchical model that predicts outcomes with recursive binary partitions based on an ordering of the importance of the predictors. At each node/leaf of the classification tree, the feature capable of reducing the maximal amount of Gini impurity, a criterion to measure the mixture of different classes of the outcome, is selected to partition the data [44, 75, 73]. Predictors that appear in earlier nodes are considered more important in predicting the outcome [14]. The ordering of importance of predictors is crucial in our study, as we aim to accurately predict the presence of infections in residence buildings and to reveal important and interpretable features from wastewater test results to aid in decision making for campus-wide recommendations and mandates. To avoid overfitting and improve interpretability, we apply constraints on the model complexity using a penalty parameter *cp* [13, 89, 6]. In addition, the classification tree mitigates collinearity among predictors as a result of its variable selection mechanism based on feature importance [92].

We also incorporate a re-weighting mechanism in our model to address the important issue of imbalance in the outcome. There are many more negative than positive individual test results in the data, which represents a typical imbalance in the outcome of individual testing of COVID-19 in similar communities. Models optimizing prediction accuracy when trained with the data without any adjustment tend to classify all outcomes as negative due to over-representation of the negative outcomes. To address this issue we re-weight the data by allocating larger weights to positive than to negative outcomes in training the classification tree models. This approach is similar to over-sampling the minority class and under-sampling the majority class, which has been shown to achieve better classifier performance [100, 19, 18].

To evaluate the performance of the proposed approach, we partition the data from 11/23/20 - 11/13/21 into a training and a testing set. The training set includes data from 11/23/20 to 04/30/21 and the testing set includes data from 06/30/21 to 11/13/21. The partition of the dataset is not random: it preserves the chronological ordering of the dates of the test results as definitions of the features extracted from the wastewater samples rely on the chronological ordering of the dates. In addition, results in the same period are expected to behave similarly as the policies, circulating variants, and other pandemic conditions vary with the chronologic time of measurement. Comparing model performance in the training and testing sets also provides insight into the influence of these factors on the effectiveness of the wastewater surveillance system. We exclude the samples in May and June due to potential data quality issues; further investigation of the wastewater results during this period is needed. In the Appendix, we present a sensitivity analysis that includes data from this period, and we obtain the same model as described in the following section. This analysis serves to demonstrate the robustness of our results.

3.3 Results

3.3.1 Classification Tree Trained with the Training Set

Figure 3.2 shows the result of classification tree trained with the training set. From the top (root) to the bottom (leaves) of the tree, we show the features selected to predict the outcome; features closer to the root are considered to be more important. The branches of each node, visualized by the arrows, describe the features and the two possible conditions used for binary partitioning of the data according to which condition is satisfied. The color of each node indicates the predicted outcome for records partitioned into the category corresponding to the node: red indicates a positive predicted outcome of at least one infection in associated buildings, and blue,

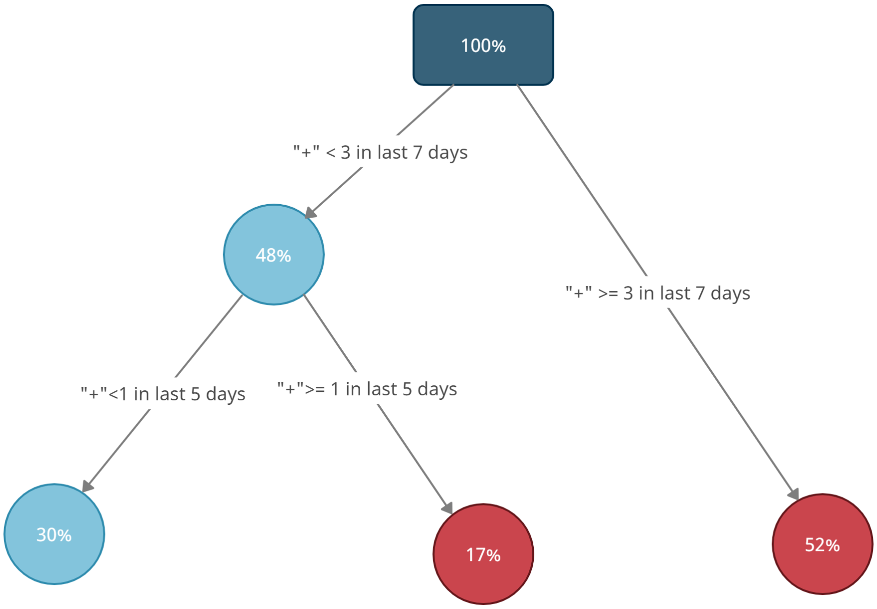


Figure 3.2. Classification tree model trained with the training set only. Wastewater time series features are used to predict individual COVID-19 test results. The red node means a positive predicted outcome and the blue node means a negative predicted outcome. The value inside each node denotes the percentage of the total data records that falls in the category of the node. “+” means number of positive wastewater results. For example: “+ < 3 in last 7 days” means there were less than (<) 3 days of positive wastewater results in the last 7 days of wastewater testing.

a negative predicted outcome. The value in the circle of each node indicates the percentage of the partitioned data records in the whole data.

The model in Figure 3.2 indicates the most important feature in predicting the outcome is whether fewer than (<) 3 days in the last 7 had positive wastewater test results. The outcome is predicted to be positive if wastewater results are positive in at least 3 out of the past 7 days, and negative otherwise. Given positive wastewater results on fewer than 3 out of the past 7 days, the second most important feature is whether none of the past 5 days have positive wastewater results. If yes then the outcome is predicted to be negative, otherwise to be positive.

The decision tree in Figure 3.2 is fitted with weights of positive outcomes equal to (2 / # positive classes) and weights of negative outcomes equal to (1 / # of negative classes). Note the weights are standardized by the total number of positive and negative outcomes, respectively, and then multiplied by scalers based on the importance placed on correctly predicting the positive and negative outcomes. Our choice of weights reflects the priority of sensitivity (true positive rate) over specificity (true negative rate) in predicting positive individual infections. A sensitivity analysis using weights equal to the reciprocal of class sizes for both classes is performed in Appendix. The value of the penalty parameter on model complexity $cp = 0.02$ is chosen to balance optimal performance in the training set as suggested by cross-validation while maintaining a small number of nodes in the tree for model interpretability. A sensitivity analysis using $cp = 0.001$ to train the model is available in the Appendix to further investigate the influence of model complexity on the prediction performance and the trade-off between model complexity and interpretability.

Table 3.1 shows the confusion matrix of the predictions when applying the model to the training set. The sensitivity (True Positive Rate, TPR = TP/(TP+FN)) is 83.7% and the specificity (True Negative Rate, TNR = TN/(TN+FP)) is 58.5%. Note that the calculations of sensitivity and specificity are unaffected by the weights allocated to positive and negative outcome classes as the weights appear in both numerators and denominators and cancel out. The overall weighted prediction accuracy is 75.3%, which is calculated by

$$\frac{\sum_{i=1}^{n} w_i [I\left(predict\ positive \mid positive\right) + I\left(predict\ negative \mid negative\right)]}{\sum_{i=1}^{n} w_i}$$

where w_i denotes the weight of sample i, $I\left(predict\ positive \mid positive\right)$ denotes the indicator function that sample i has a positive outcome that is predicted to be positive, and $I\left(predict\ negative \mid negative\right)$ denotes the indicator function of sample i has a negative outcome that is predicted to be negative. It is expected to observe a higher estimated sensitivity than specificity as we are over-sampling the positive outcome class compared to the negative class.

To evaluate the prediction performance of the classification tree, we then apply the model to the set-apart testing set in the period of 06/30/21 - 11/13/21. The confusion matrix is provided

Table 3.1. Confusion matrix of results obtained from applying the model (trained with the training set) to the training set.

	Predict positive	Predict negative
Actual positive	83.7%	16.3%
Actual negative	41.5%	58.5%

Table 3.2. Confusion matrix of results obtained from applying the model (trained with training set only) to the testing set.

	Predict positive	Predict negative
Actual positive	77.1%	22.9%
Actual negative	37.2%	62.8%

in Table 3.2. For the testing set, the sensitivity decreased from 83.7% to 77.1% while the specificity increased from 58.5% to 62.8%. The overall weighted prediction accuracy is 72.3%. The testing set contains the period in which most of the student residents had already received vaccination and the wave of the highly infectious SARS-CoV-2 Omicron variant had not yet arrived [11]. Therefore, fewer infected cases were observed and thus underrepresented the total population. Despite the evolving nature of the pandemic, the model performed well and was able to predict individual infections with satisfactory accuracy and high sensitivity. We also trained a model on the testing set alone and compared it with the model trained with the training set; the comparison of results is available in the Appendix.

3.3.2 Influence of Weights

In this section we investigate the role of relative weights of positive and negative outcomes in the prediction. For simplicity of notation, we denote a relative weight of (*a* / #positive classes) : (*b* / #negative classes) for positive vs. negative outcomes as *a:b*. For example, the model in Figure 2 is trained with weights 2:1; this weighting places a double amount of emphasis on records with positive outcomes compared to those with negative outcomes after standardizing by

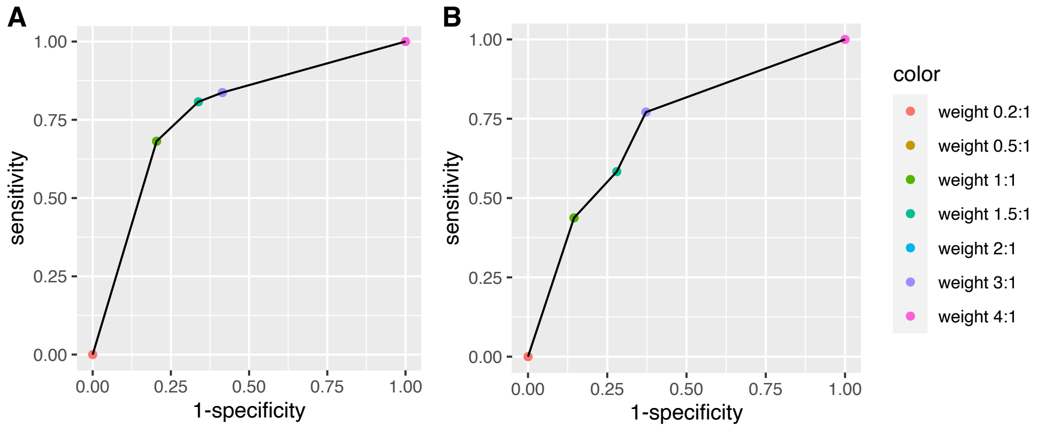


Figure 3.3. ROC (Receiver Operating Characteristic) curves of models trained with different relative weights for positive and negative outcome classes using data of the training set only. The left panel shows results obtained from applying the models to the training data. The right panel shows results of applying the models trained with the training set to the testing set.

the total numbers of positive and negative outcomes. The trained decision tree model for relative weights 1:1 is available in the Appendix as a sensitivity analysis.

Figure 3.3 displays the receiver operating characteristics (ROC) curve [40, 88], which demonstrates a trade-off between sensitivity and specificity; the *x*-axis indicates one minus the specificity, and the *y*-axis indicates the sensitivity. This curve permits a comparison of the performance of models trained with varying weights. Detailed results are provided in Table 3. With relative weights on the positive class as small as 0.2:1, all the outcomes are predicted to be negative; hence, the sensitivity is 0 and the specificity is 1. As the weight for positive class increases, the sensitivity also increases, and the specificity decreases. With relative weights of 4:1 or greater, all outcomes are predicted to be positive, yielding sensitivity of 1 and specificity of 0.

Table 3.4 summarizes the importance of features in models trained with different weights given by orders of nodes appearing in the classification trees. For results to be comparable, *cp* value of 0.02 is used in training all models with different weights; this approach leads to different numbers of nodes under different weight settings. For all models, the root nodes are

Table 3.3. Detailed values of Sensitivity and (1-Specificity) for ROC curves in Figure 3.3.

Relative Weight (positive vs. negative outcome)	Sensitivity (training set performance)	1-Specificity (training set performance)	Sensitivity (testing set performance)	1-Specificity (testing set performance)
0.2:1	0	0	0	0
0.5:1	68.1%	20.4%	43.8%	14.5%
1:1	68.1%	20.4%	43.8%	14.5%
1.5:1	80.7%	33.8%	58.3%	28.0%
2:1	83.7%	41.5%	77.1%	37.2%
3:1	83.7%	41.5%	77.1%	37.2%
4:1	100%	100%	100%	100%

Table 3.4. Importance of features extracted from wastewater time series given by models trained with different relative weights. "*a*_out_*b*" in the table represents the dichotomous feature of whether there were at least *a* out of the previous *b* days with positive wastewater test results.

Relative weights	0.5:1	1:1	1.5:1	2:1	3:1
1^{st} level feature	3_out_7	3_out_7	3_out_7	3_out_7	3_out_7
2^{nd} level feature			1_out_5	1_out_5	1_out_5
3^{rd} level feature			2_out_7		

defined by whether or not fewer than 3 out of the past 7 days have positive wastewater signals; this is consistently the most predictive wastewater feature for predicting individual COVID-19 infections. In all models with a lower level node/leaf, the next most important feature is whether or not none of the previous 5 days have positive wastewater signals. Combined with the result of the root node, a predictive model that is robust to the choice of weights consistently includes the dichotomous features: 3 or more out of 7 days wastewater positive (yes/no) and 1 to 5 of the previous days wastewater positive (vs 0 days). This model leverages features characterizing wastewater results both in a long-term trend of 7 days and in shorter periods of 5 days.

3.3.3 Prediction With Random Forest Model as a Benchmark

To further evaluate the prediction performance of the proposed decision tree model, we apply a weighted random forest model [12] consisting of an ensemble of 1,000 individual weighted decision trees. As in the classification tree model, weights are applied for oversampling the positive individual cases. The random forest is known for its high prediction accuracy but lacks the interpretability of the classification trees. Comparing the performance of the proposed model to that of the random forest enables us to assess the proposed model with a reliable benchmark and to understand the trade-off between the interpretability and prediction accuracy of models.

Figure 3.4 shows the ROC curve of sensitivity vs. (1-specificity) when applying training-set-fitted decision tree and random forest models under different weight settings to the testing data, which were not used in training the models. Detailed results are provided in Table 3.5. The proposed decision tree models generally outperform the random forest models in the same weight settings, especially when the relative weights of positive vs. negative outcomes are high. For the random forest approach, the optimal weight, with high sensitivity and relatively high specificity, is 3:1. In this case, both sensitivity and specificity equal to 68.8%, leading to a 68.8% prediction accuracy, while the proposed decision tree model has a prediction accuracy of 75.3%. One possible reason for the random forest to under-perform compared to the proposed decision tree is that the random forest is based on bootstrap (or subsampling) of the data, which breaks the chronological structure of the time series in the data and thereby potentially affects the prediction performance.

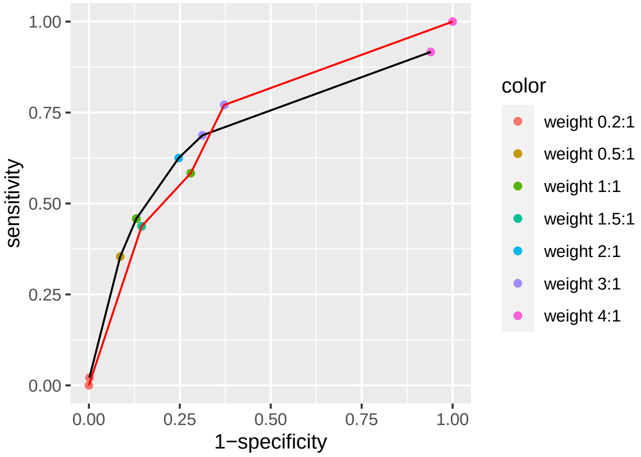


Figure 3.4. ROC curves of models trained with different relative weights for positive and negative outcome classes using random forest model (black) and classification tree model (red). The figure shows the results of applying the models trained with the training set to the test set.

Table 3.5. Detailed values of Sensitivity and (1-Specificity) in ROC curves of Figure 3.4.

Relative Weight (positive vs. negative outcome)	Sensitivity (black)	1-Specificity (black)	Sensitivity (red)	1-Specificity (red)
0.2:1	2.1%	0.1%	0	0
0.5:1	35.4%	8.6%	43.8%	14.5%
1:1	45.8%	13.1%	43.8%	14.5%
1.5:1	45.8%	13.1%	58.3%	28.0%
2:1	62.5%	24.7%	77.1%	37.2%
3:1	68.8%	31.2%	77.1%	37.2%
4:1	91.7%	94.0%	100%	100%

3.3.4 Positive Predictive Value (PPV) and Negative Predictive Value (NPV)

We further examine the positive predictive value (PPV) and negative predictive value (NPV) of the predictions of individual infections as defined below:

$$\text{Positive predictive value (PPV) of wastewater (WW) test } = \frac{\text{Sensitivity of WW test * prevalence}}{\{(\text{sensitivity * prevalence}) + (1\text{-specificity})\ (1 - \text{prevalence})\}} = \text{TP/(TP+FP)},$$

$$\text{Negative predictive value (NPV) of WW test } = \frac{\text{Specificity of WW test * (1-prevalence)}}{\{\text{specificity *(1- prevalence)} + (1\text{- sensitivity})\ (\text{prevalence})\}} = \text{TN/(TN + FN)},$$

Where TP and FP are numbers of true and false positives and TN and FN are numbers of true and false negatives in the prediction, and the prevalence is the proportion of true positives among all tested units of observation (which could be, for example, at a building or individual level).

These quantities can be particularly useful in developing policies regarding control of the COVID-19 epidemic. In the case of pooled tests, results can help in using testing resources more efficiently—by focusing intensive testing where cases are most likely to reside. In addition, the tests can provide an early warning about the potential for at least one resident of a building unit to be infected. To make best use of the wastewater tests, we estimate the probability that there is at least one infected person in a residence given a positive wastewater test. This estimate will aid in evaluating the cost-benefit of different strategies for testing the residents. In addition, knowledge of the relationship between the timing of positive wastewater tests and positive individual-level tests can inform us about when—or at what schedule–it is best to offer the latter to residents.

Our testing setting is a little more complex than usual, because the wastewater test is a pooled test that aggregates results of buildings associated with the same manholes; hence, the number who contribute to the pool varies across tests—which are done at the residence level. Furthermore the prevalence of interest is at the residence level; as noted above, we define a residence to be a true positive if there is at least 1 infected resident in the residence. Like the wastewater itself, this definition is at the residence building level.

The prevalence at the residence building level p_c can be estimated from the prevalence p at the individual level given the number of residents (n), under the assumption of independence across infection events across them: $p_c =$ prob of (>=1 infected resident) = $1-(1-p)^n$ where p is individual-level prevalence. Because most detected infection events we observed are only in a single person, we believe that violation of this assumption has little effect on our estimates. As the prevalence of COVID-19 and the number of residents vary with date, the estimates of PPV and NPV will vary with date as well. There are also possible dilution effects that could affect the estimations. For example, the detectability of SARS CoV-2 genetic material may depend on the total number of residents living in the upstream of the manholes.

Here we provide approximate building-level estimates of the PPV and NPV and demonstrate how they are affected by the number of residents in buildings associated with manholes. We focus on the period of the week before Fall 2021 quarter begins, as most student residents are in the process of moving back onto campus during that week, and are required to take individual-level tests as soon as they move into their residences. The curves of PPV and NPV as a function of the number of people in residence buildings are shown in Figure 3.5. We note that the PPV and NPV are quite sensitive to the number of residents; the usefulness of wastewater tests must be considered in this context. Negative tests are less reassuring as the number climbs near 1,000; whereas PPV only approaches 50% when the number of residents is near 250.

3.3.5 Sensitivity Analyses

As previously mentioned, multiple sensitivity analyses are conducted to examine the effects of different model parameters; these include: 1) a different definition of outcome using a longer-term time window, 2) varying weights used in training the model to balance positive and negative individual test results, and 3) different levels of model complexity. In addition, we use the testing data set alone to train the models and compare them to models trained with the

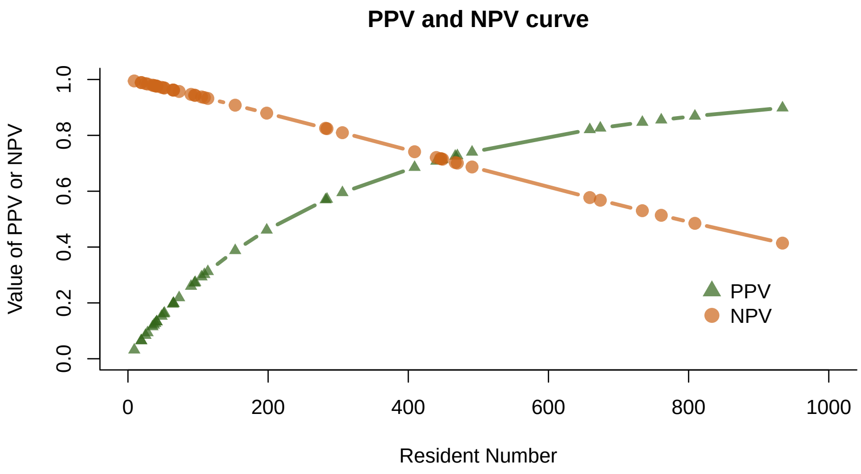


Figure 3.5. PPV and NPV curves as functions of numbers of residents in buildings associated with manholes.

training set in order to gain a deeper understanding of the difference between the two datasets and how the trained models vary with different time period of the data. All of the analysis results are available in Supporting Information.

3.4 Concluding Remarks

This paper proposes a model framework for predicting the presence of infections in residence buildings using results from wastewater surveillance systems. The goal of this study is to make use of wastewater test results to inform decision making regarding notification of wastewater results to guide public health strategies intended to control the spread of individual COVID-19 infections in communities. To this end, we extract features that characterize wastewater test results over time, develop classification/decision tree models to select important features, use them to predict probabilities that there is at least one individual infection in residences, and finally optimize the COVID-19 test notification strategy.

We used the classification tree to analyze data from the wastewater surveillance system and individual-level COVID-19 tests of residents on UCSD campus from Nov 2020 to Nov 2021. Results reveal that the best predictor of positive individual level tests in residence buildings is whether or not the wastewater results were positive in at least 3 of the past 7 days. Using a set-apart testing set, we demonstrate the accuracy of these predictions. Our results suggest that the proposed analysis approach can be useful in using wastewater to guide policies around notifications for building residents to seek individual-level testing. Features included in the model are robust to changes in weights of positive and negative individual test results, and the features discovered to be most important are consistent across different weight choices in balancing the positive and negative outcomes in the data. Discoveries from the analysis have been useful in assisting decision making in the UCSD campus-wide Return-to-Learn program and incorporated into the email notification system.

Although our approach is motivated by and developed for the UCSD Return-to-Learn program, the model framework proposed here can be readily applied to similar wastewater surveillance systems to predict individual COVID-19 infections in communities and to facilitate decision making processes in making community-wide guidelines, mandates and policies for containing transmission of the virus. In applying the proposed approach, several aspects of the model may need to be adjusted by researchers and/or policymakers according to pandemic conditions at the time of analysis. First, in defining the outcome of individual COVID-19 infections, we introduced here a time window of 3 days to account for potential lags from onset of infections to testing and mismatches between the individual infections and the wastewater results. If the required test frequency changes, the optimal performance of wastewater tests may require that the time window be adjusted accordingly. Second, pandemic conditions vary over time because of the regular appearance of new variants and changes in behavior with regard to masking and other mandates and mitigation strategies. Coverage rates of vaccinations may improve over time in some communities, but the effectiveness of older vaccines constantly wanes. Therefore we recommend an "online" learning approach in which the prediction model is updated regularly as new data become available to ensure that the model reflects prevailing conditions.

Chapter 3, in full, has been submitted for publication of the material as it may appear in

Tuo Lin, Smruthi Karthikeyan, Alysson Satterlund, Robert Schooley, Rob Knight, Victor De Gruttola, Natasha Martin, Jingjing Zou. Optimizing campus-wide COVID-19 test notifications with interpretable wastewater time-series features using machine learning models. The dissertation author was the primary author of this paper.

Chapter 4

Non-parametric Causal Inference for Mann-Whitney-Wilcoxon Rank Sum Test Using Random Forest

4.1 Introduction

Increased number of outliers is a big problem in data analysis, yielding uninterpretable and often biased results when analyzed using mean-based statistical models, including most popular models such as two-sample t-test and linear regression. Rank-based methods such as the Mann-Whitney Wilcoxon rank sum test (MWWRST) introduced in Chapter 1 and rank regression can effectively address this statistical problem without any subjective bias as in other ad-hoc methods such as winsorized estimates based on truncating the outliers to 3 times IQR [61].

In randomized control trials, MWWRST is commonly used as an alternative to two-sample t-test when there are outliers in data. In non-randomized observational studies, treatment or exposure effects cannot be estimated using MWWRST because of confounders. Rubin's potential outcome framework allows us to define and estimate the average treatment effect [78]. However, existing mean-based methods break down in the presence of outliers. In our previous paper [under review], we proposed a novel Mann-Whitney-Wilcoxon type of causal effect, $Pr(Y_i^{(0)} < Y_j^{(1)})$, to address this problem. To estimate causal effects (see details of this causal effect in Section 4.2) under this framework, a semiparametric FRM model with a doubly robust estimator, akin to the method in Chapter 2, has been proposed. Although methodologically sound, such semiparametric models require correct model specifications for either propensity score model or outcome regression model to ensure consistency. If the linear predictor for these models involve non-linear and non-additive associations with the explanatory, the doubly robust estimators will generally be biased. We consider nonparametric models to address this major weakness of semiparametric models.

The advancement of data collection and storage technology creates possibilities for us to access high dimensional data such as electronic health records and DNA sequences. Many researchers start to extract information from these types of observational data and incorporate them into clinical studies [30, 29], yet many classic statistical methods such as generalized linear model (GLM) could not be applied due to the high dimensionality of the data. Machine learning methods including random forests, support vector machine and neural network have been widely adopted to model the association between the outcomes and high dimensional variables of interest and have been shown to have good prediction performances [12, 105, 83]. While successful in prediction, most machine learning methods do not have a well-established theory for their asymptotic properties, which is important in hypotheses testing and constructing interval estimates. A recent work from Wager and Athey (2018) [96] has justified the consistency and asymptotic normality of the random forests for non-parametric linear regression analysis and extended such properties to the causal inference setting. In this project, we propose a between-subject random forests estimator for our Mann-Whitney-Wilcoxon type of causal effect and extend the results of Wager and Athey to show the asymptotic properties of our estimator.

Different from our previous work and other causal inference methods aimed at estimating the average treatment effect, our goal is to explore the heterogeneous treatment effects for pairs of subjects from the two treatment groups. Specifically, we consider applying random forests to model $Pr(Y_i^{(0)} < Y_j^{(1)} \mid X_i = X_j = x)$. This model is particularly useful in personalized medicine, when interest is centered on treatment differences at the individual-level. The approach of random forests matches coincidentally with the concept of heterogeneous treatment effects estimation, since in growing the random forests we divide the sample into a few subgroups with inference based on these subgroups. Each subgroup has a relatively small size, leading to a higher sensitivity to outliers, and our rank-based method provides an effective solution to this problem.

Although our main focus in this project is causal inference, the proposed random forests method can be easily extended to model other types of between-subject outcomes defined by a function of paired subjects' outcomes, such as viral genetic linkage network in Chapter 2 and microbiome beta diversity. The asymptotic properties derived in this work can also be extended to inference about the other types of between-subject outcomes to facilitate analyses of emerging high-dimensional data arising from different biomedical and psychosocial research areas.

4.2 Mann-Whitney-Wilcoxon Type of Causal Effect

The causal inference paradigm is introduced for a few purposes: (1) to define and characterize causal effect, (2) to explicitly identify and control for factors of confounding in studies that is not feasible or even ethical to apply randomization. Limitations of non-randomization based studies, although rather obvious on intuitive grounds, are actually nearly impossible to characterize analytically. In fact, such an inquiry raises a more fundamental question as to how treatment effects are defined in the first place.

The concept of potential outcome addresses this fundamental gap. The idea is that for every patient, there is a potential outcome for each treatment condition received, and the treatment effect is defined by the difference between the outcomes in response to the respective treatments from the same individual. Thus, treatment effect is defined for each subject based on his/her differential responses to different treatments, thereby free of any confounder.

For notational brevity, consider two treatment conditions and let $y_i^{(k)}$ denote the potential outcome for the ith subject under the kth treatment ($1 \leq i \leq n$, $k = 0,1$). For convenience, let

$k = 1$ (0) indicate the treated (controlled) condition. We observe only one of the two outcomes, $y_i^{(1)}$ or $y_i^{(0)}$, depending on the treatment received by the patient. The difference between $y_i^{(1)}$ and $y_i^{(1)}$ is attributable to the differential effect of the treatments, since there is absolutely no other factor that may also influence the outcome, or confounder, in this case. Thus, the causal effect of the intervention for each subject is:

$$\text{Individual causal effect:} \quad \Delta_i = y_i^{(1)} - y_i^{(0)}, \quad 1 \le i \le n.$$

Although the potential outcomes and individual causal effect are well-defined, the latter Δ_i cannot be computed, since only one of the $y_i^{(k)}$'s is available. This "missing data" feature in particular precludes direct applications of conventional statistical methods for inference about causal effect such as the sample mean.

By taking the mathematical expectation, we obtain:

$$\text{Average (population) causal effect:} \quad E(\Delta_i) = E\left(Y_i^{(1)} - Y_i^{(0)}\right) = E\left(Y_i^{(1)}\right) - E\left(Y_i^{(0)}\right).$$

In the above form, $E\left(Y_i^{(k)}\right)$ is expressed as the (population) mean μ_k of $Y_i^{(k)}$ for the kth treatment.

Within the context of heterogeneous treatment effects, the causal effect at x is defined as:

$$\tau(x) = E\left(Y_i^{(1)} - Y_i^{(0)} \mid X_i = x\right) = E\left(Y_i^{(1)} \mid x\right) - E\left(Y_i^{(0)} \mid x\right).$$

As discussed in Section 4.1, in our between-subject outcomes setting, the causal effect at x, $\delta(x)$, is defined as:

$$\delta(x) = E\left(I\left\{Y_i^{(0)} \le Y_j^{(1)}\right\} \mid X_i = X_j = x\right). \tag{4.1}$$

Note that unlike mean-based method, $I\left\{Y_i^{(0)} \le Y_i^{(1)}\right\}$ for a single subject i is non-identifiable because we only observe one of $Y_i^{(1)}$ and $Y_i^{(0)}$ at the same time. Thus, we replace

$Y_i^{(1)}$ by $Y_j^{(1)}$ that has the same value of X.

4.3 Between-subject Random Forests

In this section, we aim to construct the between-subject random forests in a similar manner to the within-subject random forests. Since random forests are ensemble of multiple classification/regression trees, we start by the splitting procedure for a single tree, which resembles the widely known classification and regression tree (CART) analysis for within-subject outcomes [14, 12].

Let $Z_i = (Y_i, X_i)$ be the observed data for subject i, where Y_i is the outcome of interest and X_i is a vector of independent variables. We first recursively split the feature space into a sets of leaves L, each contains a few training samples. Let L_i be the leaf that contains subjects i and $L_i^{(w)}$ as two subleaves corresponding to the two treatments around i-th subject ($w = 0, 1$). A single tree estimator $\widehat{\delta}_{tree}$ is defined as

$$\widehat{\delta}_{tree}(x) = \frac{1}{\left|L_i^{(0)}\right|\left|L_i^{(1)}\right|} \sum_{Y_{j0} \in L_i^{(0)}} \sum_{Y_{l1} \in L_i^{(1)}} I\left\{Y_{j0} < Y_{l1}\right\}.$$

By subsampling B times ($B \to \infty$) and leveraging $\widehat{\delta}_{tree}^{(b)}$, $b = 1, \ldots, B$ obtained from each subsample, we define a random forests estimator $\widehat{\delta}(x)$ of (4.1) as follows:

$$\widehat{\delta}(x) = \frac{1}{B} \sum_{b=1}^{B} \widehat{\delta}_{tree}^{(b)}(x).$$

The details of how to construct the between-subject random forests are given in Algorithm 1.

4.4 Asymptotic Theory

The asymptotic theory including the convergence and consistency of within-subject random forests has been studied in some previous work [10, 81, 97]. However, until recently,

Algorithm 1. Between-subject Random Forests

The algorithm follows the double-sampling tree algorithm [ref], which indicates splitting the whole subsample into 2 parts, half for building the tree and half for inference.

Input: a subsample (Z_i, W_i) with $i = 1, ..., s,\ s < n$, where W_i is the treatment variable with a value w.

1. Grow the tree for each of subsampling samples b by applying within-subject double-sampling (or propensity) tree algorithm. The splitting criteria follows the standard one for CART, that is minimizing mean-squared error of predictions.

2. Identify leaf node for subject i, denoted as L_i, which consists of two subleaves, $L_i^{(w)}$, corresponding to the two treatments around i-th subject $(w = 0, 1)$:

3. Estimate $\delta(x)$ by:

$$\widehat{\delta}_{tree}(x) = \frac{1}{\left|L_i^{(0)}\right|\left|L_i^{(1)}\right|} \sum_{Y_{j0} \in L_i^{(0)}} \sum_{Y_{l1} \in L_i^{(1)}} I\left\{Y_{j0} < Y_{l1}\right\},$$

   where $\left|L_i^{(w)}\right|$ denotes the size of $L_i^{(w)}$. If we use double-sampling trees, this estimation should be based on the set-apart testing set.

4. Aggregate all trees and calculate

$$\widehat{\delta}(x) = \frac{1}{B} \sum_{b=1}^{B} \widehat{\delta}_{tree}^{(b)}(x).$$

the asymptotic normality has been shown by Wager and Athey in their seminal paper [96]. In our work, we have successfully extended their work to show the asymptotic unbiasedness and normality of our between-subject random forests. Before diving deep into the asymptotic theory, let us first look at a few definitions for a single tree that consists the between-subject random forests.

Definition 4. *A tree grown based on a subsample $(Z_1, \ldots, Z_s)$ is honest if the tree uses half of the subsample to grow the tree and another half of the subsample to make inference, that is saying, there are no overlap samples that are used for both tasks.*

To ensure the randomness of the spit, in other words, to ensure the size of a terminal node is no larger than other terminal nodes, we need the definition of a random-split tree.

Definition 5. *A tree is a random-split tree if at every step of the tree-growing procedure, the probability that the next split occurs along any one of the features is bounded below by π/d for some* $0 < \pi \leq 1$.

Finally, to ensure the size of each terminal node not diminishing, we have the following regularity definition.

Definition 6. *A tree predictor is (α, k)-regular for some $\alpha > 0$ if each split leaves at least a fraction α of the available samples on each side of the split and there are between k and $2k-1$ observations in each terminal node of the tree.*

Definition 7. *A predictor is symmetric if the (possibly randomized) output of the predictor does not depend on the order $(i = 1, 2, \ldots)$ in which the training examples are indexed.*

4.4.1 Bias

After introducing the splitting algorithm and definitions of regression trees, we start by studying the asymptotic unbiasedness of such regression trees. We use the same definition of a diameter of a leaf $L(x)$ as in graph theory, which is the length of the longest path contained in

$L(x)$. The following Lemma 8 indicates the diameter of the terminal node is bounded away from 0 in probability.

Lemma 8. *Let T be (α,k)-regular, random-split tree and let $L(x)$ denote its leaf containing x. Suppose that $X_1,...,X_s \sim U([0,1]^d)$ independently. Let $s_1(s_0)$ be the number of cases (controls) and $s = s_1 + s_0$, $s_{min} = \min(s_1, s_0)$. Then, for any $0 < \eta < 1$, and for large enough s,*

$$\Pr\left[\operatorname{diam}_j(L(x)) \geq \left(\frac{\varepsilon s}{2k-1}\right)^{-\frac{0.99(1-\eta)log((1-\alpha)^{-1})}{log(\alpha^{-1})}\frac{\pi}{d}}\right] \leq \left(\frac{s_{min}}{2k-1}^{-\frac{\eta^2}{2}\frac{1}{log(\alpha^{-1})}\frac{\pi}{d}}\right),$$

where ε is the asymptotic lower bound of s_{min}/s to ensure positivity assumption.

Lemma 8 can imply an asymptotic bound on the bias of a single tree estimator $\widehat{\delta}_{tree}(x)$. Since $\widehat{\delta}$ is the average of independent $\widehat{\delta}_{tree}(x)$, we derive the following Theorem 9 for the asymptotic unbiasedness.

Theorem 9. *Under conditions of Lemma (8), suppose moreover that $\delta(x)$ is Lipschitz continuous and the trees in the random forest are honest. Then, provided that $\alpha \leq 0.2$, the bias of the random forest at x is bounded by*

$$\left|E\left[\widehat{\delta}(x)\right] - \delta(x)\right| = O\left((\varepsilon s)^{-\frac{1}{2}\frac{log((1-\alpha)^{-1})}{log(\alpha^{-1})}\frac{\pi}{d}}\right).$$

The proof of Lemma 8 and Theorem 9 can be found in Appendix.

4.4.2 Asymptotic Normality

The proof of the asymptotic normality relies heavily on U-statistics theory. First, the random forest (RF) is defined as

$$RF(\mathbf{x}; Z_1, \ldots, Z_n) = E_{\xi \sim \Xi}[T(\mathbf{x}; \xi, Z_{i1}, \ldots, Z_{is})] = \binom{n}{s}^{-1} \sum_{(i_1,\ldots,i_s) \in C_s^n} T(\mathbf{x}; Z_{i1}, \ldots, Z_{is}). \quad (4.2)$$

And the Hájek projection of RF is defined as

$$\overset{\circ}{RF}(\mathbf{x}; Z_1, \ldots, Z_n) = \sum_{i=1}^{n} E\left[RF(\mathbf{x}; Z_1, \ldots, Z_n) \mid Z_i\right] - (n-1)\,\theta(\mathbf{x}) \\ = \sum_{i=1}^{n} \binom{n}{s}^{-1} \sum_{(i_1,\ldots,i_s)\in C_s^n} E\left[T(\mathbf{x}; Z_{i1}, \ldots, Z_{is}) \mid Z_i\right] - (n-1)\theta(\mathbf{x}) \tag{4.3}$$

where $\theta(\mathbf{x}) = E\left[T(x; Z_{i1}, \ldots, Z_{is})\right]$. The projection in (C.4) can also be expressed in a centered version by

$$\overset{\circ}{RF}(\mathbf{x}; Z_1, \ldots, Z_n) - \theta(\mathbf{x}) = \sum_{i=1}^{n} E\left\{\left[RF(\mathbf{x}; Z_1, \ldots, Z_n) - \theta(\mathbf{x})\right] \mid Z_i\right\} \\ = \sum_{i=1}^{n} \binom{n}{s}^{-1} \sum_{(i_1,\ldots,i_s)\in C_s^n} E\left\{\left[T(\mathbf{x}; Z_{i1}, \ldots, Z_{is}) - \theta(\mathbf{x})\right] \mid Z_i\right\}. \tag{4.4}$$

In classic U-statistics theory, the number of the arguments s in $T(\mathbf{x}; Z_{i1}, \ldots, Z_{is})$ is finite and constant, then the asymptotic properties can be straightforwardly derived for the Hájek projection [ref]. However, in the RF case $s \to \infty$ as $n \to \infty$, the classic U-statistics theory could not be applied directly and we seek a new proof for this more restrictive condition. To this end, we first introduce two definitions to understand the tree estimator, k-potential nearest neighbors (k-PNN) and PNN k-set [ref]. Consider a set of sample points $\mathbf{X}_1, \ldots, \mathbf{X}_s \in \mathbb{R}^d$ and a fixed $\mathbf{x} \in \mathbb{R}^d$. In Lin and Jeon 2006 [ref], they define a point $\mathbf{X}_i$ is a potential nearest neighbor (PNN) of $\mathbf{x}$ if an axis-aligned hyperrectangle defined by vertices x and $\mathbf{X}_i$ contains no other points $\mathbf{X}_j,\ j \neq x$. By extending this notion to a set of PNN point, the definition of a PNN k-set is as follow.

Definition 10. *A PNN k-set of* $\mathbf{x}$ *is a set of points* $\Lambda \subseteq \{\mathbf{X}_1, \ldots, \mathbf{X}_s\}$ *such that there exists an axis aligned hyperrectangle L, with size* $k \leq |L| < 2k-1$, *containing* $\mathbf{x}, \Lambda$, *and no other points.*

From definition 10, we could formalize a definition of k-PNN of $\mathbf{x}$ to help understanding the estimation made by the data-driven nonpararmetric tree method. Upon defining k-PNN, any decision tree that makes axis-aligned split and has leaf size between k and $2k-1$ can be viewed as a k-PNN predictor, as described next.

Definition 11. *A sample point $\mathbf{X}_i$ is called a k-PNN of $\mathbf{x}$ if there exists a PNN k-set of $\mathbf{x}$ containing $\mathbf{X}_i$. A k-PNN predictor $T\left(\mathbf{x};Z_{i1},\ldots,Z_{is}\right)$ at $\mathbf{x}\in\mathbb{R}^d$, where*

$$\{Z_{i1},\ldots,Z_{is}\}=\{(\mathbf{X}_1,Y_1),\ldots,(\mathbf{X}_s,Y_s)\}\in\left\{\mathbb{R}^d\times\mathscr{Y}\right\}^s,$$

is a predictor T always outputs the average of y_i over a k-PNN set of $\mathbf{x}$.

Prediction made by k-PNN predictor in our between-subject random forests case can be written as:

$$T^b\left(\mathbf{x};Z_1,...,Z_s\right)=\sum_{j=1}^{s_0}\sum_{l=1}^{s_1}S_{0j}S_{1l}g\left(Y_{0j},Y_{1l}\right),$$
$$g\left(Y_{0j},Y_{1l}\right)=I\left(Y_{0j}\leq Y_{1l}\right),$$

where

$$S_{0j}=\begin{cases}\left|\left\{j:X_{0j}\in L_0\left(\mathbf{x};Z\right)\text{ and }R_{0j}=0\right\}\right|^{-1} & \text{if } X_{0j}\in L_0\left(\mathbf{x};Z\right),\\ 0 & \text{else,}\end{cases}$$
$$S_{1l}=\begin{cases}\left|\left\{l:X_{1l}\in L_1\left(\mathbf{x};Z\right)\text{ and }R_{1l}=1\right\}\right|^{-1} & \text{if } X_{1l}\in L_1\left(\mathbf{x};Z\right),\\ 0 & \text{else ,}\end{cases}$$

and $R_{0j(1l)}$ is the treatment indicator, $L_0\left(\mathbf{x};Z\right)=\left\{X_{0j}:X_{0j}\in L\left(\mathbf{x};Z\right)\text{ and }R_{0j}=0\right\}$ and $L_1\left(\mathbf{x};Z\right)=\left\{X_{1l}:X_{1l}\in L\left(\mathbf{x};Z\right)\text{ and }R_{1l}=1\right\}$.

Finally, the asymptotic normality result is summarized in Theorem 12 below, with a detailed proof by leveraging the Lyapunov central limit theorem in Appendix.

Theorem 12. *Assume that $\mathbf{x}_1,...,\mathbf{x}_s$ are independent and identically distributed samples on $[0,1]^d$ with a density $f<\infty$. Suppose that $E\left[g\left(Y_{0j},Y_{1l}\right)\mid X_{0j}=\mathbf{x},X_{1l}=\mathbf{x}\right]$ and*

$E\left[g\left(Y_{0j},Y_{1l}\right)^2 \mid X_{0j}=\mathbf{x},X_{1l}=\mathbf{x}\right]$ *are Lipschitz. Let T be a symmetric k-PNN predictor that satisfies Definition 4, 5, 6, 7. Suppose, moreover, that the subsample size s_n satisfies*

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

and that $E\left\{\left|g\left(Y_{0j},Y_{1l}\right)-E\left[g\left(Y_{0j},Y_{1l}\right)\mid X_{0j},X_{1l}\right]\right|^{2+\delta} \mid X_{0j}=X_{1l}=\mathbf{x}\right\} \leq M$ *for some constants* $\delta, M > 0$*, uniformly over all* $\mathbf{x} \in [0,1]^d$ *and* $\operatorname{Var}\left[\sum_{i_1=1}^{n_1} S_{1i_1} g(Y_{0i_0},Y_{1i_1}) \mid X_{0i_0}=\boldsymbol{x}\right] > 0$*. Then, there exists a sequence* $\sigma_n(\mathbf{x}) \to 0$ *such that the between-subject random forests predictions are asymptotically Normal:*

$$\frac{\hat{\delta}(\mathbf{x})-\delta(\mathbf{x})}{\sigma_n(\mathbf{x})} \Rightarrow N(0,1)$$

where $N(0,1)$ is the standard normal distribution.

4.5 Simulation Study

To study the performance of our between-subject random forest estimator, we generate data from the following setup for the potential outcome, confounder and treatment assignment mechanism:

$$\begin{aligned}
Y_i^{(0)} &= \varepsilon_i, \quad \varepsilon_i \sim N(0,1), \quad i=1,...,500,\\
Y_j^{(1)} &= X_{1j}+\varepsilon_i \quad \varepsilon_i \sim N(0,1), \quad j=1,...,500,\\
X_i &= (X_{1i},...,X_{pi}), \quad X_{1i},...,X_{pi} \sim N(0,1),\\
X_j &= (X_{1j},...,X_{pj}), \quad X_{1j},...,X_{pj} \sim N(0,1), \quad p=10,\\
\delta(x) &= \Pr\left[Y_i^{(0)} \leq Y_j^{(1)} | X_i = X_j = x\mathbf{e}_1\right] = \Pr\left[N(x,2) \leq 0\right].
\end{aligned}$$

Figure 4.1 shows the pointwise estimation for Mann-Whitney-Wilcoxon type of causal effects defined in (4.1) by using between-subject random forests method discussed in Section 4.2. The estimation has the same linear increasing pattern and a similar slope as the true causal

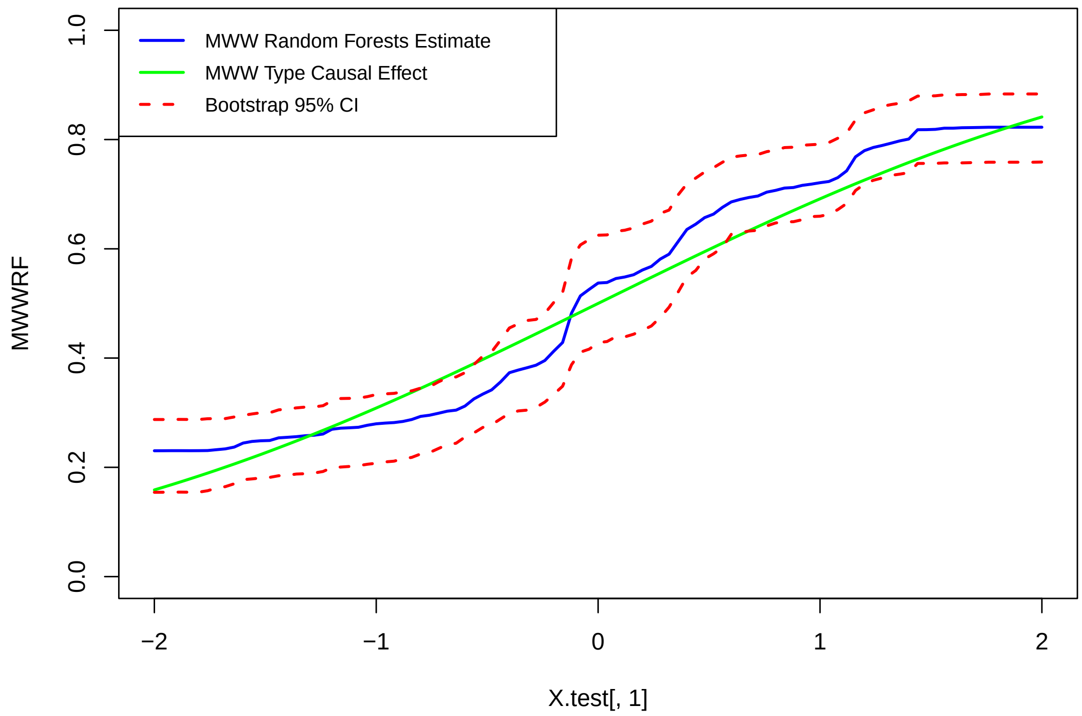


Figure 4.1. Between-subject random forests estimates of the Mann-Whitney-Wilcoxon type of causal effects based on simulated data

effect. The 95% confidence band is computed by Bootstrap, which covers the true MWW type of causal effect.

Chapter 4, in full, is currently being prepared for submission for publication of the material as it may appear in *Tuo Lin, Tsungchin Wu, Xinlian Zhang, and Xin Tu. Non-parametric Causal Inference for Mann-Whitney-Wilcoxon Rank Sum Test Using Random Forest*. The dissertation author was the primary author of this paper.

Chapter 5

Peak *p*-values for Gaussian Random Fields on a Lattice

5.1 Introduction

Statistical parametric mapping (SPM) is widely used as a tool to conduct statistical inference on neuroimaging data [43, 102, 103]. Recently, [37, 36] investigated the validity of cluster size and voxelwise inference based on random field theory (RFT) and found that a number of the assumptions that have been traditionally made do not hold in practice. One of these important assumptions, which we address in this work, is that the data is sufficiently smooth so that it can be treated as a continuous random field. Inference based on peaks or local maxima, recognized as topological features of the statistical summary maps [26, 25, 43, 79, 23] strongly relies on this assumption. In this paper, we circumvent this assumption and develop a method for performing peak inference that is valid for data observed on a regular lattice.

The traditional approach to obtaining peak p-values in fMRI analysis has been to assume that the data is distributed as a smooth stationary Gaussian random field. Given this, [69, 2, 21] showed that the distribution of the height of peaks above a peak-defining threshold $u \in \mathbb{R}$ is asymptotically exponential (as $u \to \infty$). The choice of u is somewhat arbitrary and this result only holds in practice for reasonably large choices of u. Recently, [21] obtained more general formula to calculate the exact height distribution of local maxima in an isotropic Gaussian random field, that is valid for all peak heights and does not require a pre-threshold. This distribution can be used to compute a p-value at each local maximum based on its height. The formula has a single parameter κ, which only depends on the shape of the auto-correlation function near the origin, and is invariant under spatial scaling. While elegant, the formula is only accurate when the Gaussian random field is sampled on a continuous domain, instead of a discrete lattice grid, which in practice can require a high level of applied smoothing. To give context, [80] suggest that this formula is imprecise when data has an intrinsic FWHM that is lower than 7 voxels. However, since the typical smoothing kernel in an fMRI study has an FWHM of 3 voxels, using this formula provides conservative p-values in practice. Moreover, the isotropic assumption is rather strong and is unlikely to hold in practice. Thus it is desirable to directly calculate the height distribution of local maxima sampled on a discrete lattice, which we shall refer to as discrete local maxima (DLM).

In order to address the difference between a discrete lattice and a smooth random field, [101] and [87] introduced a method that targets the distribution of the global maximum on a lattice in order to provide control of the voxelwise family-wise error rate. Although this method aims to infer on the global maximum, it can also be used, after some modifications which we develop here, to compute the height distribution of local maxima. However, their approach is limited in that it is only valid for a narrow class of Gaussian random fields, namely the ones that arise as the result of convolving Gaussian white noise with a separable kernel. In addition, they require local maxima to be defined as those voxels with height values larger than its immediate neighbors along the coordinate axes (i.e. excluding diagonally adjacent neighbors). Figure 5.1 gives a rough idea of why this assumption is restrictive in practice by comparing density plots from peak height distributions calculated from both Worsley and Taylor's analytical DLM approach (which we shall refer to as ADLM) and Cheng and Schwartzman's continuous RFT approach.

To address these issues, we propose a simulation-based method called Monte Carlo DLM (MCDLM) that works for any stationary Gaussian random field under arbitrary connectivity (i.e. where local maxima are defined with respect to any desired neighborhood). This improves

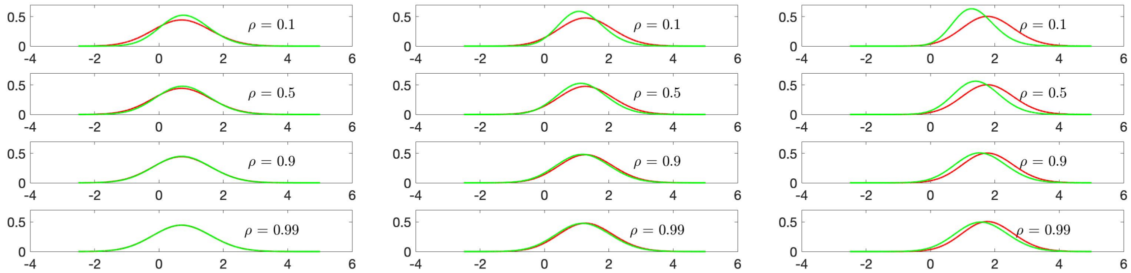


Figure 5.1. Theoretical peak height density function for local maxima. Left: 1D, Middle: 2D, Right: 3D. Each row is calculated using a different correlation between adjacent voxels. In each plot the green curve is from ADLM and the red curve is from the continuous RFT method. In the 1D case, when ρ is small, the ADLM density is narrower, but as ρ increases, the discrepancy disappears and the two methods converge. In 2D and 3D, the differences between the two methods remain for high ρ, with the ADLM density shifted to the left. This occurs because ADLM does not consider diagonal voxels as neighbors, so distribution of local maxima obtained from this method consists of smaller height values and is thus left shifted relative to the continuous method as dimension increases. Note that in 1D there are no diagonal voxels and so convergence occurs.

upon ADLM in that it allows the diagonal neighbors to be considered and makes relatively few assumptions, allowing the accurate computation of the height distribution of local maxima on a lattice. Our approach works by calculating (either theoretically or via empirical estimation) the joint covariance of a voxel and its neighbors, and then simulating many times from a multivariate Gaussian distribution with this covariance and storing the iterations for which the voxel takes a value that is larger than its neighbors. This provides an empirical cdf for the height of local maxima via numerical integration. A p-value for an observed peak in data can be obtained by interpolating the cdf. We also extend this approach to calculate the height distribution of local maxima of t-fields, by generating the simulations from a multivariate t-distribution.

For practical purposes, we recommend using MCDLM to calculate the p-values for peaks in a stationary Gaussian random field that is observed on a lattice. When it is safe to assume that the random field has arisen by convolving Gaussian white noise with a separable kernel and if we only consider partial connectivity, ADLM may also be a good choice to calculate the peak p-value. Alternatively if the field is smooth sufficiently and isotropic, then we recommend using the method discussed in [21] to calculate the peak p-values as it provides a precise formula for the peak height distribution which can be quickly and accurately calculated. In addition, the covariance matrix of a voxel and its neighbors in this case is nearly singular, leading to incorrect p-values when applying estimated covariance in MCDLM.

The structure of this paper is as follows. Section 5.2 provides details about how to calculate the peak height distribution using continuous RFT, the ADLM and the MCDLM method. Section 5.3 and 5.4 apply the MCDLM to the isotropic Gaussian random fields, t-fields and stationary Gaussian fields with unknown covariance and compares its performance with the other two methods. Section 5.3 provides the simulation setup and Section 5.4 includes all the simulation results. Section 5.5 gives the concluding remarks. All codes used in this paper are available on GitHub (https://github.com/tuolin123/DLM-Code) and the RFTtoolbox (https://github.com/sjdavenport/RFTtoolbox).

5.2 Theory and methods for calculating the height distribution of local maxima

Let $\{Z(s), s \in S\}$ be a real-valued stationary Gaussian random field parametrized on a D-dimensional set S, where $D \in \mathbb{Z}^+$. We assume that S is a regularly spaced discrete lattice, in particular that

$$S \subset \left\{ \sum_{d=1}^{D} v_d e_d : v_d \in \mathbb{Z} \text{ for } 1 \leq d \leq D \right\},$$

where $(e_d)_{1 \leq d \leq D}$ is the standard basis in $\mathbb{R}^D$ and v_d represents the step size in the dth direction. Our interest lies in calculating the peak height distribution, which for $u \in \mathbb{R}$, is defined as

$$P(Z(s) > u | s \text{ is a local maximum}) = P[Z(s) > u | Z(t) < Z(s), \forall t \in \mathscr{N}(s)], \tag{5.1}$$

where $\mathscr{N}(s)$ denotes the set of neighbors of $s \in S$ in the discrete lattice. The most relevant neighborhoods are the partially connected and fully connected ones that are respectively defined as

$$\mathcal{N}_{PC}(s) = \{s + k_d v_d e_d : k_d \in \{-1, 1\} \text{ for } 1 \le d \le D\} \text{ and} \tag{5.2}$$

$$\mathcal{N}_{FC}(s) = \left\{ s + \sum_{d=1}^{D} k_d v_d e_d : k_d \in \{-1, 0, 1\} \text{ for } 1 \le d \le D \right\} \Big\backslash \{s\}. \tag{5.3}$$

Figure 5.2 illustrates the two types of neighborhoods for a point, s_5, on a 2D regular lattice. If s_5 is partially connected to the adjacent pixels in the horizontal and vertical directions, then $\mathcal{N}_{PC}(s_5) = \{s_2, s_4, s_6, s_8\}$, shown in the left of Figure 5.2. If s_5 is fully connected, meaning it is connected to pixels in the horizontal, vertical and diagonal directions, then $\mathcal{N}_{FC}(s_5) = \{s_1, s_2, s_3, s_4, s_6, s_7, s_8, s_9\}$, shown in the right of Figure 5.2.

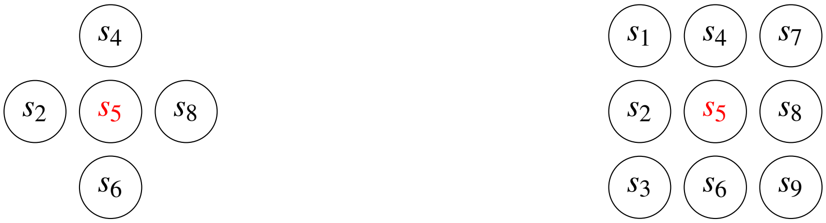


Figure 5.2. Local pixel neighborhood in 2D. The partially and fully connected neighborhoods are shown on the left and right respectively. The point s_5 (colored in red) is considered a local maximum if its value is larger than its neighbors.

5.2.1 Analytical DLM method

The DLM approach of [101] and [87] provides closed form expressions for the familywise error rate in testing the signals of data that consists of Gaussian white noise smoothed with a Gaussian kernel. They do not explicitly focus on the peak height distribution, however in inferring on the global maximum they calculate probabilities of the form

$$P\left[\{Z(s) > u\} \cap_{t \in \mathcal{N}_{PC}(s)} \{Z(t) < Z(s)\}\right].$$

These probabilities can be used to calculate a peak height distribution since (5.1) can be written as

$$P[Z(s) > u | Z(t) < Z(s), \forall t \in \mathscr{N}_{PC}(s)] = \frac{P\left[\{Z(s) > u\} \cap_{t \in \mathscr{N}_{PC}(s)} \{Z(t) < Z(s)\}\right]}{P[Z(t) < Z(s), \forall t \in \mathscr{N}_{PC}(s)]}. \tag{5.4}$$

Using partial results in [101] and [87], we expand the left hand side of (5.4) as $\int_u^\infty f_{\mathrm{DLM}}(z)dz$, where $f_{\mathrm{DLM}}(z)$ is the density function of the height distribution. Under the assumption that the smoothing kernel is Gaussian and $Z(s)$ is locally stationary, $f_{\mathrm{DLM}}(z)$ has the form

$$f_{\mathrm{DLM}}(z) = \frac{\prod_{d=1}^{D} Q(\rho_d, z)\phi(z)}{\int_{-\infty}^{\infty} \left(\prod_{d=1}^{D} Q(\rho_d, z)\right)\phi(z)dz}, \tag{5.5}$$

where

$$h_d = \sqrt{\frac{1-\rho_d}{1+\rho_d}}, \alpha_d = \sin^{-1}\left(\sqrt{(1-\rho_d^2)/2}\right), z^+ = \max(z,0)$$
$$Q(\rho_d, z) = 1 - 2\Phi(h_d z^+) + \frac{1}{\pi}\int_0^{\alpha_d} \exp\left(-\frac{1}{2}h_d^2 z^2 / \sin^2\theta\right) d\theta,$$

and ρ_d is the correlation between two voxels along each axis direction d, given by $\rho_d = \rho(s, s + v_d e_d)$, where $\rho(\cdot,\cdot)$ is introduced in (5.6). This approach also allows for the calculation of the height of local maxima on the boundary of the image or a mask by substituting $Q(\rho_d, z)$ with $1 - \Phi(h_d z)$ if a voxel on the boundary only has one neighbor, and with 1 if it has no neighbors. Further details regarding the derivation of (5.5) are provided in Appendix D.1.2. Since this method provides a closed form density function, we call it analytical DLM (ADLM) method.

One critical assumption of the ADLM approach is that the correlation function has a specific separable structure. Under this assumption things simplify because conditioned on the center voxel, the distribution of the height of neighboring voxels along a given axis are conditionally independent of the distribution of the height at neighboring voxels along the other (perpendicular) axis directions, described in the proposition below.

Proposition 13. *Given data* $\{Z(s), s \in S\}$ *such that the spatial correlation between* s *and* t *is*

$$\rho(s,t) = \exp[-(s-t)'\Lambda(s-t)/2], \tag{5.6}$$

where $\Lambda = \mathrm{diag}(1/(2\eta_1^2), ..., 1/(2\eta_D^2))$*, and* $(\eta_d)_{d=1,...,D}$ *is the standard deviation of the Gaussian kernel in the dth direction. Assume that* $d_1, d_2 \in \{1, ..., D\}$*,* e_d *is the standard basis in* $\mathbb{R}^D$ *and* v_d *represents the step size in the dth direction and* $s \pm v_{d_1}e_{d_1}, s \pm v_{d_2}e_{d_2} \in S$*, then*

$$\begin{pmatrix} Z(s - v_{d_1}e_{d_1}) \\ Z(s + v_{d_1}e_{d_1}) \end{pmatrix} \perp\!\!\!\perp \begin{pmatrix} Z(s - v_{d_2}e_{d_2}) \\ Z(s + v_{d_2}e_{d_2}) \end{pmatrix} \Bigg| Z(s).$$

Correlation function in (5.6) arises, for example, from integration of continuous white noise against a Gaussian kernel. The result in Proposition 13 is stated in [87] and we provide a short proof in Appendix D.1.1. To visualize it more precisely, in Figure 5.2, under the required assumptions, we have

$$\begin{pmatrix} Z(s_4) \\ Z(s_6) \end{pmatrix} \perp\!\!\!\perp \begin{pmatrix} Z(s_2) \\ Z(s_8) \end{pmatrix} \Bigg| Z(s_5).$$

This conditional independence result holds along the horizontal and vertical axes and allows for an expansion for the distribution of partially connected local maxima. However, it does not imply independence when the diagonals are included, i.e.,

$$(Z(s_1), Z(s_3), Z(s_7), Z(s_9))^\top \not\perp\!\!\!\perp (Z(s_2), Z(s_4), Z(s_6), Z(s_8))^\top | Z(s_5).$$

Thus, their method can only be used to calculate the height distribution of peaks that are greater than their directly adjacent neighbors.

Under the stationary assumption, the correlation ρ_d between two adjacent voxels along each lattice axis d can be simplified from (5.6) to

$$\rho_d = \rho(s, s + v_d e_d) = \exp\left[-\frac{1}{2}\left(\frac{v_d^2}{2\eta_d^2}\right)\right] = \exp\left[-\frac{v_d^2}{4\eta_d^2}\right], \tag{5.7}$$

and if we assume $Z(s)$ is isotropic with a common standard deviation of the Gaussian kernel $\eta_d = \eta$ and $v_d = 1$, ρ_d can be further simplified to $\rho_d = \exp[-1/4\eta^2]$, a function that does not depend on d.

The ADLM approach allows the calculation of the height distribution of local maxima on a discrete lattice. However, the method makes restrictive assumptions and its validity is limited to partial connectivity.

5.2.2 The correlation function on the lattice

The methods which we will develop in what follows rely strongly on the correlation function. In this section we provide some explicit expansions of this function under the assumption that the fields are derived by smoothing i.i.d white noise with a kernel (we will relax this assumption later on).

Define the correlation function $\rho(s,t) : S \times S \to \mathbb{R}$ to be the function that maps $s,t \in S$ to $\text{corr}(Z(s), Z(t))$. As a step toward our goal of calculating peak p-values for a Gaussian random field on a regular discrete lattice, we shall calculate the spatial correlation of a special type of Gaussian random field over the lattice analytically. Assume that $W : S \to \mathbb{R}$ is a Gaussian random field consisting of i.i.d. unit variance white noise and for some kernel $K : \mathbb{R}^D \to \mathbb{R}$,

$$Z(s) = \sum_{l \in S} K(s-l)W(l) \quad \text{for each} \quad s \in S.$$

The correlation function $\rho(s,t)$ is then

$$\begin{aligned}
\rho(s,t) &= \frac{E\left[\sum_{l\in S} K(s-l)W(l)\sum_{l'\in S} K(t-l')W(l')\right]}{\sqrt{\operatorname{Var}\left[\sum_{l\in S} K(s-l)W(l)\right]\operatorname{Var}\left[\sum_{l'\in S} K(t-l')W(l')\right]}} \\
&= \frac{E\left[\sum_{l\in S}\sum_{l'\in S} K(s-l)K(t-l')W(l)W(l')\right]}{\operatorname{Var}\left[\sum_{l\in S} K(s-l)W(l)\right]} \\
&= \frac{\sum_{l\in S} K(s-l)K(t-l)}{\sum_{l\in S} K(s-l)^2}
\end{aligned}$$

since $E[W(l)W(l')] = 0$ for $l \neq l'$ and $EW(l)^2 = 1$ for all l. In particular when K is an isotropic Gaussian kernel, i.e., $K(s) = \frac{1}{\eta^D}\phi_D\left(\frac{||s||}{\eta}\right)$, for some $\eta > 0$ and each $s \in S$,

$$\rho(s,t) = \frac{\sum_{l\in S} \frac{1}{\eta^{2D}}\phi_D\left(\frac{||s-l||}{\eta}\right)\phi_D\left(\frac{||t-l||}{\eta}\right)}{\sum_{l\in S} \frac{1}{\eta^{2D}}\left[\phi_D\left(\frac{||s-l||}{\eta}\right)\right]^2}, \tag{5.8}$$

where ϕ_D is the density function for the D dimensional standard Gaussian distribution. As is common in fMRI analysis we will typically refer to this kernel using its full width at half maximum (FWHM) which is defined as $\text{FWHM} = 2\sqrt{2\ln 2}\eta$. Applying (5.8) as the correlation function to ADLM defined in (5.5) improves the performance of ADLM approach by applying correlation (5.6).

More generally if K is an elliptical Gaussian kernel, i.e., $K(s) = \prod_{j=1}^{D}\frac{1}{\eta_j}\phi_1\left(\frac{[s]_j}{\eta_j}\right)$, then

$$\rho(s,t) = \frac{\sum_{l\in S}\prod_{j=1}^{D}\frac{1}{\eta_j^2}\phi_1\left(\frac{[s-l]_j}{\eta_j}\right)\phi_1\left(\frac{[t-l]_j}{\eta_j}\right)}{\sum_{l\in S}\prod_{j=1}^{D}\frac{1}{\eta_j^2}\phi_1\left(\frac{[s-l]_j}{\eta_j}\right)^2}, \tag{5.9}$$

where $[s-w]_j$ and $[s+v-w]_j$ refer to the jth dimension of array $s-w$ and $s+v-w$ respectively.

5.2.3 Monte Carlo DLM method

In this section we introduce a new method based on Monte Carlo simulation to calculate the height distribution of local maxima on a discrete lattice. Our approach is based on the observation that the probability that $s \in S$ is a local maximum based entirely on the distribution of s with its neighbors as in (5.1). Define

$$\mathbf{Z}(s) = (Z(s), Z(n_1(s)), ..., Z(n_k(s)))^\top,$$

where we have enumerated the neighborhood of s as $\mathscr{N}(s) = \{n_1(s), \ldots, n_k(s)\}$ for some $k \in \mathbb{N}$. $\mathscr{N}(s)$ can be either $\mathscr{N}_{PC}(s)$ or $\mathscr{N}_{FC}(s)$. In the case of the partially connected neighborhood $k = 2D$ and for the fully connected neighborhood, $k = 3^D - 1$. Under the stationarity assumption $\mathbf{Z}(s) \sim N(\mathbf{0}, \Sigma)$ for each s, where $\Sigma = \mathrm{Cov}(\mathbf{Z}(s))$ is constant over the domain. The covariance matrix Σ can be derived analytically under certain assumptions or estimated from the data, see below. Given Σ, the method calculates the peak height distribution via Monte Carlo simulation as described in Algorithm 2.

Algorithm 2. MCDLM

```
Require: The number of iterations $M \in \mathbb{N}$, and the covariance matrix $\Sigma$
procedure SIMULATELOCMAX($M, \Sigma$)
    for $m = 1, \ldots, M$ do
        Generate $X_m \sim N(0, \Sigma)$
        if $X_{m1} > \max_{2 \le j \le k} X_{mj}$ then
            $h = [h, X_{m1}]$
        end if
    end for
    return $h$
end procedure
```

After obtaining the vector $h = (h_1, h_2, ..., h_N)^\top$, for $u \in \mathbb{R}$, the MCDLM approximation to the peak height distribution can be calculated as

$$\hat{F}_N(u) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}\{h_i \leq u\},$$

where $\mathbb{1}\{\cdot\}$ denotes the indicator function and N is the length of h. For an observed peak of height u, a peak p-value can be computed as $1 - \hat{F}_N(u)$. In order to make this empirical p-value as accurate as possible, N should be taken to be as large as possible. In our examples we choose M large enough to ensure that the number of resulting empirical samples is at least $N = 10^6$ for FWHM < 11.7 and at least $N = 2 \times 10^5$ for FWHM $= 11.7$.

Now we address the question of how to calculate Σ. We will first do this in a specific parametric case discussed in the previous section of smoothing i.i.d Gaussian white noise with an isotropic Gaussian kernel. In that case, assuming a fully connected neighborhood, it can be shown that, up to reindexing of $\mathbf{Z}$,

$$\Sigma = \underbrace{A \otimes A \otimes \ldots \otimes A}_{D \text{ terms of } A} = A^{\otimes D} \tag{5.10}$$

where

$$A = \begin{pmatrix} 1 & \rho & \rho^4 \\ \rho & 1 & \rho \\ \rho^4 & \rho & 1 \end{pmatrix}.$$

with ρ the correlation between adjacent voxels as defined in (5.7). The proof of (5.10) is in Appendix D.2.

Equation (5.10) holds under restrictive assumptions. For a general stationary field we can instead use the data to estimate Σ. To do so we first center and standardize i.i.d Gaussian random fields $Z_1, \ldots, Z_n$ on S at every $s \in S$. We wish to infer on the distribution of peaks of the mean $\frac{1}{n}\sum_{i=1}^{n} Z_n$. Next we estimate the Σ from the data as follows. For each $s,t \in S$,

$$\text{Cov}\left(\frac{1}{n}\sum_{i=1}^{n} Z_i(s), \frac{1}{n}\sum_{i=1}^{n} Z_i(t)\right) = \frac{1}{n}\text{Cov}(Z_1(s), Z_1(t)), \tag{5.11}$$

as such, using the assumption of stationarity, we can estimate this covariance as

$$\widehat{\text{Cov}}(Z_1(s), Z_1(t)) = \frac{1}{n|L|}\sum_{i=1}^{n}\sum_{(s',t')\in L} Z_i(s')Z_i(t') \tag{5.12}$$

where $L = \{(s',t') \in S \times S : s' - t' = s - t\}$. If we assume that the fields are isotropic, then we can improve the accuracy of this estimate by taking $L = \{(s',t') \in S \times S : ||s' - t'|| = ||s - t||\}$.

5.2.4 Continuous Gaussian random field theory method

Historically [26, 25, 80] it has been common to use the results of continuous Gaussian random fields [2] to perform inference on the lattice. [21] derived a closed form for the distribution of the height of a local maximum. We briefly outline how this works in what follows and explain how it can be used to provide peak height distributions for local maxima in a sufficiently smooth random field. To do so, we now assume that $S \subset \mathbb{R}^D$ is a continuous domain, i.e., compact with non-empty interior $\mathring{S}$. Suppose that Z is a C^3 random field on S and let

$$\begin{aligned} \nabla Z(s) &= \left(\frac{\partial Z(s)}{\partial s_1}, \ldots, \frac{\partial Z(s)}{\partial s_D}\right), \\ \nabla^2 Z(s) &= \left(\frac{\partial Z(s)}{\partial s_{ij}}\right)_{1\leq i,j\leq D}. \end{aligned}$$

Then the local maxima of Z are defined to be the points $s \in \mathring{S}$ such that $\nabla Z(s) = 0$ and $\nabla^2 Z(s) < 0$.

The event that a local maximum is observed at a given $s \in \mathring{S}$ has probability zero. As such, in order to obtain a conditional peak height distribution, Palm distributions must be used (see [21] for details). For $u \in \mathbb{R}$, they provide formulae to calculate

$$\mathbb{P}[Z(s) > u \mid \nabla Z(s) = 0 \text{ and } \nabla^2 Z(s) < 0]. \tag{5.13}$$

In general these expressions are difficult to evaluate. However under the assumption of isotropy, [22] showed that they can be obtained explicitly. Recently, [24] extended these results to the case where the field arises as a diffeomorphic transformation of an isotropic field. Details of how to apply these methods to perform peak inference in fMRI data can be found in [80].

5.2.5 MCDLM for t-fields

We can also use our MCDLM approach to calculate the height distribution of local maxima of a t-field. The t-fields are generated by voxelwise calculation of t-statistic using

$$T(s) = \frac{\varepsilon(s)}{\sqrt{\sum_{i=1}^{N} Z_i^2/N}}, \tag{5.14}$$

where $Z_1, ..., Z_N$ and $\varepsilon(s)$ are i.i.d isotropic Gaussian fields. In practice, the t-statistic is typically used as the test statistic for regression coefficients.

In this setting the local neighborhood has a multivariate t-distribution. Thus, if we know the estimate of the neighborhood covariance we can again apply MCDLM algorithm 2 by changing the simulation from a multivariate Gaussian distribution to a multivariate t-distribution.

This approach works well in practice (see Section 5.4.2) however it is somewhat computational (especially as ρ and the degrees of freedom increase). To get around this we consider a voxelwise Gaussianization transformation of the t-fields (as in [80]) which acts using the distribution function as follows:

$$Z(s) = -\Phi^{-1}[F_{t,\nu}(-T(s))], \tag{5.15}$$

where $F_{t,\nu}$ is the cdf of the t-distribution with ν degrees of freedom. We then apply the MCDLM method for the Gaussian field to the transformed t-field. As we will see in the simulations in the next two sections, the changes in the covariance matrix do not influence the results much.

5.3 Simulation Setup

In this section we describe the different simulation settings we have considered in order to compare the performance of the three methods introduced in Section 5.2, i.e. the ADLM, MCDLM and continuous RFT method. For each simulation setting considered we generate a large number of stationary Gaussian random fields (or t-fields), collect the heights of the peaks across all fields and combine these to obtain a reference peak height distribution, which will allow us to test the validity of each of the approaches.

For each method, we calculate a p-value at each peak. Local maxima were selected based on the criteria that their height values are larger than their neighbors - we separately consider both the fully connected and partially connected neighborhoods. We compare the validity and accuracy of these p-values using *pp* plots which compare the sorted p-values to the tail probability of the true peak height distribution. These are formally defined in Appendix D.3. The closer each plot is to the identity function, the closer the approximation is to the true distribution. The plots lying below the identity function correspond to conservative p-values and plots above the identity function correspond to liberal p-values. We use these *pp* plots to compare the performance of the three approaches in all of our simulation studies.

5.3.1 Isotropic Gaussian random fields

Our first set of simulations consists of isotropic Gaussian fields which are obtained by convolving Gaussian white noise with a Gaussian kernel with specified FWHM and normalizing so that the resulting fields have unit variance. To avoid any boundary effects the fields were initially generated on a D-dimensional large grid of size $(50+2\times\lceil 4*\eta\rceil)$ at each direction and the central subset is taken, as described in [33]. We do this in 2D and in 3D. The resulting 2D images are of size 50×50 and the resulting 3D images are of size $50\times 50\times 50$. In the simulations we choose the FWHM so that the correlation between adjacent voxels in each perpendicular direction is equal to $\rho \in \{0.01, 0.5, 0.99\}$ - this corresponds to FWHMs of 0.7, 1.5 and 11.7 voxels. In each setting we generate 10,000 random fields and compare the different approaches using *pp* plots - as described above and in Appendix D.3. See Section 5.4.1 for the corresponding results.

5.3.2 Isotropic t-fields

In this section, we consider the performance of the different approaches when it comes to evaluating the height distribution of peaks of t-fields. To do so we generate fields with ν degrees of freedom (taking $\nu = 20, 50, 100$) by simulating i.i.d isotropic Gaussian fields $Z_1, ..., Z_\nu$ and $\varepsilon(s)$ in (5.14) as in Section 5.3.1. We generate 10,000 t-fields in both 2D and 3D. In each setting we calculate peak-height p-values using the MCDLM approach for t-fields discussed in Section 5.2.4. We also calculate p-values using the continuous RFT approach. Note that this is designed for Gaussian random fields so we would not expect it to work as well in this setting. We compute a final set of p-values using the actual simulated Gaussian random fields - as the number of degrees of freedom of the t-field goes to infinity the t-field converges to a Gaussian random field with the original covariance function so this provides a measure of the convergence of the fields. Note that in practice this is not a viable measure for calculating a peak height distribution as it requires us to have the original fields. We compare the p-values obtained using these different approaches using *pp*-plots. The results are described in Section 5.4.2.

As discussed in Section 5.2.5, we also consider using Gaussianization transformation of the t-field to improve the computational efficiency. We perform the same set of simulations but where each of the t-fields is Gaussianized, with the same t-field as described above. We then calculate p-values using the MCDLM for Gaussian fields and the continuous RFT approach. The results are described in Appendix D.4.

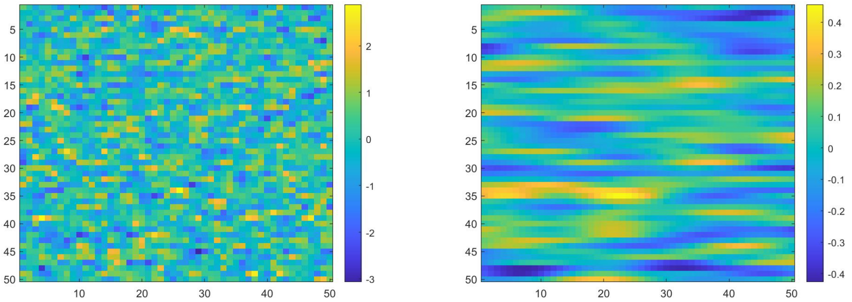


Figure 5.3. Examples of stationary Gaussian random fields which are obtained by convolving white noise with an elliptical Gaussian kernel with $\rho_1 = 0.01, \rho_2 = 0.5$ in the left plot and $\rho_1 = 0.5, \rho_2 = 0.99$ in the right plot.

5.3.3 Stationary Gaussian fields with unknown covariance

To test the performance of MCDLM, we consider two different simulation settings. In the first we use n fields ($n = 20, 50, 100, 200$ for small ρ and $n = 20, 50, 100, 200, 1000$ for large ρ) to estimate the neighborhood covariance using the isotropic version of equation (5.12). We compare the performance of MCDLM with this estimated covariance across different sample sizes (once again using 10,000 simulations in each of the settings described in Section 5.3.1). In the second we consider 2D non-isotropic Gaussian fields. To generate these we smooth Gaussian white noise with an elliptical Gaussian kernel with smoothing FWHM in each direction chosen such that the correlation between adjacent voxels in the vertical and horizontal directions is ρ_1 and ρ_2 respectively. Again, we estimate the neighborhood covariance using equation (5.9) using n fields ($n = 20, 50, 100, 200$). We consider two 2D examples, one where $\rho_1 = 0.01$ and $\rho_2 = 0.5$ and a second where $\rho_1 = 0.5$ and $\rho_2 = 0.99$ (example realizations of these fields are illustrated in Figure 5.3). In each setting We compare the p-values obtained using different sample size and using a theoretical covariance function by pp plots. The results are shown in Section 5.4.3.

In practice the spatial covariance of the observed random field is unknown and it may not be reasonable to assume the field is isotropic. Thus Σ must be estimated from the data as described in Section 5.2.3. Once the covariance function has been estimated we must ensure that it is positive semi-definite (p.s.d). We proceed by pushing the negative eigenvalues of estimated covariance matrix to a small positive value, 1×10^{-10}. In a stationary Gaussian field, there is a lot of structure that can be taken advantage of when estimating Σ. Under stationarity the neighborhood covariance matrix has a block Toeplitz structure which makes it easier to estimate, see examples in Appendix D.2.

5.4 Simulation Results

5.4.1 Results for isotropic Gaussian random fields

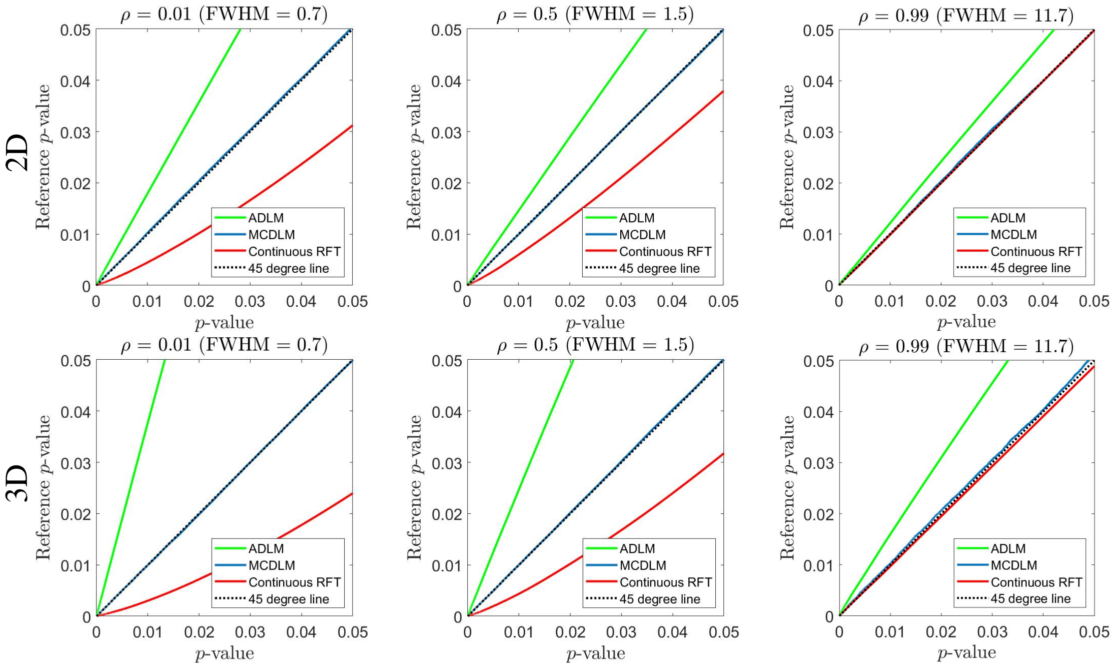


Figure 5.4. *pp* plots which compare the different methods of computing peak height p-values in the isotropic Gaussian random field scenario. 2D and 3D results are displayed in the first and second rows respectively. The correlations between adjacent voxels are $\rho = 0.01, 0.5, 0.99$. The plots compare the performance of ADLM, MCDLM and the continuous RFT approach.

Results comparing all three methods in the isotropic Gaussian random fields setting (described in Section 5.3.1) are presented in Figure 5.4. From this figure we see that the MCDLM method obtains p-values which are uniformly distributed and thus provides accurate and valid inference at all smoothness levels. The continuous approach is valid but conservative unless the data is very smooth (large FWHM). The ADLM method gives liberal p-values at all smoothness levels though the severity of this reduces when the smoothness is very large. In 3D the results are similar though at the highest smoothness level the curve corresponding to the MCDLM method is slightly rough, this could be made sharper if desired by increasing the number of Monte Carlo runs used. As discussed in Section 5.2, we focus on the covariance function calculated by (5.8) in the calculation of the MCDLM and ADLM distributions since it is the actual covariance of the data. The results are slightly worse if (5.7) is used instead (as was done in [101, 87]). The results for this are discussed in Appendix D.4.2. We also generate a look-up table, which provides the same results under reduced computation time (for the case of Gaussian white noise smoothed with an isotropic Gaussian kernel). The results of using the lookup table are shown in Appendix D.4.3. Here we only consider the case where the neighborhood is fully connected, results for partially connected neighborhoods (in which the ADLM method performs much better than MCDLM method) are presented in Appendix D.4.1.

To quantify the difference of p-values from all three methods, we use the Root Mean Squared Error (RMSE) to calculate the difference between p-value from each of the method and 45 degree line. Our comparison focus on the region of p-value ≤ 0.05. The RMSE results calculated from 2D isotropic Gaussian random fields with different ρ are shown in Table 5.1. From the table, MCDLM performs better than ADLM and continuous RFT approach in all low smoothness cases (FWHM ≤ 5.2). The MCDLM is outperformed by continuous RFT approach when smoothness level is high (FWHM $= 8.3$ and 11.7).

5.4.2 Isotropic t-fields

The *pp* plots comparing the different methods in the setting of isotropic t-fields (described in Section 5.3.2) are presented in Figure 5.5 (2D) and Figure 5.6 (3D). As shown in these two figures, MCDLM obtains p-values which are uniformly distributed at all smoothness levels and different degrees of freedom for 2D and lower smoothness levels for 3D. In the 3D case at high smoothness levels, the MCDLM approach becomes a rough approximation because the number of peaks generated is not sufficient. As such the height distribution computed is inaccurate,

Table 5.1. RMSE results from 2D isotropic Gaussian random fields for comparing the p-values from MCDLM, ADLM and continuous RFT methods. The smallest value in each row is highlighted in red color.

	MCDLM	ADLM	Continuous RFT
$\rho = 0.01$ (FWHM $= 0.7$)	1.71×10^{-4}	5.77×10^{-3}	8.18×10^{-3}
$\rho = 0.1$ (FWHM $= 1$)	1.91×10^{-4}	5.48×10^{-3}	7.75×10^{-3}
$\rho = 0.3$ (FWHM $= 1.2$)	6.16×10^{-5}	4.83×10^{-3}	6.53×10^{-3}
$\rho = 0.5$ (FWHM $= 1.5$)	6.34×10^{-5}	4.16×10^{-3}	4.99×10^{-3}
$\rho = 0.7$ (FWHM $= 2$)	1.22×10^{-4}	3.58×10^{-3}	3.01×10^{-3}
$\rho = 0.9$ (FWHM $= 3.6$)	2.05×10^{-4}	2.91×10^{-3}	8.61×10^{-4}
$\rho = 0.95$ (FWHM $= 5.2$)	1.12×10^{-4}	2.56×10^{-3}	5.04×10^{-4}
$\rho = 0.98$ (FWHM $= 8.3$)	1.57×10^{-4}	2.51×10^{-3}	1.02×10^{-4}
$\rho = 0.99$ (FWHM $= 11.7$)	2.87×10^{-4}	2.54×10^{-3}	1.29×10^{-4}

as shown in the noisy subfigure in the bottom right of Figure 5.6. The continuous method is designed for Gaussian fields rather than t-fields, so it is too liberal when ν is small while too conservative when ν is large and ρ is small. Although in the case that the number of degrees of freedom is large and FWHM large, the Gaussian field can approximate the t-field and continuous method has improved performance, the MCDLM method always outperforms the continuous method in all circumstances as long as we generate enough local maxima.

As discussed in Section 5.3.2, the Gaussianization approach is introduced to save computation time. The results for this are shown in Appendix D.4. They show that the MCDLM method performs well when the number of degrees of freedom is large whereas continuous method requires both the degrees of freedom and FWHM to be large.

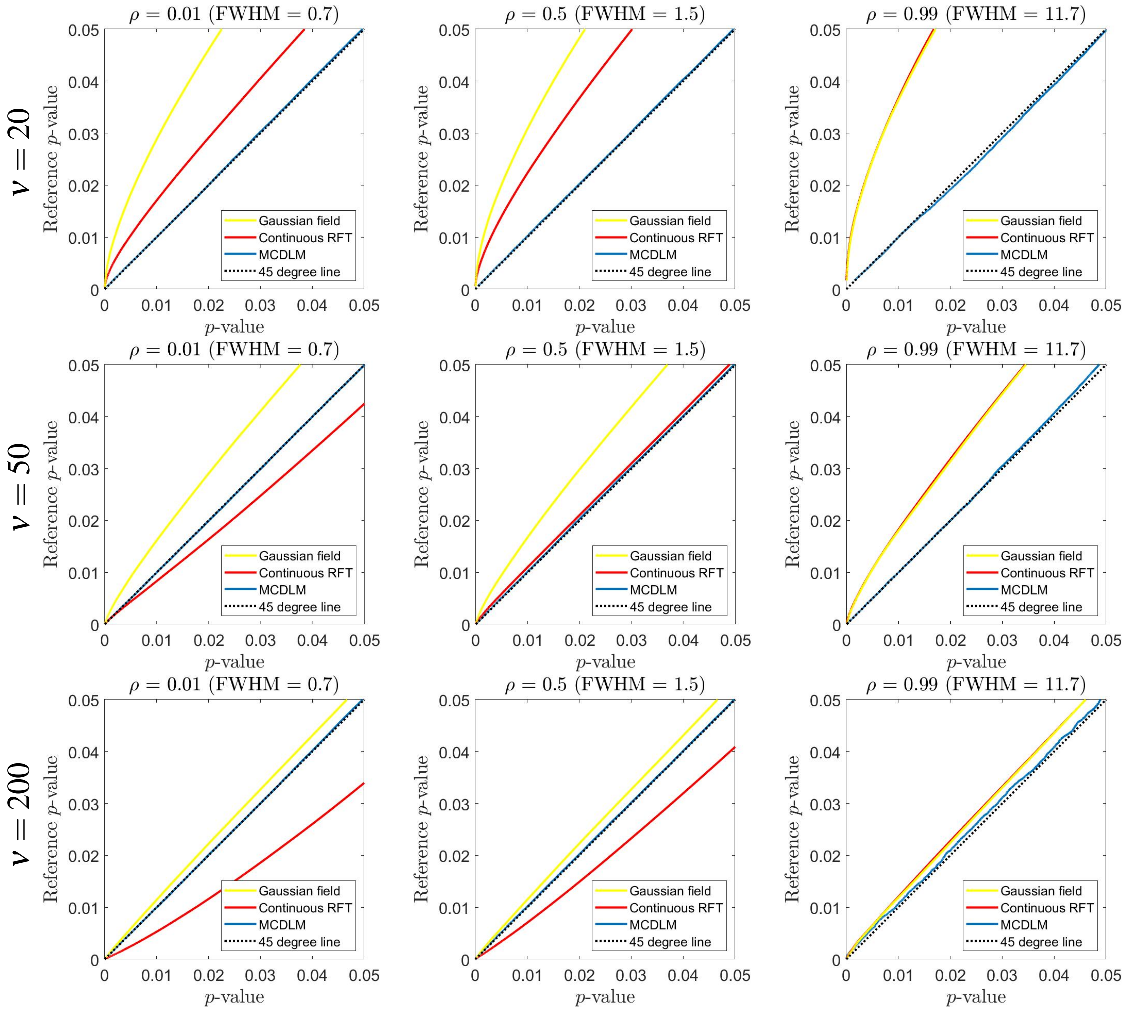


Figure 5.5. Comparing methods for calculating the peak height distribution of a 2D t-field with ν degrees of freedom. From left to right: spatial correlation $\rho = 0.01, 0.5, 0.99$. From top to bottom: $\nu = 20, 50, 200$. The figure is generated based on the comparison of the p-values calculated using Gaussian field with same correlation ρ, continuous RFT and MCDLM approach. The reference is the true peak height distribution generated from the t field.

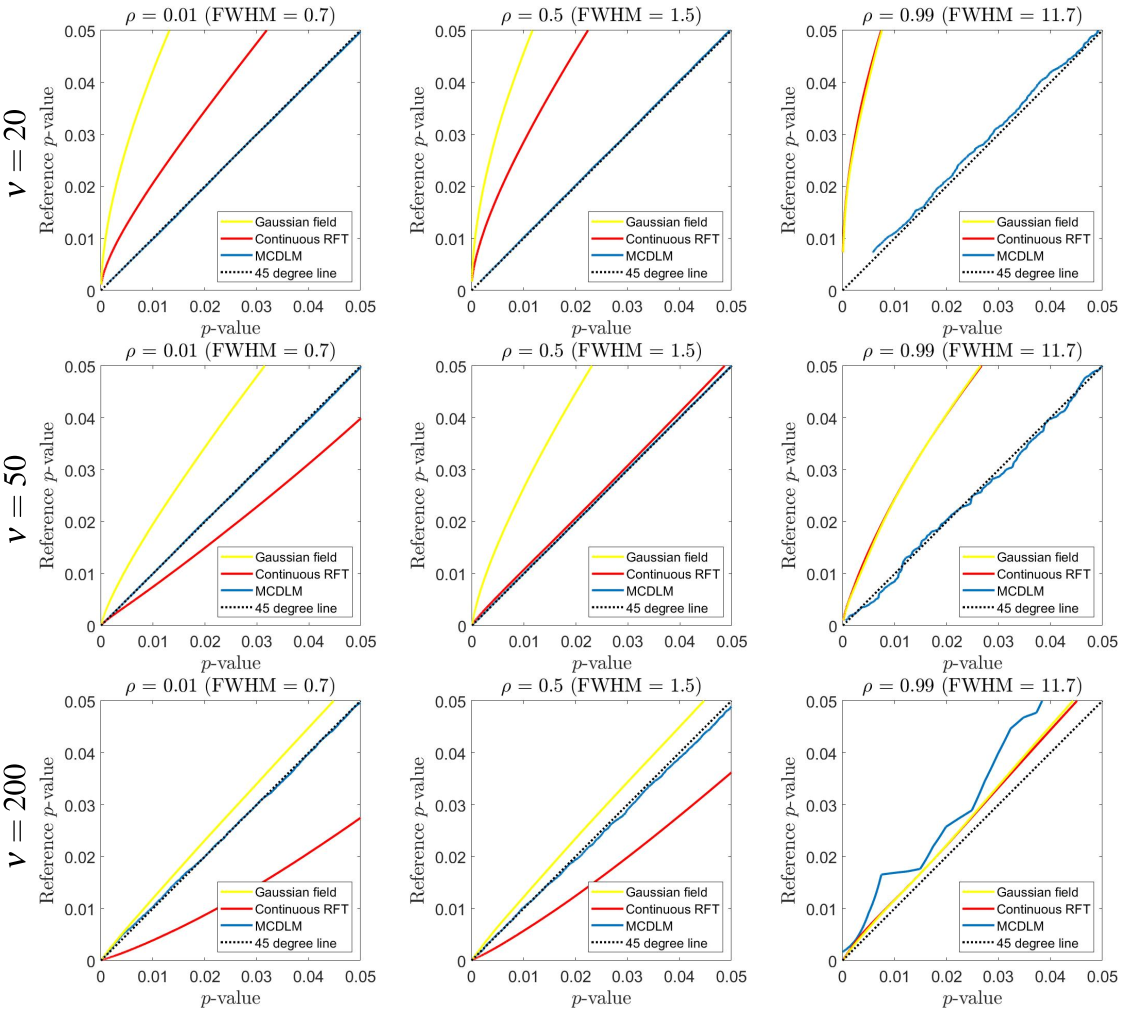


Figure 5.6. Comparing methods for calculating the peak height distribution of a 3D t-field with ν degrees of freedom. From left to right: spatial correlation $\rho = 0.01, 0.5, 0.99$. From top to bottom: $\nu = 20, 50, 200$. The figure is generated based on the comparison of the p-values calculated using Gaussian field with same correlation ρ, continuous RFT and MCDLM approach. The reference is the true peak height distribution generated from the t field.

5.4.3 Stationary Gaussian fields with unknown covariance

In Figure 5.7, we compare the peak height distribution calculated from the MCDLM method using both theoretical (i.e. true) neighborhood covariance in (5.8) and estimated neighborhood covariance displayed in (5.12). Figure 5.7 shows that MCDLM with the estimated neighborhood covariance performs as well as MCDLM with the theoretical neighborhood covariance when $\rho = 0.01$ and 0.5. When ρ increases to 0.99, the MCDLM method with estimated covariance function requires a large number of simulated peaks before it converges. This number decreases with the number of voxels in the image (which is why the performance of the 3D simulations are substantially better than the 2D ones), even when a very large sample size is used to estimate it. Since with $\rho = 0.99$, even with 1000 instances to estimate the neighborhood covariance the MCDLM method still performs poorly, we investigate more scenarios with $\rho = 0.9, 0.93, 0.95$. The detailed results are included in Appendix D.4.5. Based on the results, we recommend using the MCDLM method with estimated covariance function when $\rho < 0.95$, or FWHM < 5.2 in practice.

The results for the second (non-isotropic) simulations discussed in Section 5.3.3 are shown in Figure 5.8. From this plot we see that the estimated version works well when $\rho_1 = 0.01$ and $\rho_2 = 0.5$, and requires a larger number of realizations to converge when $\rho_1 = 0.5$ and $\rho_2 = 0.99$.

5.5 Discussion

In this paper, we have proposed a new Monte Carlo method to calculate the distribution of the height of a peak of a discrete Gaussian random field which works under minimal assumptions. When inferring on the heights of the peaks of Gaussian field MCDLM performed well compared to other approaches. Historically, continuous RFT method was used to calculate the distribution of the height of local maxima in a continuous random field. However, in practice we observe data on a lattice. As shown in [80], when the data is sufficiently smooth (FWHM ≥ 7), the

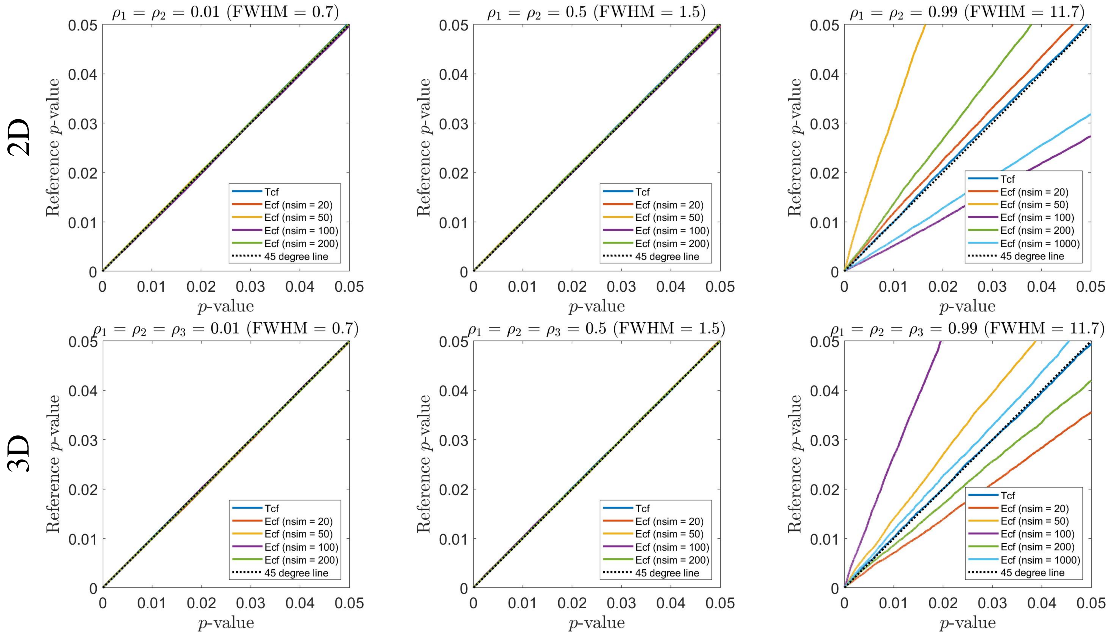


Figure 5.7. Comparison of the peak height distribution calculated from using MCDLM with different neighborhood covariance for 2D and 3D isotropic Gaussian fields. The covariance functions used here are theoretical covariance function (Tcf) and empirically estimated covariance function (Ecf). The number of random fields used to estimate the covariance function is denoted using nsim. From left to right: $\rho = 0.01, 0.5, 0.99$.

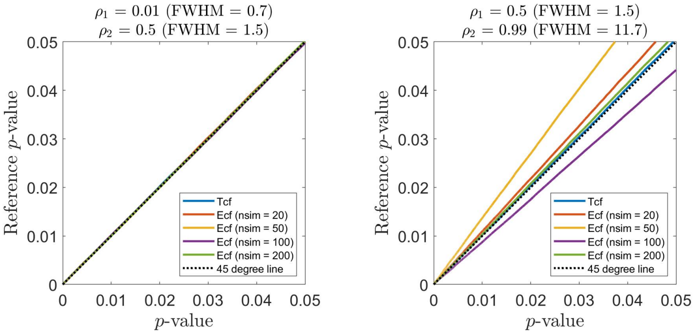


Figure 5.8. Comparison of the peak height distribution calculated from using MCDLM method with different covariance functions for 2D anisotropic stationary Gaussian fields. The left: $\rho_1 = 0.01$ and $\rho_2 = 0.5$ and the right: $\rho_1 = 0.5$ and $\rho_2 = 0.99$.

continuous formulae provide a good approximation to the height of local maxima, but in many realistic situations (FWHM < 7) the data is not sufficiently smooth and using the continuous formulae can lead to conservative inference. Furthermore, the continuous formulae only work for an isotropic field or a field that can be deformed to an isotropic field, but in practice this assumption is too restrictive. Additionally, the height distribution will be different for points on the boundary of the domain and these cases are not considered when using the continuous methods.

We showed that ADLM was liberal while the continuous RFT approach was conservative. We thus recommend using MCDLM to infer on peak height at all smoothness levels. However, when the data is very smooth and when it is reasonable to assume the data is isotropic, there may not be much gain relative to the continuous RFT approach and since the latter is very efficient we recommend it in that setting. In this case the covariance matrix is nearly singular, which causes problems when applied to our MCDLM method with estimated covariance function. A detailed running-time table for all methods under different scenarios is provided in Appendix D.5. The advantage of MCDLM is that it works well for local maxima on a rough lattice. ADLM only performs well under restrictive assumptions and for the partially connected neighborhood.

The proposed MCDLM method also works for t-fields, but it takes quite long time to implement this approach when the number of degrees of freedom is large. To improve the computational efficiency, we recommend using a Gaussianization transformation and then applying MCDLM to the Gaussianized field. The continuous RFT approach works better when both smoothness and degrees of freedom are high, but even when degrees of freedom increases to 200, it is still outperformed by MCDLM. Our approach can also easily be extended to obtain the peak height of two-sample t-statistic and F-fields of a lattice. The proposed method is limited to stationary Gaussian or Gaussian-derived random fields. However extensions to locally stationary and non-stationary fields are possible and are an interesting avenue for future research.

Chapter 5, in full, is currently being prepared for submission for publication of the material as it may appear in *Tuo Lin, Armin Schwartzman and Samuel Davenport. Peak p-values* *for Gaussian random fields on a lattice*. The dissertation author was the primary author of this paper.

Appendix A

Appendix for Chapter 1

A.1 Proof of equivalence of the nulls in (1.4) and (1.5).

Taking expectation on both sides of (1.3) yields:

$$\sum_{i=1}^{n_k} E\left(R_{ki}\right) = n_1 n_2 \Delta + \frac{n_k\left(n_k+1\right)}{2}. \tag{A.1}$$

Since the R_{ki}'s are natural numbers ranging over $[1, n_1+n_2]$, $E\left(R_{ki}\right) = \sum_{i=1}^{n_k} R_{ki}/n_k$.

If $E\left(R_{1i}\right) = E\left(R_{2j}\right)$, then substituting $\sum_{i=1}^{n_1} R_{1i}/n_1$ in place of $\sum_{j=1}^{n_2} R_{2j}/n_2$ in the following identity:

$$\frac{\sum_{i=1}^{n_1} R_{1i} + \sum_{j=1}^{n_2} R_{2j}}{n_1+n_2} = \frac{1}{2}\left(n_1+n_2+1\right),$$

and simplifying yields:

$$E\left(R_{1i}\right) = \frac{\sum_{i=1}^{n_1} R_{1i}}{n_1} = \frac{1}{2}\left(n_1+n_2+1\right).$$

Thus, $E\left(R_{1i}\right) = E\left(R_{2j}\right) = \left(n_1+n_2+1\right)/2$. Substituting $\left(n_1+n_2+1\right)/2$ in place of $E\left(R_{1i}\right)$ in (A.1) and simplifying yields: $\Delta = 1/2$.

If $\Delta = 1/2$, then (A.1) reduces to $E(R_{ki}) = (n_1+n_2+1)/2$, i.e., the groups have the same mean rank.

A.2 Proof of Theorem 1

a) The null of equal median implies $H_0 : \Delta = \frac{1}{2}$. We have:

$$\begin{aligned}
\Delta &= \int I(\xi \leq \eta) f_1(\xi) f_2(\eta) d\xi d\eta \\
&= \int_{\{\xi \geq m, \eta \leq m\}} I(\xi \leq \eta) f_1(\xi) f_2(\eta) d\xi d\eta + \int_{\{\xi \leq m, \eta \geq m\}} I(\xi \leq \eta) f_1(\xi) f_2(\eta) d\xi d\eta + \\
&\quad + \int_{\{\xi \geq m, \eta \geq m\}} I(\xi \leq \eta) f_1(\xi) f_2(\eta) d\xi d\eta + \int_{\{\xi \leq m, \eta \leq m\}} I(\xi \leq \eta) f_1(\xi) f_2(\eta) d\xi d\eta \\
&= \Pr(y_{1i} \geq m) \Pr\left(y_{2j} \geq m\right) + J,
\end{aligned}$$

where

$$J = \int_{\{\xi \geq m, \eta \geq m\}} I(\xi \leq \eta) f_1(\xi) f_2(\eta) d\xi d\eta + \int_{\{\xi \leq m, \eta \leq m\}} I(\xi \leq \eta) f_1(\xi) f_2(\eta) d\xi d\eta.$$

Since $\Pr(y_{1i} \geq m) \Pr\left(y_{2j} \geq m\right) = 1/4$, we only need to show that $J = 1/4$.

Let $y'_{1i} = 2m - y_{1i}$ and $y'_{2j} = 2m - y_{2j}$. Since y_{ki} have symmetric distribution, we have:

$$\begin{gathered}
f'_1(\xi) = f_1(\xi), \quad f'_2(\eta) = f_2(\eta), \quad d\xi' = -d\xi, \quad d\eta' = -d\eta, \\
\{y_{1i} \leq m\} = \{y'_{1i} \geq m\}, \quad \{y_{2j} \leq m\} = \{y'_{2j} \geq m\}, \quad I(\xi \leq \eta) = I\left(\xi' \leq \eta'\right).
\end{gathered}$$

Thus, we have:

$$\int_{\{\xi \leq m, \eta \leq m\}} I(\xi \leq \eta) f_1(\xi) f_2(\eta) d\xi d\eta = \int_{\{\xi' \geq m, \eta' \geq m\}} I\left(\xi' \geq \eta'\right) f_1\left(\xi'\right) f_2\left(\eta'\right) d\xi' d\eta'.$$

It the follows that

$$\begin{aligned}J &= \int_{\{\xi \geq m, \eta \geq m\}} [I(\xi \leq \eta) + I(\xi \geq \eta)] f_1(\xi) f_2(\eta) d\xi d\eta \\ &= \int_{\{\xi \geq m, \eta \geq m\}} f_1(\xi) f_2(\eta) d\xi d\eta \\ &= 1/4.\end{aligned}$$

b) The condition $H_0 : \Delta = \frac{1}{2}$ implies equal median. We show this by contradiction: if y_{ki} have different medians, then $\Delta \neq 1/2$.

Let m_k denote the median of y_{ki} $(k = 1, 2)$. If $m_1 \neq m_2$, then $y_{1i} - m_1 + m_2$ have the same median m_2 as y_{2j}. From a), we have

$$pr\left(y_{1i} - m_1 + m_2 \leq y_{2j}\right) = \frac{1}{2}.$$

Suppose that $m_1 > m_2$. Then,

$$pr(y_{1i} - m_1 + m_2 \leq c) > pr(y_{1i} \leq c)$$

for some c in the support of y_{1i}, $y_{1i} - m_1 + m_2$ and y_{2j}. Since y_{1i} and y_{2j} are independent,

$$\begin{aligned}pr\left(y_{1i} \leq y_{2j}\right) &= pr(y_{1i} \leq c)\, pr\left(y_{2j} \geq c\right), \\ pr\left(y_{1i} - m_1 + m_2 \leq y_{2j}\right) &= pr(y_{1i} - m_1 + m_2 \leq c)\, pr\left(y_{2j} \geq c\right),\end{aligned}$$

it follows that

$$\Delta = pr\left(y_{1i} \leq y_{2j}\right) < pr\left(y_{1i} - m_1 + m_2 \leq y_{2j}\right) = \frac{1}{2},$$

contradicting $\Delta = 1/2$. Similarly, if $m_1 < m_2$, we can show that

$$\Delta = pr\left(y_{1i} \leq y_{2j}\right) > pr\left(y_{1i} - m_1 + m_2 \leq y_{2j}\right) = 1/2.$$

A.3 Proof of Theorem 2

Let

$$V_n = \frac{1}{n_1}\frac{1}{n_2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2} w_{1i}w_{2j}I\left(y_{1i} \leq y_{2j}\right) = \frac{1}{n_1}\frac{1}{n_2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2} h\left(y_{1i}, w_{1i}; y_{2j}, w_{2j}\right). \qquad \text{(A.2)}$$

Then V_n is a two-sample U-statistic with kernel, $h\left(y_{1i}, w_{1i}; y_{2j}, w_{2i}\right) = w_{1i}w_{2j}I\left\{y_{1i} \leq y_{2j}\right\}$, and two bivariate arguments, (y_{1i}, w_{1i}) from the first and $\left(y_{2j}, w_{2j}\right)$ from the second sample [55]. Let x_i denote a vector of design variables that define the unique sampling weights such that $y_{ki} \perp w_{ki} \mid x_{ki}$ $(k = 1,2)$, where $\perp$ denotes stochastic independence. We then have:

$$\begin{aligned}
\delta &= E\left[h\left(y_{1i}, w_{1i}; y_{2j}, w_{2j}\right)\right] && \text{(A.3)}\\
&= E\left\{E\left[w_{1i}w_{2i}I\left(y_{1i} \leq y_{2j}\right) \mid x_{1i}, x_{2j}\right]\right\}\\
&= E\left\{E\left(w_{1i}w_{2i} \mid x_{1i}, x_{2j}\right)E\left[I\left(y_{1i} \leq y_{2j}\right) \mid x_{1i}, x_{2j}\right]\right\}\\
&= E\left[E\left(w_{1i}w_{2i} \mid x_{1i}, x_{2j}\right)\right]E\left\{E\left[I\left(y_{1i} \leq y_{2j}\right) \mid x_{1i}, x_{2j}\right]\right\}\\
&= E\left(w_{1i}\right)E\left(w_{2i}\right)E\left[I\left(y_{1i} \leq y_{2j}\right)\right]\\
&= E\left(w_{1i}\right)E\left(w_{2i}\right)\Delta.
\end{aligned}$$

Let

$$\begin{aligned}
&h_{1k}\left(y_{ki}, w_{ki}\right) = E\left[h\left(y_{1i}, w_{1i}; y_{2j}, w_{2j}\right) \mid y_{ki}, w_{ki}\right],\\
&\widetilde{h}_{1k}\left(y_{ki}, w_{ki}\right) = h_{1k}\left(y_{ki}, w_{ki}\right) - \delta, \quad \sigma_{h_k}^2 = Var\left(\widetilde{h}_{k1}\left(y_{ki}\right)\right), \quad k = 1,2.
\end{aligned}$$

By applying the theory of U-statistics, V_n has the same asymptotic distribution as its projection [55]:

$$\begin{aligned}
\widehat{V}_n &= \sum_{k=1}^{2} \frac{1}{n_k} \sum_{i=1}^{n_k} \widetilde{h}_{1k}\left(y_{ki}, w_{ki}\right) \\
&= \frac{\sqrt{n}}{\sqrt{n_1}} \frac{\sqrt{n_1}}{n_1} \sum_{i=1}^{n_1} \widetilde{h}_{11}\left(y_{1i}, w_{1i}\right) + \frac{\sqrt{n}}{\sqrt{n_2}} \frac{\sqrt{n_2}}{n_2} \sum_{i=1}^{n_2} \widetilde{h}_{12}\left(y_{2i}, w_{2i}\right) \\
&= S_{n1} + S_{n2}.
\end{aligned}$$

By applying the central limit theorem to S_{nk}, we have:

$$S_{nk} \to_d N\left(0, \rho_k^2 \sigma_{h_k}^2\right), \quad k = 1, 2.$$

Since S_{nk} are independent of each other, it follows from Slutsky's theorem that

$$\sqrt{n}\widehat{V}_n = S_{n1} + S_{n2} \to_d N\left(0, \sigma_V^2 = \rho_1^2 \sigma_{h_1}^2 + \rho_2^2 \sigma_{h_2}^2\right). \tag{A.4}$$

A consistent estimate of the asymptotic variance σ_V^2 is given by:

$$\widehat{\sigma}_V^2 = \frac{n}{n_1} \widehat{\sigma}_1^2 + \frac{n}{n_2} \widehat{\sigma}_2^2. \tag{A.5}$$

By expressing the MWW test statistic with sampling weights in (1.9) as a function of V_n, we have:

$$U_n = \frac{1}{\overline{w}_{1\cdot} \overline{w}_{2\cdot}} V_n. \tag{A.6}$$

It follows from the LLN, properties of convergence in probabilities and (A.3) that

$$\begin{aligned}
&\overline{w}_{k\cdot} \to_p E\left(w_{ki}\right), \\
&U_n = \frac{1}{\overline{w}_{1\cdot}} \frac{1}{\overline{w}_{2\cdot}} V_n \\
&\quad \to_p \frac{1}{E\left(w_{1i}\right) E\left(w_{2i}\right)} E\left(w_{1i}\right) E\left(w_{2i}\right) \Delta \\
&\quad = \Delta,
\end{aligned}$$

establishing consistency in Theorem 2/a.

Also,

$$\begin{aligned}
\sqrt{n}\left(U_n - \Delta\right) &= \sqrt{n}\left(\frac{1}{\overline{w}_1.\overline{w}_{2\cdot}} V_n - \Delta\right) \qquad \text{(A.7)} \\
&= \sqrt{n}\left(\frac{1}{\overline{w}_1.\overline{w}_{2\cdot}} V_n - \frac{\overline{w}_1.\overline{w}_{2\cdot}}{\overline{w}_1.\overline{w}_{2\cdot}} \Delta\right) \\
&= \sqrt{n}\left[\frac{1}{\overline{w}_1.\overline{w}_{2\cdot}} \left(V_n - \overline{w}_1.\overline{w}_2.\Delta\right)\right] \\
&= \frac{1}{\overline{w}_1.\overline{w}_{2\cdot}} \sqrt{n}\left[\frac{1}{n_1}\frac{1}{n_2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2} w_{1i}w_{2i} I\left(y_{1i} \leq y_{2j}\right) - \frac{1}{n_1}\frac{1}{n_2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2} w_{1i}w_{2i}\Delta\right] \\
&= \frac{1}{\overline{w}_1.\overline{w}_{2\cdot}} \sqrt{n}\left\{\frac{1}{n_1}\frac{1}{n_2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2} w_{1i}w_{2i}\left(\left[I\left(y_{1i} \leq y_{2j}\right) - \Delta\right]\right)\right\} \\
&= \frac{1}{\overline{w}_1.\overline{w}_{2\cdot}} \sqrt{n} V_n.
\end{aligned}$$

Theorem/(b) follows from (A.4) and an applications of Slutsky's theorem to the last equality in (A.7). Theorem/(c) follows from Theorem/(b) and (A.5).

Appendix B

Appendix for Chpater 2

B.1 Consistency Proof

Consistency of $\widehat{\rho}$ in (2.3). Under $r_{ij} \perp y_{ij}$, it follows from the theory of U-statistics

$$\begin{aligned}
\binom{n}{2}^{-1} \sum_{(i,j)\in C_2^n} r_{ij} &\to_p E\left(r_{12}\right), \\
\binom{n}{2}^{-1} \sum_{(i,j)\in C_2^n} r_{ij} y_{ij} &\to_p E\left(r_{12} y_{12}\right) = E\left(r_{12}\right) E\left(y_{12}\right) = E\left(r_{12}\right) \rho .
\end{aligned}$$

Thus, we have:

$$\widehat{\rho} = \left(\sum_{(i,j)\in C_2^n} r_{ij} \right)^{-1} \sum_{(i,j)\in C_2^n} r_{ij} y_{ij} \to_p \frac{E\left(r_{12}\right) E\left(y_{12}\right)}{E\left(r_{12}\right)} = E\left(y_{12}\right) = \rho .$$

Consistency of $\widehat{\rho}_n^{IPW}$ in (2.4). By the theory of U-statistics, we have

$$\widehat{\rho}_n^{IPW} = \binom{n}{2}^{-1} \sum_{(i,j)\in C_2^n} \frac{r_{ij}}{\pi_{ij}} y_{ij} \to_p E\left(\frac{r_{ij}}{\pi_{ij}} y_{ij} \right) .$$

Under $r_{ij} \perp y_{ij} \mid \mathbf{z}_i, \mathbf{z}_j$, it follows from the iterated conditional expectation

$$\begin{aligned} E\left(\frac{r_{ij}}{\pi_{ij}} y_{ij}\right) &= E\left[E\left(\frac{r_{ij}}{\pi_{ij}} y_{ij}\right) \mid y_{ij}, \mathbf{z}_i, \mathbf{z}_j\right] = E\left[\frac{1}{\pi_{ij}} y_{ij} E\left(r_{ij} \mid y_{ij}, \mathbf{z}_i, \mathbf{z}_j\right)\right] \\ &= E\left[\left(\frac{1}{\pi_{ij}} y_{ij} \pi_{ij}\right)\right] = E\left(y_{ij}\right) \\ &= \rho. \end{aligned}$$

Consistency of $\widehat{\rho}_n^{MSI}$ in (2.7). By the theory of U-statistics, we have

$$\widehat{\rho}_n^{MSI} = \binom{n}{2}^{-1} \sum_{\mathbf{i} \in C_2^n} \left\{r_{ij} y_{ij} + \left(1 - r_{ij}\right) g_{ij}\right\} \to_p E\left[r_{ij} y_{ij} + \left(1 - r_{ij}\right) g_{ij}\right].$$

Under $r_{ij} \perp y_{ij} \mid \mathbf{z}_i, \mathbf{z}_j$, it follows from the iterated conditional expectation

$$\begin{aligned} E\left[r_{ij} y_{ij} + \left(1 - r_{ij}\right) g_{ij}\right] &= E\left[E\left[\left(r_{ij} y_{ij} + \left(1 - r_{ij}\right) g_{ij}\right) \mid \mathbf{z}_i, \mathbf{z}_j\right]\right] \\ &= E\left[E\left[r_{ij}\left(y_{ij} - g_{ij}\right) \mid \mathbf{z}_i, \mathbf{z}_j\right] + g_{ij}\right] \\ &= E\left[E\left[r_{ij} \mid \mathbf{z}_i, \mathbf{z}_j\right] E\left[y_{ij} - g_{ij} \mid \mathbf{z}_i, \mathbf{z}_j\right]\right] + E\left[g_{ij}\right] \\ &= \rho. \end{aligned}$$

Consistency of $\widehat{\rho}_n^{DR}$ in (2.8). By the theory of U-statistics, we have

$$\widehat{\rho}_n^{DR} = \binom{n}{2}^{-1} \sum_{\mathbf{i} \in C_2^n} \left\{\frac{r_{ij}}{\pi_{ij}} y_{ij} + \left(1 - \frac{r_{ij}}{\pi_{ij}}\right) g_{ij}\right\} \to_p E\left[\frac{r_{ij}}{\pi_{ij}} y_{ij} + \left(1 - \frac{r_{ij}}{\pi_{ij}}\right) g_{ij}\right]$$

Under $r_{ij} \perp y_{ij} \mid \mathbf{z}_i, \mathbf{z}_j$, it follows from the iterated conditional expectation

$$\begin{aligned}
E\left[\frac{r_{ij}}{\pi_{ij}} y_{ij} + \left(1 - \frac{r_{ij}}{\pi_{ij}}\right) g_{ij}\right] &= E\left[E\left[\frac{r_{ij}}{\pi_{ij}} y_{ij} + \left(1 - \frac{r_{ij}}{\pi_{ij}}\right) g_{ij} \mid \mathbf{z}_i, \mathbf{z}_j\right]\right] \\
&= E\left[E\left[\frac{r_{ij}}{\pi_{ij}} \left(y_{ij} - g_{ij}\right) \mid \mathbf{z}_i, \mathbf{z}_j\right] + g_{ij}\right] \\
&= E\left[E\left[\frac{r_{ij}}{\pi_{ij}} \mid \mathbf{z}_i, \mathbf{z}_j\right] E\left[y_{ij} - g_{ij} \mid \mathbf{z}_i, \mathbf{z}_j\right]\right] + E\left[g_{ij}\right] \qquad \text{(B.1)}
\end{aligned}$$

Now, if $g_{ij} = E\left[y_{ij} \mid \mathbf{z}_i, \mathbf{z}_j\right]$, then (B.1) reduces to

$$E\left[g_{ij}\right] = \rho.$$

On the other hand, if $\pi_{ij} = E\left[r_{ij} \mid \mathbf{z}_i, \mathbf{z}_j\right] = \pi_{ij}$, then (B.1) reduces to

$$E\left[E\left[y_{ij} - g_{ij} \mid \mathbf{z}_i, \mathbf{z}_j\right]\right] + E\left[g_{ij}\right] = E\left[y_{ij}\right] = \rho.$$

B.2 Proof of Theorem 2

Let's first take a look at expressions for variance $V_{\mathbf{i}k}$ specified in the estimating equation for the doubly robust estimator.

Expressions for Variance $V_{\mathbf{i}k}$. For $V_{\mathbf{i}1}$ and $V_{\mathbf{i}2}$, we have:

$$\begin{aligned}
V_{ij1} &= Var\left(f_{ij1} \mid \mathbf{z}_i, \mathbf{z}_j\right) = \pi\left(\mathbf{z}_i; \gamma\right)_i \left(1 - \pi\left(\mathbf{z}_i; \gamma\right)\right) + \pi\left(\mathbf{z}_j; \gamma\right)\left(1 - \pi\left(\mathbf{z}_j; \gamma\right)\right) \\
V_{ij2} &= Var(f_{\mathbf{i}2} \mid \mathbf{z}_i, \mathbf{z}_j) = g\left(\mathbf{z}_i, \mathbf{z}_j; \beta\right)\left(1 - g\left(\mathbf{z}_i, \mathbf{z}_j; \beta\right)\right).
\end{aligned}$$

For V_{ij3},

$$\begin{aligned}
V_{ij3} &= Var\left(f_{ij3} \mid \mathbf{z}_i, \mathbf{z}_j\right) \\
&= E\left(f_{ij3}^2 \mid \mathbf{z}_i, \mathbf{z}_j\right) - E^2\left(f_{ij3} \mid \mathbf{z}_i, \mathbf{z}_j\right) \\
&= E\left[\left\{\frac{r_{ij}^2}{\pi_{ij}^2}\left(y_{ij} - g\left(\mathbf{z}_i, \mathbf{z}_j, \beta\right)\right)^2\right\} \mid \mathbf{z}_i, \mathbf{z}_j\right] \\
&= E\left[\left\{\frac{r_{ij}^2}{\pi_{ij}^2} g\left(\mathbf{z}_i, \mathbf{z}_j, \beta\right)\left(1 - g\left(\mathbf{z}_i, \mathbf{z}_j, \beta\right)\right)\right\} \mid \mathbf{z}_i, \mathbf{z}_j\right] \\
&= E\left[\left\{\frac{\left(r_i\left(1-r_j\right)\right)^2}{\left(\pi_i\left(1-\pi_j\right)\right)^2} g\left(\mathbf{z}_i, \mathbf{z}_j, \beta\right)\left(1 - g\left(\mathbf{z}_i, \mathbf{z}_j, \beta\right)\right)\right\} \mid \mathbf{z}_i, \mathbf{z}_j\right] \\
&= E\left[\left\{\frac{r_i\left(1-r_j\right)}{(\pi_i(1-\pi_j))^2} g\left(\mathbf{z}_i, \mathbf{z}_j, \beta\right)\left(1 - g\left(\mathbf{z}_i, \mathbf{z}_j, \beta\right)\right)\right\} \mid \mathbf{z}_i, \mathbf{z}_j\right] \\
&= \frac{1}{\pi_i(1-\pi_j)} g\left(\mathbf{z}_i, \mathbf{z}_j, \beta\right)\left(1 - g\left(\mathbf{z}_i, \mathbf{z}_j, \beta\right)\right).
\end{aligned}$$

Theorem 2 proof.

Proof. We first prove within-village case, where the two samples are from the same village. A Taylor's series expansion gives:

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

Let $D_{ij} = \frac{\partial}{\partial \theta^\top} h_{ij}(\theta)$ and $B = E\left(D_{ij}^\top V_{ij}^{-1} D_{ij}\right)$. From the theory of multivariate U-statistics [55], we have

$$\begin{aligned}
\frac{\partial}{\partial \theta^\top}\mathbf{U}_{n,ij}(\theta) &= \binom{n}{2}^{-1} \sum_{(i,j)\in C_2^n} \frac{\partial}{\partial \theta^\top}\left(-D_{ij}^\top V_{ij}^{-1} h_{ij}(\theta)\right) \\
&\to_p E\left[\left(-D_{ij}^\top V_{ij}^{-1}\right)\frac{\partial}{\partial \theta^\top} h_{ij}(\theta)\right] \\
&= -E\left(D_{ij}^\top V_{ij}^{-1} D_{ij}\right) \\
&= -B,
\end{aligned}$$

and

$$\begin{aligned}\sqrt{n}\mathbf{U}_{n,ij} &= \binom{n}{2}^{-1} \sum_{(i,j)\in C_2^n} \mathbf{U}_{n,ij} = \sqrt{n}\frac{1}{n}\sum_{i=1}^{n} 2E\left(\mathbf{U}_{n,ij} \mid \mathbf{Y}_i, \mathbf{X}_i\right) + \mathbf{o}_p(1) \\ &= \sqrt{n}\frac{1}{n}\sum_{i_1=1}^{n} \widetilde{\mathbf{v}}_i + \mathbf{o}_p(1) \\ &\rightarrow_d N\left(\mathbf{0}, \Sigma_U\right).\end{aligned}$$

By Slutsky's Theorem, we obtain

$$\begin{aligned}\sqrt{n}(\widehat{\theta}-\theta) &= \left(-\frac{\partial}{\partial\theta^\top}\mathbf{U}_{n,ij}(\theta)\right)^{-1} \sqrt{n}\mathbf{U}_{n,ij}(\theta) + \mathbf{o}_p(1) \\ &= B^{-1}\sqrt{n}\frac{1}{n}\sum_{i_1=1}^{n} \mathbf{v}_i + \mathbf{o}_p(1) \\ &\rightarrow_d N\left(\mathbf{0}, \Sigma_\theta\right),\end{aligned}$$

where $\Sigma_\theta = B^{-1}\Sigma_U B^{-\top}$. Consistency can be shown straightforwardly by WLLN from above.

Next, we prove between-village case, where the two samples are from different villages. Let's denote $n = n_r + n_s$. Again by Taylor's series expansion, we get

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

Similarly, from the theory of multivariate U-statistics, we have

$$\begin{aligned}\frac{\partial}{\partial\theta^\top}\mathbf{U}_{n,ij}(\theta) &= \left[\binom{n_r}{1}\binom{n_s}{1}\right]^{-1} \sum_{i=1}^{n_r}\sum_{j=1}^{n_s} \frac{\partial}{\partial\theta^\top}\left(-D_{ij}^\top V_{ij}^{-1} h_{ij}(\theta)\right) \\ &\rightarrow_p E\left[\left(-D_{ij}^\top V_{ij}^{-1}\right)\frac{\partial}{\partial\theta^\top} h_{ij}(\theta)\right] \\ &= -E\left(D_{ij}^\top V_{ij}^{-1} D_{ij}\right) \\ &= -B,\end{aligned}$$

and

$$\begin{aligned}
\sqrt{n}\mathbf{U}_{n,ij} &= \left[\binom{n_r}{1}\binom{n_s}{1}\right]^{-1}\sum_{i=1}^{n_r}\sum_{j=1}^{n_s}\mathbf{U}_{n,ij} \\
&= \frac{\sqrt{n}}{\sqrt{n_r}}\frac{\sqrt{n_r}}{n_r}\sum_{i=1}^{n_r}E\left(\mathbf{U}_{n,ij} \mid \mathbf{Y}_{ri},\mathbf{X}_{ri}\right)+\frac{\sqrt{n}}{\sqrt{n_s}}\frac{\sqrt{n_s}}{n_s}\sum_{j=1}^{n_s}E\left(\mathbf{U}_{n,ij} \mid \mathbf{Y}_{sj},\mathbf{X}_{sj}\right)+\mathbf{o}_p(1) \\
&= \rho_r\frac{\sqrt{n_r}}{n_r}\sum_{i=1}^{n_r}\mathbf{v}_{ri}+\rho_s\frac{\sqrt{n_s}}{n_s}\sum_{j=1}^{n_s}\mathbf{v}_{sj}+\mathbf{o}_p(1) \\
&\to_d N\left(\mathbf{0},\Sigma_U\right),
\end{aligned}$$

where $\Sigma_U = \rho_r^2 Var(\mathbf{v}_{ri})+\rho_s^2 Var(\mathbf{v}_{sj}) = \rho_r^2\Sigma_r+\rho_s^2\Sigma_s$.

By Slutsky's Theorem, we obtain

$$\begin{aligned}
\sqrt{n}(\widehat{\theta}-\theta) &= \left(-\frac{\partial}{\partial\theta^\top}\mathbf{U}_{n,ij}(\theta)\right)^{-1}\sqrt{n}\mathbf{U}_{n,ij}(\theta)+\mathbf{o}_p(1) \\
&= B^{-1}\left(\rho_r\frac{\sqrt{n_r}}{n_r}\sum_{i=1}^{n_r}\mathbf{v}_{ri}+\rho_s\frac{\sqrt{n_s}}{n_s}\sum_{j=1}^{n_s}\mathbf{v}_{sj}\right)+\mathbf{o}_p(1) \\
&\to_d N\left(\mathbf{0},\Sigma_\theta\right),
\end{aligned}$$

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

Appendix C

Appendix for Chapter 4

C.1 Proof of Asymptotic Unbiasedness

Proof of Lemma 8. Let $c(x)$ be the number of splits leading to the leaf $L(x)$ and $c_j(x)$ be the number of these splits along the j-th coordinates. By α-regular, we have $s_{min}\alpha^{c(x)} \leq 2k-1$, which implies $c(x) \geq log(s_{min}/(2k-1))/log(\alpha^{-1})$ for $0 < \alpha < 1$. Thus, the stochastic lower bound for $c_j(x)$ is:

$$c_j(x) \geq \text{Binom}\left(\frac{\log(s_{min}/(2k-1))}{\log\left(\alpha^{-1}\right)}; \frac{\pi}{d}\right).$$

By Chernoff's inequality and $s_{min}/s \gtrsim \varepsilon$ by positivity assumption, it follows that

$$\begin{aligned}\mathbb{P}\left[c_j(x) \leq \frac{\pi}{d}\frac{\log(s_{min}/(2k-1))}{\log\left(\alpha^{-1}\right)}(1-\eta)\right] &\leq \exp\left[-\frac{\eta^2}{2}\frac{\log(s_{min}/(2k-1))}{\pi^{-1}d\log\left(\alpha^{-1}\right)}\right]\\ &\leq \exp\left[-\frac{\eta^2}{2}\frac{\log(\varepsilon s/(2k-1))}{\pi^{-1}d\log\left(\alpha^{-1}\right)}\right]\\ &= \left(\frac{\varepsilon s}{2k-1}\right)^{-\frac{\eta^2}{2}\frac{1}{\pi^{-1}\log\left(\alpha^{-1}\right)}}.\end{aligned}$$

From Wager and Walther (2015) [ref], we have

$$\text{diam}_j(L(x)) \leq (1-\alpha)^{0.991c_j(x)}.$$

Combining with the Chernoff's inequality we get the conclusion of lemma 8. □

Proof of Theorem 9. Define $g(Y_i^{(0)}, Y_i^{(1)}) = I\{Y_i^{(0)} \leq Y_i^{(1)}\}$, then

$$\delta(x) = E\left[g(Y_i^{(0)}, Y_i^{(1)}) \middle| X_i = x\right].$$

In addition, $\widehat{\delta}_{tree}^{(b)}(x)$ can be written as

$$\widehat{\delta}_{tree}^{(b)}(x) = \frac{1}{\left|L_i^{(0)}\right|\left|L_i^{(1)}\right|} \sum_{Z_{j0} \in L_i^{(0)}} \sum_{Z_{l1} \in L_i^{(1)}} s_{j0} s_{l1} g(Y_{j0}, Y_{l1}),$$

where $s_{j0} = I\{Z_{j0} \in L_i^{(0)}\}$ and $s_{l1} = I\{Z_{l1} \in L_i^{(1)}\}$. Thus, by honesty,

$$\begin{aligned} &E[\widehat{\delta}_{tree}^{(b)}(x)] - E\left[g(Y_i^{(0)}, Y_i^{(1)}) \middle| X_i = x\right] \\ &\quad = E\left[E[g(Y_{j0}, Y_{l1}) \mid X_j, X_l \in L_i] - E\left[g(Y_i^{(0)}, Y_i^{(1)}) \middle| X_i = x\right]\right]. \end{aligned}$$

By Lipschitz continuity of $\delta(x)$,

$$\left| E[g(Y_{j0}, Y_{l1}) \mid X_j, X_l \in L_i] - E\left[g(Y_i^{(0)}, Y_i^{(1)}) \middle| X_i = x\right] \right| \leq \sqrt{2} C \operatorname{diam}(L_i).$$

Thus it suffices to show that the average diameter of leaf L_i is bounded. To do so, let $\eta = \sqrt{\log\left((1-\alpha)^{-1}\right.}$. Since we assume $\alpha \leq 0.2$, we have $\eta \leq 0.48$ and so $0.99 \cdot (1-\eta) \geq 0.51$. Since by Pythagorean theorem,

$$\begin{aligned} &\left\{ \operatorname{diam}(L_i) \geq \sqrt{d} \left(\frac{\varepsilon s}{2k-1} \right)^{-0.51 \frac{\log\left((1-\alpha)^{-1}\right)}{\log\left(\alpha^{-1}\right)} \frac{\pi}{d}} \right\} \\ &\qquad \subset \bigcup_j \left\{ \operatorname{diam}_j(L_i) \geq \left(\frac{\varepsilon s}{2k-1} \right)^{-0.51 \frac{\log\left((1-\alpha)^{-1}\right)}{\log\left(\alpha^{-1}\right)} \frac{\pi}{d}} \right\}. \end{aligned}$$

Applying lemma (8) and union bound, we obtain, for large enough s,

$$\mathbb{P}\left[\operatorname{diam}(L_i) \geq \sqrt{d}\left(\frac{\varepsilon s}{2k-1}\right)^{-0.51\frac{\log\left((1-\alpha)^{-1}\right)}{\log\left(\alpha^{-1}\right)}\frac{\pi}{d}}\right] \leq d\left(\frac{\varepsilon s}{2k-1}\right)^{-\frac{1}{2}\frac{\log\left((1-\alpha)^{-1}\right)}{\log\left(\alpha^{-1}\right)}\frac{\pi}{d}}.$$

Let A be the event that $\operatorname{diam}(L_i) \geq \sqrt{d}\left(\frac{\varepsilon s}{2k-1}\right)^{-0.51\frac{\log\left((1-\alpha)^{-1}\right)}{\log\left(\alpha^{-1}\right)}\frac{\pi}{d}}$. For large s, we have

$$\begin{aligned} & E\left|E[g(Y_{j0},Y_{l1}) \mid X_j, X_l \in L_i] - E\left[g(Y_i^{(0)},Y_i^{(1)})\middle|X_i = x\right]\right| P(A) \\ & \leq d\left(\frac{\varepsilon s}{2k-1}\right)^{-\frac{1}{2}\frac{\log\left((1-\alpha)^{-1}\right)}{\log\left(\alpha^{-1}\right)}\frac{\pi}{d}} \end{aligned} \tag{C.1}$$

and

$$\begin{aligned} & E\left|E[g(Y_{j0},Y_{l1}) \mid X_j, X_l \in L_i] - E\left[g(Y_i^{(0)},Y_i^{(1)})\middle|X_i = x\right]\right| P(A^c) \\ & \leq \sqrt{2}C\sqrt{d}\left(\frac{\varepsilon s}{2k-1}\right)^{-0.51\frac{\log\left((1-\alpha)^{-1}\right)}{\log\left(\alpha^{-1}\right)}\frac{\pi}{d}}. \end{aligned} \tag{C.2}$$

Combining the above (C.1) and (C.2), we get

$$\begin{aligned} & E\left|E[g(Y_{j0},Y_{l1}) \mid X_j, X_l \in L_i] - E\left[g(Y_i^{(0)},Y_i^{(1)})\middle|X_i = x\right]\right| \\ \leq \quad & (d + \sqrt{2}C\sqrt{d})\left(\frac{\varepsilon s}{2k-1}\right)^{-\frac{1}{2}\frac{\log\left((1-\alpha)^{-1}\right)}{\log\left(\alpha^{-1}\right)}\frac{\pi}{d}}. \end{aligned}$$

Thus the conclusion for Theorem 9 follows. □

C.2 Proof of Theorem 12

C.2.1 RF U-statistics

The random forest (RF) is defined as

$$RF\left(\mathbf{x};Z_1,\ldots,Z_n\right)=E_{\xi\sim\Xi}\left[T\left(\mathbf{x};\xi,Z_{i1},\ldots,Z_{is}\right)\right]=\binom{n}{s}^{-1}\sum_{(i_1,\ldots,i_s)\in C_s^n}T\left(\mathbf{x};Z_{i1},\ldots,Z_{is}\right). \quad \text{(C.3)}$$

And the Hajek projection of RF is defined as

$$\begin{aligned}\mathring{RF}\left(\mathbf{x};Z_1,\ldots,Z_n\right)&=\sum_{i=1}^{n}E\left[RF\left(\mathbf{x};Z_1,\ldots,Z_n\right)\mid Z_i\right]-(n-1)\,\theta\left(\mathbf{x}\right) \quad \text{(C.4)}\\ &=\sum_{i=1}^{n}\binom{n}{s}^{-1}\sum_{(i_1,\ldots,i_s)\in C_s^n}E\left[T\left(\mathbf{x};Z_{i1},\ldots,Z_{is}\right)\mid Z_i\right]-(n-1)\theta\left(\mathbf{x}\right)\end{aligned}$$

where $\theta\left(\mathbf{x}\right)=E\left[T\left(x;Z_{i1},\ldots,Z_{is}\right)\right]$. The projection in (C.4) can also be expressed in a centered version by

$$\begin{aligned}\mathring{RF}\left(\mathbf{x};Z_1,\ldots,Z_n\right)-\theta\left(\mathbf{x}\right)&=\sum_{i=1}^{n}E\left\{\left[RF\left(\mathbf{x};Z_1,\ldots,Z_n\right)-\theta\left(\mathbf{x}\right)\right]\mid Z_i\right\} \quad \text{(C.5)}\\ &=\sum_{i=1}^{n}\binom{n}{s}^{-1}\sum_{(i_1,\ldots,i_s)\in C_s^n}E\left\{\left[T\left(\mathbf{x};Z_{i1},\ldots,Z_{is}\right)-\theta\left(\mathbf{x}\right)\right]\mid Z_i\right\}\end{aligned}$$

Let

$$\begin{aligned}&T_1\left(Z_i\right)=E\left[T\left(\mathbf{x};Z_{i1},\ldots,Z_{is}\right)\mid Z_i\right],\quad \widetilde{T}_1\left(Z_i\right)=T_1\left(Z_i\right)-\theta\left(\mathbf{x}\right),\\ &e_n=\sqrt{n}\left(RF-\mathring{RF}\right),\quad \sigma_h^2=\operatorname{Var}\left[\widetilde{T}_1\left(Z_i\right)\right],\end{aligned}$$

We will show:

1) $e_n\to_p 0$ as $n\to\infty$ so that $\sqrt{n}\left(RF-\theta\left(\mathbf{x}\right)\right)$ and $\sqrt{n}\left(\mathring{RF}-\theta\left(\mathbf{x}\right)\right)$ have the same limit distribution;

2) Find the limitting distribution of $\sqrt{n}\left(\mathring{RF}-\theta\left(\mathbf{x}\right)\right)$.

To this end, we first derive a decomposition of RF in terms of the projection of RF onto an filtration defined by $(Z_1, \ldots, Z_k)$ for $1 \leq k \leq s$. This decomposition can also be expressed for a single tree $T(\mathbf{x}; Z_1, \ldots, Z_s)$ in terms of its projection onto the filtration.

Note that $\mathring{RF}$ is the projection of RF onto the filtration defined by Z_1.

Let $\mathscr{F}_0 = \{\emptyset, \Omega\}$ and $\mathscr{F}_k = \sigma(Z_1, \ldots, Z_k)$ for $1 \leq k \leq s$ be a sequence of σ-field of $Z_1, \ldots, Z_k$. For an integer k $(1 \leq k \leq s)$, let

$$\widetilde{T} = T - \theta(\mathbf{x}), \quad T_k(Z_{i1}, \ldots, Z_{is}) = E(T \mid Z_1, \ldots, Z_k), \quad \widetilde{T}_k = T_k - \theta(\mathbf{x}).$$

Then, $T_k(Z_1, \ldots, Z_k)$ is a random variable since it is the conditional expectation of T given $Z_1, \ldots, Z_k$. By law of iterated conditional expectation, it follows that

$$\begin{aligned} E(T_k \mid \mathscr{F}_{k-1}) &= E[E(T \mid \mathscr{F}_k) \mid \mathscr{F}_{k-1}] = E(T \mid \mathscr{F}_{k-1}) = T_{k-1} \\ E\left(\mathring{T}_k \mid \mathscr{F}_{k-1}\right) &= E(T_k \mid \mathscr{F}_{k-1}) - \theta(\mathbf{x}) = T_{k-1} - \theta(\mathbf{x}) = \mathring{T}_{k-1}. \end{aligned} \tag{C.6}$$

For notation brevity we express $E(T_k \mid Z_1, \ldots, Z_{k-1})$ using $E(T_k \mid \mathscr{F}_{k-1})$.

For $1 < k \leq s$, define

$$\begin{aligned} g_1(Z_1) &= \widetilde{T}_1 \\ g_k(Z_1, \ldots, Z_k) &= \widetilde{T}_k - \sum_{l=1}^{k-1} \sum_{(i_1, \ldots, i_l) \in C_l^k} g_l(Z_{i_1}, \ldots, Z_{i_l}). \end{aligned} \tag{C.7}$$

By (C.6), we can readily show that $E[g_k(Z_1, \ldots, Z_k) \mid \mathscr{F}_{k-1}] = 0$ $(1 \leq k \leq s)$. Also, since $\widetilde{T}_k$ are all centered, we have $E[g_k(Z_1, \ldots, Z_K)] = 0$ $(1 \leq k \leq s)$. Now, by Theorem 2 in Chapter 3 of Kowalski and Tu (2007) [ref], let

$$S_{kn} = \sum_{(i_1, \ldots, i_k) \in C_k^n} g_k(Z_{i_1}, \ldots, Z_{i_k}), \quad 1 \leq k \leq s.$$

The RF has the following representation:

$$RF - \theta(\mathbf{x}) = \sum_{k=1}^{s} \binom{s}{k} \binom{n}{k}^{-1} S_{kn}. \tag{C.8}$$

Since $\binom{s}{k}\binom{n}{k}^{-1} = \binom{n}{s}^{-1}\binom{n-k}{s-k}$ (see below), by (C.6) we can also express the decomposition in (C.8) in terms of the s-argument kernel, or a single tree, $T(\mathbf{x}; Z_1, \ldots, Z_s)$ (replacing k with s), which is also known as the ANOVA decomposition in Effron [ref]:

$$T(\mathbf{x}; Z_1, \ldots, Z_s) = \theta(\mathbf{x}) + \sum_{i=1}^{s} g_1(Z_i) + \sum_{(i_1,i_2)\in C_2^s} g_2(Z_{i_1}, Z_{i_2}) + \cdots + g_s(Z_1, \ldots, Z_s) \tag{C.9}$$

To see $\binom{s}{k}\binom{n}{k}^{-1} = \binom{n}{s}^{-1}\binom{n-k}{s-k}$, first we express it equivalently as: $\binom{n}{s}\binom{s}{k} = \binom{n}{k}\binom{n-k}{s-k}$. We can view $\binom{n}{s}\binom{s}{k}$ as the total of ways of choosing s from n subjects and then choosing k from s subjects, which is the same as choosing k from n subjects and then choosing $s-k$ from $n-k$ subjects.

Thus we have from the definition of random forest in (C.3)

$$
\begin{aligned}
RF(\mathbf{x};Z_1,\dots,Z_n) &= \binom{n}{s}^{-1} \sum_{(i_1,\dots,i_s)\in C_s^n} T(\mathbf{x};Z_{i1},\dots,Z_{is}) \\
&= \binom{n}{s}^{-1} \sum_{(i_1,\dots,i_s)\in C_s^n} \left[\theta(\mathbf{x}) + \sum_{i=1}^{s} g_1(Z_i) + \sum_{(i_1,i_2)\in C_2^s} g_2(Z_{i_1},Z_{i_2}) + \right. \\
&\quad + \dots + g_s(Z_1,\dots,Z_s)] \\
&= \theta(\mathbf{x}) + \binom{n}{s}^{-1} \left[\binom{n-1}{s-1} \sum_{i=1}^{n} g_1(Z_i) + \binom{n-2}{s-2} \sum_{(i_1,i_2)\in C_2^n} g_2(Z_{i_1},Z_{i_2}) + \right. \\
&\quad \left. + \dots + \sum_{(i_1,\dots,i_s)\in C_s^n} g_s(Z_{i_1},\dots,Z_{i_s}) \right] \\
&= \theta(\mathbf{x}) + \sum_{k=1}^{s} \binom{s}{k}\binom{n}{k}^{-1} \sum_{(i_1,\dots,i_k)\in C_k^n} g_k(Z_{i_1},\dots,Z_{i_k}) \\
&= \theta(\mathbf{x}) + \frac{s}{n}\sum_{i=1}^{n} g_1(Z_i) + \sum_{k=2}^{s} \binom{s}{k}\binom{n}{k}^{-1} \sum_{(i_1,\dots,i_k)\in C_k^n} g_k(Z_{i_1},\dots,Z_{i_k}) . \quad \text{(C.10)}
\end{aligned}
$$

Since

$$
\begin{aligned}
E(g_1(Z_1)\mid Z_1) &= g_1(Z_1) = \widetilde{T}_1 \\
E(g_2(Z_1,Z_2)\mid Z_1) &= E(\widetilde{T}_2\mid Z_1) - E[g_1(Z_1)\mid Z_1] - E[g_1(Z_2)\mid Z_1] \\
&= \widetilde{T}_1 - \widetilde{T}_1 - E[g_1(Z_2)] \\
&= 0 \\
E(g_3(Z_1,Z_2,Z_3)\mid Z_1) &= E(\widetilde{T}_3\mid Z_1) - \sum_{i=1}^{3} E[g_1(Z_i)\mid Z_1] - \sum_{(i,j)\in C_2^3} E\left[g_2(Z_i,Z_j)\mid Z_1\right] \\
&= \widetilde{T}_1 - \widetilde{T}_1 - \sum_{i=2}^{3} E[g_1(Z_i)] - E[g_2(Z_1,Z_2)\mid Z_1] - E[g_2(Z_1,Z_3)\mid Z_1] + \\
&\quad - E[g_2(Z_2,Z_3)] \\
&= 0,
\end{aligned}
$$

and $E(g_k(Z_1,\ldots,Z_k) \mid Z_1) = 0$ for $k \geq 3$ by proving in a similar manner, $E(S_{kn} \mid Z_1) = 0$ for $k \geq 2$. Thus, we can express (C.10) as:

$$\begin{aligned}
\mathring{R}F - \theta(\mathbf{x}) &= \sum_{i=1}^{n} E\left\{\left[RF(\mathbf{x}; Z_1, \ldots, Z_n) - \theta(\mathbf{x})\right] \mid Z_i\right\} \\
&= \frac{1}{n}\sum_{i=1}^{n} s g_1(Z_i).
\end{aligned}$$

And (C.10) can also be expressesd in lieu of (C.8) as

$$\begin{aligned}
RF - \theta(\mathbf{x}) &= \binom{s}{1}\binom{n}{1}^{-1} S_{1n} + \sum_{k=2}^{s}\binom{s}{k}\binom{n}{k}^{-1} S_{kn} \\
&= \frac{1}{n}\sum_{i=1}^{n} s g_1(Z_i) + \sum_{k=2}^{s}\binom{s}{k}\binom{n}{k}^{-1} S_{kn} \\
&= \mathring{R}F - \theta + \sum_{k=2}^{s}\binom{s}{k}\binom{n}{k}^{-1} S_{kn}.
\end{aligned}$$

We can also define the Hajek projection $\mathring{T}$ of a single tree, or kernel, T as

$$\begin{aligned}
\mathring{T} - \theta(\mathbf{x}) &= \sum_{i=1}^{s} E\left[\left(T(\mathbf{x}; Z_1, \ldots, Z_s) - \theta(\mathbf{x})\right) \mid Z_i\right] - (s-1)\theta(\mathbf{x}) \\
&= \sum_{i=1}^{s} g_1(Z_i).
\end{aligned}$$

Then, we can express $T(\mathbf{x}; Z_1, \ldots, Z_s)$ and $RF(\mathbf{x}; Z_1, \ldots, Z_n)$ in terms of their projections as:

$$\begin{aligned}
T - \theta(\mathbf{x}) &= \mathring{T} + \sum_{(i_1,i_2)\in C_2^s} g_2(Z_{i_1}, Z_{i_2}) + \cdots + g_s(Z_1, \ldots, Z_s), \qquad \text{(C.11)} \\
RF - \theta(\mathbf{x}) &= \mathring{R}F - \theta(\mathbf{x}) + \sum_{k=2}^{s}\binom{s}{k}\binom{n}{k}^{-1} \sum_{(i_1,\ldots,i_k)\in C_k^n} g_k(Z_{i_1}, \ldots, Z_{i_k}).
\end{aligned}$$

To show that $RF - \theta(\mathbf{x})$ and $\mathring{R}F - \theta(\mathbf{x})$ have the same asymptotic distribution, it follows from (C.11)

$$\sqrt{n}\left(RF - \theta\left(\mathbf{x}\right)\right) = \sqrt{n}\left(\mathring{R}F - \theta\left(\mathbf{x}\right)\right) + e_n, \tag{C.12}$$
$$e_n = \sqrt{n}\sum_{k=2}^{s}\binom{s}{k}\binom{n}{k}^{-1}\sum_{(i_1,\ldots,i_k)\in C_k^n} g_k\left(Z_{i_1},\ldots,Z_{i_k}\right),$$

where e_n is the normalized difference between RF and $\mathring{R}F$. By iterated conditional expectation, we can easily show that $E\left[g_{s_1}(Z_{i_1},...,Z_{i_{s_1}})g_{s_2}(Z_{j_1},...,Z_{j_{s_2}})\right] = 0$ for all $s_1 \neq s_2$ or $\{i_1,...,i_{s_1}\} \neq \{j_1,...,j_{s_2}\}$ with $1 \leq s_1, s_2 \leq s$.

In classic U-statistics theory,

$$\begin{aligned}
E\left(e_n^2\right) &= n\sum_{k=2}^{s}\binom{s}{k}^2\binom{n}{k}^{-1}\operatorname{Var}\left[g_k(Z_{i_1},\ldots,Z_{i_k})\right] \\
&= n\frac{\binom{s}{2}^2}{\binom{n}{2}}\operatorname{Var}\left[g_2(Z_{i_1},Z_{i_2})\right] + \cdots + n\frac{\binom{s}{1}^2}{\binom{n}{s}}\operatorname{Var}\left[g_s(Z_{i_1},\ldots,Z_{i_s})\right] \\
&= nO(n^{-2}) \\
&\rightarrow_p 0
\end{aligned}$$

since s is a finite fixed constant. Thus, the U-statistic and its projection have the same asymptotic distribution. In addition, $\operatorname{Var}(sg_1\left(Z_i\right)) = s^2\operatorname{Var}\left(g_1\left(Z_i\right)\right) = s\operatorname{Var}\left(\mathring{T}\right)$ is a finite constant. Then the asymptotic distribution of $\sqrt{n}\left(\mathring{R}F - \theta\left(\mathbf{x}\right)\right)$ is readily obtained by applying the classic Lindeberg CLT for i.i.d. random sequence Z_i to $\mathring{R}F - \theta\left(\mathbf{x}\right)$ and is given by $N\left(0, s^2Var\left(g_1\left(Z_i\right)\right)\right)$.

Within the context of random forests, $s \rightarrow \infty$. Moreover, since only the Z_i's with X_i close to $\mathbf{x}$ contribute to $\mathring{T} - \theta\left(\mathbf{x}\right)$, $s^2Var\left(g_1\left(Z_i\right)\right)$ is no longer a constant. Under this case , $\mathring{R}F - \theta\left(\mathbf{x}\right)$ may not converge to 0 at the rate $n^{-\frac{1}{2}}$. However, $N\left(0, \frac{s^2Var(g_1(Z_i))}{n}\right)$ is a valid asymptotic distribution if $\frac{s^2Var(g_1(Z_i))}{n} \rightarrow 0$, as $n, s \rightarrow \infty$. In this paper, contraints are imposed on the rate of $s \rightarrow \infty$ to ensure $\frac{s^2Var(g_1(Z_i))}{n} \rightarrow 0$, as $n, s \rightarrow \infty$. Thus, unlike regular and asymptotically normal estimators for semiparametric models, $\mathring{R}F - \theta\left(\mathbf{x}\right) \rightarrow 0$ no longer have $n^{-\frac{1}{2}}$ convergence rate.

Thus we introduce the Lyapunov CLT to show asymptotic normality of $\mathring{RF} - \theta(\mathbf{x})$, rather than the Lindeberg CLT.

C.2.2 Lyapunov CLT

Let's first show the normalized difference between RF and $\mathring{RF}$ is $o_p(1)$. Let $\sigma_{ns}^2 = \sum_{i=1}^n \operatorname{Var}[sg_1(Z_i)] = \sum_{i=1}^n s^2 \operatorname{Var}[g_1(Z_i)] = ns^2 \operatorname{Var}[g_1(Z_i)]$. Multiplying $\frac{1}{\sqrt{\frac{\sigma_{ns}^2}{n}}}$ on both side of (C.12) we obtain

$$\frac{\sqrt{n}}{\sqrt{\frac{\sigma_{ns}^2}{n}}}(RF - \theta(\mathbf{x})) = \frac{\sqrt{n}}{\sqrt{\frac{\sigma_{ns}^2}{n}}}\left(\mathring{RF} - \theta(\mathbf{x})\right) + \frac{1}{\sqrt{\frac{\sigma_{ns}^2}{n}}}e_n.$$

Then it is equivalent to show that

$$E\left[\frac{n}{\sigma_{ns}^2}e_n^2\right] \to_p 0.$$

To this end,

$$\begin{aligned}
E\left[\frac{n}{\sigma_{ns}^2}e_n^2\right] &= \frac{n}{\sigma_{ns}^2}E\left[e_n^2\right] \\
&= \frac{n}{s^2 \operatorname{Var}[g_1(Z_i)]}\sum_{k=2}^{s}\binom{s}{k}^2\binom{n}{k}^{-1}\operatorname{Var}[g_k(Z_{i_1},\ldots,Z_{i_k})] \\
&\le \frac{n}{s\operatorname{Var}\left[\mathring{T}\right]}\frac{\binom{s}{2}}{\binom{n}{2}}\sum_{k=2}^{s}\binom{s}{k}\operatorname{Var}[g_k(Z_{i_1},\ldots,Z_{i_k})] \\
&= \frac{s-1}{n-1}\frac{\operatorname{Var}[T]}{\operatorname{Var}\left[\mathring{T}\right]}. \qquad \text{(C.13)}
\end{aligned}$$

The last equality is because

$$\operatorname{Var}[T] = \sum_{k=1}^{s}\binom{s}{k}\operatorname{Var}[g_k(Z_{i_1},\ldots,Z_{i_k})].$$

Thus, it is critical to bound $\frac{\operatorname{Var}[T]}{\operatorname{Var}[\mathring{T}]}$ asymptotically so that (C.13) converge to 0 in probability. Note that as we discussed in Section C.2.1, in classic U-statistics when s is a finite constant,

$\frac{\mathrm{Var}[T]}{\mathrm{Var}[\mathring{T}]} = O(1)$. We will show the bound $\frac{\mathrm{Var}[T]}{\mathrm{Var}[\mathring{T}]}$, which is named as ν-incrementality in Wager and Athey (2018) [ref], in Section (C.2.3).

Next, we apply Lyapunov CLT to derive the asymptotic normality of $\mathring{RF}$. We first show that $s^2\,\mathrm{Var}(g_1(Z_i)) \to \infty$ at the rate of $\frac{s}{\log(s)^d}$, as $s \to \infty$.

$$\begin{aligned} s^2\,\mathrm{Var}(g_1(Z_i)) &= s^2\,\mathrm{Var}\left[E(T \mid Z_i)\right] \\ &\geq s^2\,\mathrm{Var}\left[E(S_1 \mid Z_1)\right]\mathrm{Var}\left(Y \mid X = x\right) \\ &\sim \Omega\left(s^2 \frac{1}{s\log(s)^d}\right) \\ &\sim \Omega\left(\frac{s}{\log(s)^d}\right) \\ &\to \infty, \quad \text{as } s \to \infty. \end{aligned}$$

where the second inequality is from the intermediate steps in Section C.2.3. Thus, $\frac{s^2 Var(g_1(Z_i))}{n} \to 0$ as $n, s \to \infty$.

By applying Lyapunov CLT, to show

$$\frac{\sqrt{n}}{\sqrt{\frac{\sigma_{ns}^2}{n}}}\left(\mathring{RF} - \theta(\mathbf{x})\right) = \frac{\sqrt{n}}{\sqrt{\frac{\sigma_{ns}^2}{n}}}\frac{1}{n}\sum_{i=1}^{n} sg_1(Z_i) = \frac{1}{\sigma_{ns}}\sum_{i=1}^{n} sg_1(Z_i) \to_d N(0,1),$$

we need to verify the following condition:

$$\lim_{n\to\infty} \frac{1}{\left(\sigma_{ns}^2\right)^{1+\frac{\delta}{2}}}\sum_{i=1}^{n} E\left(\left|sg_1(Z_i)\right|^{2+\delta}\right) = 0, \tag{C.14}$$

for $\delta > 0$.

Note that unlike regular and asymptotically normal estimators for semiparametric models, $\frac{\sigma_{ns}^2}{n}$ does not converge to a constant as $n, s \to \infty$. However, since $\frac{\sigma_{ns}^2}{n^2} = \frac{s^2\,\mathrm{Var}[sg_1(Z_i)]}{n} \to 0$ as

$n, s \to \infty$,

$$\frac{\sqrt{n}}{\sqrt{\frac{\sigma_{ns}^2}{n}}}\left(\mathring{RF} - \theta\left(\mathbf{x}\right)\right)$$

has a limiting normal, but $\mathring{RF} - \theta\left(\mathbf{x}\right)$ converges to 0 at a rate slower than $\frac{1}{\sqrt{n}}$.

Given that $E\left(T \mid Z_i\right) - \theta\left(\mathbf{x}\right) = sg_1\left(Z_i\right)$, $Es\left[g_1(Z_i)\right] = 0$ and $E\left(T\right) = \theta\left(\mathbf{x}\right)$, (C.14) is equivalent to

$$\lim_{n\to\infty}\frac{1}{\left(\sum_{i=1}^{n}\operatorname{Var}\left[E\left(T \mid Z_i\right)\right]\right)^{1+\frac{\delta}{2}}}\sum_{i=1}^{n}E\left[\left|E\left(T \mid Z_i\right) - \theta\left(\mathbf{x}\right)\right|^{2+\delta}\right] = 0. \qquad \text{(C.15)}$$

We first bound $E\left[\left|E\left(T \mid Z_i\right) - \theta\left(\mathbf{x}\right)\right|^{2+\delta}\right]$. By the definition of k-PNN,

$$T = \sum_{j=1}^{n_0}\sum_{l=1}^{n_1}S_{0j}S_{1l}g\left(Y_{0j}, Y_{1l}\right).$$

Let i_0 be the set of i that includes all the samples in group 0, and i_1 be the set of i that includes all the samples in group 1. The Hajek projection of T is defined as

$$E\left(T \mid Z_i\right) = E\left(T \mid Z_{0j}\right)I\left\{i \in i_0\right\} + E\left(T \mid Z_{1l}\right)I\left\{i \in i_1\right\}.$$

We consider $E\left(T \mid Z_{0j}\right)I\left\{i \in i_0\right\}$ first, and $E\left(T \mid Z_{1l}\right)I\left\{i \in i_1\right\}$ will be considered using similar strategy. By adding and subtract $\sum_{j=1}^{n_0}\sum_{l=1}^{n_1} S_{0j}S_{1l}E\left[g\left(Y_{0j},Y_{1l}\right)\mid X_{0j},X_{1l}\right]$, we get

$$\begin{aligned}
E\left(T\mid Z_{0j}\right)-E\left(T\right) &= E\left(T\mid Z_{01}\right)-E\left(T\right)\\
&= E\left(\sum_{j=1}^{n_0}\sum_{l=1}^{n_1} S_{0j}S_{1l}g\left(Y_{0j},Y_{1l}\right)\right. \qquad \text{(C.16)}\\
&\quad \left.-\sum_{j=1}^{n_0}\sum_{l=1}^{n_1} S_{0j}S_{1l}E\left[g\left(Y_{0j},Y_{1l}\right)\mid X_{0j},X_{1l}\right]\middle| Z_{01}\right)\\
&\quad +E\left(\sum_{j=1}^{n_0}\sum_{l=1}^{n_1} S_{0j}S_{1l}E\left[g\left(Y_{0j},Y_{1l}\right)\mid X_{0j},X_{1l}\right]\middle| Z_{01}\right)-E\left(T\right)\\
&= E\left(\sum_{j=1}^{n_0}\sum_{l=1}^{n_1} S_{0j}S_{1l}\left(g\left(Y_{0j},Y_{1l}\right)-E\left[g\left(Y_{0j},Y_{1l}\right)\mid X_{0j},X_{1l}\right]\right)\middle| Z_{01}\right)+\\
&\quad +E\left(\sum_{j=1}^{n_0}\sum_{l=1}^{n_1} S_{0j}S_{1l}E\left[g\left(Y_{0j},Y_{1l}\right)\mid X_{0j},X_{1l}\right]\middle| Z_{01}\right)-E\left(T\right) \qquad \text{(C.17)}
\end{aligned}$$

By honesty, we can verify that for any index $j>1$,

$$\begin{aligned}
&E\left(S_{0j}S_{1l}\left(g\left(Y_{0j},Y_{1l}\right)-E\left[g\left(Y_{0j},Y_{1l}\right)\mid X_{0j},X_{1l}\right]\right)\middle| Z_{01}\right)\\
&=E\left(E\left(S_{0j}S_{1l}\left(g\left(Y_{0j},Y_{1l}\right)-E\left[g\left(Y_{0j},Y_{1l}\right)\mid X_{0j},X_{1l}\right]\right)\mid Z_{01},X_{0j},X_{1l}\right)\mid Z_{01}\right)\\
&=E\left(E\left(S_{0j}S_{1l}\mid Z_{01},X_{0j},X_{1l}\right)E\left(\left\{g\left(Y_{0j},Y_{1l}\right)-E\left[g\left(Y_{0j},Y_{1l}\right)\mid X_{0j},X_{1l}\right]\right\}\mid Z_{01},X_{0j},X_{1l}\right)\mid Z_{01}\right)\\
&=0,
\end{aligned}$$

since

$$\begin{aligned}
&E\left(\left\{g\left(Y_{0j},Y_{1l}\right)-E\left[g\left(Y_{0j},Y_{1l}\right)\mid X_{0j},X_{1l}\right]\right\}\mid Z_{01},X_{0j},X_{1l}\right)\\
&=E\left(\left\{g\left(Y_{0j},Y_{1l}\right)-E\left[g\left(Y_{0j},Y_{1l}\right)\mid X_{0j},X_{1l}\right]\right\}\mid X_{0j},X_{1l}\right)\\
&=0.
\end{aligned}$$

Note: Although $\left(Y_{0j},X_{0j}\right)\perp Z_{01}$ for any index $j>1$, $S_{0j}\perp Z_{01}$ may not be true, since S_{0j} may contain Z_{01}.

Therefore (C.17) reduces to

$$E\left(T \mid Z_{0j}\right) - E\left(T\right) = E\left(\sum_{l=1}^{n_1} S_{01}S_{1l}\left(g\left(Y_{01},Y_{1l}\right) - E\left[g\left(Y_{01},Y_{1l}\right) \mid X_{01},X_{1l}\right]\right) \middle| Z_{01}\right) + \\ + E\left(\sum_{j=1}^{n_0}\sum_{l=1}^{n_1} S_{0j}S_{1l}E\left[g\left(Y_{0j},Y_{1l}\right) \mid X_{0j},X_{1l}\right] \middle| Z_{01}\right) - E\left(T\right).$$

Note that the two right-hand-side terms above both have mean-zero, because

$$\begin{aligned}
&E\left[E\left(\sum_{l=1}^{n_1} S_{01}S_{1l}\left(g\left(Y_{01},Y_{1l}\right) - E\left[g\left(Y_{01},Y_{1l}\right) \mid X_{01},X_{1l}\right]\right) \middle| Z_{01}\right)\right] \\
&= E\left(\sum_{l=1}^{n_1} S_{01}S_{1l}\left(g\left(Y_{01},Y_{1l}\right) - E\left[g\left(Y_{01},Y_{1l}\right) \mid X_{01},X_{1l}\right]\right)\right) \\
&= \sum_{l=1}^{n_1} E\left(S_{01}S_{1l}\left(g\left(Y_{01},Y_{1l}\right) - E\left[g\left(Y_{01},Y_{1l}\right) \mid X_{01},X_{1l}\right]\right)\right) \\
&= \sum_{l=1}^{n_1} E\left(S_{01}S_{1l}g\left(Y_{01},Y_{1l}\right)\right) - \sum_{l=1}^{n_1} E\left(S_{01}S_{1l}E\left[g\left(Y_{01},Y_{1l}\right) \mid X_{01},X_{1l}\right]\right) \\
&= \sum_{l=1}^{n_1} E\left(S_{01}S_{1l}g\left(Y_{01},Y_{1l}\right)\right) - \sum_{l=1}^{n_1} E\left(E\left(S_{01}S_{1l}E\left[g\left(Y_{01},Y_{1l}\right) \mid X_{01},X_{1l}\right] \mid X_{01},X_{1l}\right)\right) \\
&= \sum_{l=1}^{n_1} E\left(S_{01}S_{1l}g\left(Y_{01},Y_{1l}\right)\right) - \sum_{l=1}^{n_1} E\left(E\left(S_{01}S_{1l}g\left(Y_{01},Y_{1l}\right) \mid X_{01},X_{1l}\right)\right) \\
&= 0
\end{aligned}$$

Similarly, we can get

$$E\left(E\left(\sum_{j=1}^{n_0}\sum_{l=1}^{n_1} S_{0j}S_{1l}E\left[g\left(Y_{0j},Y_{1l}\right) \mid X_{0j},X_{1l}\right] \middle| Z_{01}\right)\right) = E\left(T\right)$$

Let

$$\begin{aligned}
w_1 &= E\left(\sum_{l=1}^{n_1} S_{01}S_{1l}\left(g\left(Y_{01},Y_{1l}\right) - E\left[g\left(Y_{01},Y_{1l}\right) \mid X_{01},X_{1l}\right]\right) \middle| Z_{01}\right) \\
w_2 &= E\left(\sum_{j=1}^{n_0}\sum_{l=1}^{n_1} S_{0j}S_{1l}E\left[g\left(Y_{0j},Y_{1l}\right) \mid X_{0j},X_{1l}\right] \middle| Z_{01}\right) - E\left(T\right).
\end{aligned}$$

It follows from Jensen's inequality that

$$\left(\frac{w_1+w_2}{2}\right)^{2+\delta} \leq \frac{1}{2}w_1^{2+\delta}+\frac{1}{2}w_2^{2+\delta},$$

which yields

$$\begin{aligned}
&2^{-(1+\delta)}\left(E\left(T \mid Z_{0j}\right)-E\left(T\right)\right)^{2+\delta} \\
&\quad \leq E\left(\sum_{l=1}^{n_1} S_{01}S_{1l}\left(g\left(Y_{01},Y_{1l}\right)-E\left[g\left(Y_{01},Y_{1l}\right) \mid X_{01},X_{1l}\right]\right)\middle|Z_{01}\right)^{2+\delta} \\
&\quad +\left[E\left(\sum_{j=1}^{n_0}\sum_{l=1}^{n_1} S_{0j}S_{1l}E\left[g\left(Y_{0j},Y_{1l}\right) \mid X_{0j},X_{1l}\right]\middle|Z_{01}\right)-E\left(T\right)\right]^{2+\delta}.
\end{aligned}$$

It then follows by triangular inequality that

$$\begin{aligned}
&2^{-(1+\delta)}E\left\{\left|E\left(T \mid Z_{0j}\right)-E\left(T\right)\right|^{2+\delta}\right\} \\
&\quad \leq E\left[\left|E\left(\sum_{l=1}^{n_1} S_{01}S_{1l}\left(g\left(Y_{01},Y_{1l}\right)-E\left[g\left(Y_{01},Y_{1l}\right) \mid X_{01},X_{1l}\right]\right)\middle|Z_{01}\right)\right|^{2+\delta}\right] \\
&\quad +E\left[\left|E\left(\sum_{j=1}^{n_0}\sum_{l=1}^{n_1} S_{0j}S_{1l}E\left[g\left(Y_{0j},Y_{1l}\right) \mid X_{0j},X_{1l}\right]\middle|Z_{01}\right)-E\left(T\right)\right|^{2+\delta}\right].
\end{aligned}$$

Now, again by honesty, $E\left(S_{01}S_{1l} \mid Z_{01},Z_{1l}\right)=E\left(S_{01}S_{1l} \mid X_{01},X_{1l}\right)$.

Since we assume that $E\left\{\left|g\left(Y_{0j},Y_{1l}\right)-E\left[g\left(Y_{0j},Y_{1l}\right) \mid X_{0j},X_{1l}\right]\right| \mid X_{0j}=X_{1l}=\mathbf{x}\right\} \leq M$, we have

$$\begin{aligned}
&E\left[\left|E\left(\sum_{l=1}^{n_1} S_{01}S_{1l}\left(g\left(Y_{01},Y_{1l}\right)-E\left[g\left(Y_{01},Y_{1l}\right)\mid X_{01},X_{1l}\right]\right)\middle|Z_{01}\right)\right|^{2+\delta}\right]\\
&=E\left[\left|E\left(E\left(\sum_{l=1}^{n_1} S_{01}S_{1l}\mid X_{01},X_{1l}\right)E\left(g\left(Y_{01},Y_{1l}\right)-E\left[g\left(Y_{01},Y_{1l}\right)\mid X_{01},X_{1l}\right]\mid X_{01},X_{1l}\right)\middle|Z_{01}\right)\right|^{2+\delta}\right]\\
&\leq E\left[\left|E\left(E\left(\sum_{l=1}^{n_1} S_{01}S_{1l}\mid X_{01},X_{1l}\right)M\middle|Z_{01}\right)\right|^{2+\delta}\right]\\
&\leq E\left[M^{2+\delta}E\left(\sum_{l=1}^{n_1} S_{01}S_{1l}\mid Z_{01}\right)^{2+\delta}\right]\\
&\leq M^{2+\delta}E\left[\left(\sum_{l=1}^{n_1} S_{01}S_{1l}\mid Z_{01}\right)^{2}\right]
\end{aligned}$$

since $S_{01}, S_{1l} \leq 1$ for all $1 \leq l \leq n_1$. Also, since $E\left[g\left(Y_{0j},Y_{1l}\right)\mid X_{0j}=\mathbf{x},X_{1l}=\mathbf{x}\right]$ is Lipschitz, let $u = \sup\left\{\left|E\left[g\left(Y_{0j},Y_{1l}\right)\mid X_{0j}=\mathbf{x},X_{1l}=\mathbf{x}\right]\right| : \mathbf{x}\in[0,1]^d\right\}$. Then,

$$\begin{aligned}
&E\left[\left|E\left(\sum_{j=1}^{n_0}\sum_{l=1}^{n_1} S_{0j}S_{1l}E\left[g\left(Y_{0j},Y_{1l}\right)\mid X_{0j},X_{1l}\right]\middle|Z_{01}\right)-E\left(T\right)\right|^{\delta}\right]\\
&\leq E\left[\left|u\left\{E\left(\sum_{j=1}^{n_0}\sum_{l=1}^{n_1} S_{0j}S_{1l}\middle|Z_{01}\right)-E\left(\sum_{j=1}^{n_0}\sum_{l=1}^{n_1} S_{01}S_{1l}\right)\right\}\right|^{\delta}\right]\\
&\leq \left(2u\right)^{\delta},
\end{aligned}$$

and

$$\begin{aligned}
&E\left[\left|E\left(\sum_{j=1}^{n_0}\sum_{l=1}^{n_1} S_{0j}S_{1l}E\left[g\left(Y_{0j},Y_{1l}\right)\mid X_{0j},X_{1l}\right]\middle|Z_{01}\right)-E\left(T\right)\right|^2\right]\\
&=\operatorname{Var}\left[E\left(\sum_{j=1}^{n_0}\sum_{l=1}^{n_1} S_{0j}S_{1l}E\left[g\left(Y_{0j},Y_{1l}\right)\mid X_{0j},X_{1l}\right]\middle|Z_{01}\right)\right]\\
&\le u^2\operatorname{Var}\left[E\left[\sum_{j=1}^{n_0}\sum_{l=1}^{n_1} S_{0j}S_{1l}\middle|Z_{01}\right]\right]\\
&=u^2\operatorname{Var}\left[E\left[\sum_{l=1}^{n_1} S_{01}S_{1l}\middle|Z_{01}\right]+E\left[\sum_{j=2}^{n_0}\sum_{l=1}^{n_1} S_{0j}S_{1l}\middle|Z_{01}\right]\right]\\
&\le 2u^2\left(\operatorname{Var}\left[E\left[\sum_{l=1}^{n_1} S_{01}S_{1l}\middle|Z_{01}\right]\right]+\operatorname{Var}\left[E\left[\sum_{j=2}^{n_0}\sum_{l=1}^{n_1} S_{0j}S_{1l}\middle|Z_{01}\right]\right]\right).
\end{aligned}$$

The last inequality above could be derived by

$$\begin{aligned}
&2\operatorname{Cov}\left(E\left[\sum_{l=1}^{n_1} S_{01}S_{1l}\middle|Z_{01}\right],E\left[\sum_{j=2}^{n_0}\sum_{l=1}^{n_1} S_{0j}S_{1l}\middle|Z_{01}\right]\right)\\
&\le\operatorname{Var}\left(E\left[\sum_{l=1}^{n_1} S_{01}S_{1l}\middle|Z_{01}\right]\right)+\operatorname{Var}\left(E\left[\sum_{j=2}^{n_0}\sum_{l=1}^{n_1} S_{0j}S_{1l}\middle|Z_{01}\right]\right).
\end{aligned}$$

Summarizing the above, we have

$$\begin{aligned}
&E\left[\left|E\left(\sum_{j=1}^{n_0}\sum_{l=1}^{n_1} S_{0j}S_{1l}E\left[g\left(Y_{0j},Y_{1l}\right)\mid X_{0j},X_{1l}\right]\middle|Z_{01}\right)-E\left(T\right)\right|^{2+\delta}\right]\\
&\le(2u)^{\delta}2u^2\left(\operatorname{Var}\left[E\left[\sum_{l=1}^{n_1} S_{01}S_{1l}\middle|Z_{01}\right]\right]+\operatorname{Var}\left[E\left[\sum_{j=2}^{n_0}\sum_{l=1}^{n_1} S_{0j}S_{1l}\middle|Z_{01}\right]\right]\right)\\
&\le(2u)^{\delta}2u^2\left(2\operatorname{Var}\left[E\left[\sum_{l=1}^{n_1} S_{01}S_{1l}\middle|Z_{01}\right]\right]\right)\\
&\le(2u)^{2+\delta}E\left[E\left(\sum_{l=1}^{n_1} S_{01}S_{1l}\middle|Z_{01}\right)^2\right].
\end{aligned}$$

Thus, we conclude

$$E\left[\left|E\left(T\mid Z_{0j}\right)-\theta(\mathbf{x})\right|^{2+\delta}\right]\le C_1E\left[E\left(\sum_{l=1}^{n_1} S_{01}S_{1l}\middle|Z_{01}\right)^2\right],$$

for some constant $C_1 < \infty$. Similarly, we can also show

$$E\left[\left|E\left(T \mid Z_{1l}\right)-\theta(\mathbf{x})\right|^{2+\delta}\right] \leq C_2 E\left[E\left(\sum_{j=1}^{n_2} S_{0j}S_{11}\middle| Z_{11}\right)^2\right],$$

for another constant $C_2 < \infty$.

The condition in (C.14) or (C.15) can thus be simplifies to

$$\lim_{n\to\infty} \frac{nE\left[E\left(\sum_{l=1}^{n_1} S_{01}S_{1l}\middle| Z_{01}\right)^2\right]}{\left\{nVar\left(E\left(T \mid Z_{01}\right)\right)\right\}^{1+\delta/2}} = 0, \tag{C.18}$$

when $i \in i_0$, and

$$\lim_{n\to\infty} \frac{nE\left[E\left(\sum_{j=1}^{n_2} S_{0j}S_{11}\middle| Z_{11}\right)^2\right]}{\left\{nVar\left(E\left(T \mid Z_{11}\right)\right)\right\}^{1+\delta/2}} = 0, \tag{C.19}$$

when $i \in i_1$.

Let's focus on showing (C.18) and (C.19) can be shown in a similar way. From theorems and lemmas in Section C.2.3 and Lemma 4 and Theorem 5 in Wager and Athey (2018), we have

$$\begin{aligned}
&Var\left(E\left(T \mid Z_{01}\right)\right) = \Omega\left(E\left[E\left(\sum_{l=1}^{n_1} S_{01}S_{1l}\middle| Z_{01}\right)^2\right] \mathrm{Var}\left[\sum_{i_1=1}^{n_1} S_{1i_1} g(Y_{0i_0}, Y_{1i_1}) \mid X_{0i_0} = \mathbf{x}\right]\right), \\
&\mathrm{Var}\left[\sum_{i_1=1}^{n_1} S_{1i_1} g(Y_{0i_0}, Y_{1i_1}) \mid X_{0i_0} = \mathbf{x}\right] > 0, \\
&\left\{nE\left[E\left(\sum_{l=1}^{n_1} S_{01}S_{1l}\middle| Z_{01}\right)^2\right]\right\}^{-\delta/2} \lesssim \left(\frac{C_{f,d}}{2k}\frac{n}{s\log(s)^d}\right)^{-\delta/2} \to 0\,.
\end{aligned}$$

Therefore the condition in (C.14) further reduces to

$$
\begin{aligned}
&\lim_{n\to\infty} \frac{nE\left[E\left(\sum\limits_{l=1}^{n_1} S_{01}S_{1l}\middle|Z_{01}\right)^2\right]}{\left\{n\Omega\left(E\left[E\left(\sum\limits_{l=1}^{n_1} S_{01}S_{1l}\middle|Z_{01}\right)^2\right]\operatorname{Var}\left[\sum_{i_1=1}^{n_1} S_{1i_1}g(Y_{0i_0},Y_{1i_1}) \mid X_{0i_0}=\mathbf{x}\right]\right)\right\}^{1+\delta/2}} \\
&\leq \lim_{n\to\infty} \frac{1}{n^{\delta/2}E\left[E\left(\sum\limits_{l=1}^{n_1} S_{01}S_{1l}\middle|Z_{01}\right)^2\right]^{\delta/2}\operatorname{Var}\left[\sum_{i_1=1}^{n_1} S_{1i_1}g(Y_{0i_0},Y_{1i_1}) \mid X_{0i_0}=\mathbf{x}\right]^{1+\delta/2}} \\
&\to 0.
\end{aligned}
$$

C.2.3 ν-incrementality

Recall that our between-subjects causal tree is defined as

$$
\begin{aligned}
T(\mathbf{x};Z_1,...,Z_s) &= \sum_{j=1}^{n_0}\sum_{l=1}^{n_1} S_{0j}S_{1l}g(Y_{0j},Y_{1l}), \\
g(Y_{0j},Y_{1l}) &= I(Y_{0j}\leq Y_{1l}),
\end{aligned}
$$

where

$$
S_{0j} = \begin{cases} \left|\left\{j : X_{0j}\in L_0(\mathbf{x};Z) \text{ and } R_{0j}=0\right\}\right|^{-1} & \text{if } X_{0j}\in L_0(\mathbf{x};Z), \\ 0 & \text{else,} \end{cases}
$$

$$
S_{1l} = \begin{cases} \left|\{l : X_{1l}\in L_1(\mathbf{x};Z) \text{ and } R_{1l}=1\}\right|^{-1} & \text{if } X_{1l}\in L_1(\mathbf{x};Z), \\ 0 & \text{else ,} \end{cases}
$$

and $R_{0j(1l)}$ is the treatment indicator, $L_0(\mathbf{x};Z) = \left\{X_{0j} : X_{0j}\in L(\mathbf{x};Z) \text{ and } R_{0j}=0\right\}$ and $L_1(\mathbf{x};Z) = \{X_{1l} : X_{1l}\in L(\mathbf{x};Z) \text{ and } R_{1l}=1\}$.

First we focus on $\operatorname{Var}[E[T(\mathbf{x};Z_1,\ldots,Z_s)\mid Z_1]]$. An independent term of the Hájek projection of $T(\mathbf{x};Z_1,\ldots,Z_s)$ can be written as

$$
\begin{aligned}
E\left(T(\mathbf{x};Z_1,\ldots,Z_s)\mid Z_{01}\right) &= E\left[\sum_{j=1}^{n_0}\left(\sum_{l=1}^{n_1} S_{0j}S_{1l}g(Y_{0j},Y_{1l})\right)\middle| Z_{01}\right] \\
&= E\left[\sum_{j=1}^{n_0} S_{0j}\left(\sum_{l=1}^{n_1} S_{1l}g(Y_{0j},Y_{1l})\right)\middle| Z_{01}\right].
\end{aligned}
$$

By Theorem 5 of Wager and Athey 2018, we have

$$
\operatorname{Var}\left[E\left[T(\mathbf{x};Z_1,\ldots,Z_s)\mid Z_{01}\right]\right] \gtrsim \operatorname{Var}\left[E\left(S_{01}\mid Z_{01}\right)\right]\operatorname{Var}\left[\sum_{i_1=1}^{n_1} S_{1i_1}g(Y_{0i_0},Y_{1i_1})\mid X_{0i_0}=\mathbf{x}\right].
$$

Combining with Lemma 4 in Wager and Athey 2018, we obtain

$$
\operatorname{Var}\left[E\left[T(\mathbf{x};Z_1,\ldots,Z_s)\mid Z_{01}\right]\right] \gtrsim \frac{1}{k}\frac{\nu(s)}{s}\operatorname{Var}\left[\sum_{i_1=1}^{n_1} S_{1i_1}g(Y_{0i_0},Y_{1i_1})\mid X_{0i_0}=\mathbf{x}\right].
$$

By Lipschitz continuity,

$$
\left|E\left[\sum_{l=1}^{n_1} S_{1l}g(Y_{0j},Y_{1l})\mid X_{0j}\in L(x;Z)\right]-E\left[\sum_{i_1=1}^{n_1} S_{1i_1}g(Y_{0i_0},Y_{1i_1})\mid X_{0i_0}=\mathbf{x}\right]\right| \le \sum_{l=1}^{n_1} S_{1l}C\,\mathrm{diam}(L_i),
$$

for some constant C. Combining with Lemma 8 and $E\left[\sum_{i_1=1}^{n_1} S_{1i_1}g(Y_{0i_0},Y_{1i_1})\mid X_{0i_0}=\mathbf{x}\right]<\infty$ since $\left|E\left[g(Y_{0i_0},Y_{1i_1})\mid X_{0i_0}=\mathbf{x}\right]\right|$ is bounded, we have

$$
E\left[\sum_{l=1}^{n_1} S_{1l}g(Y_{0j},Y_{1l})\mid X_{0j}\in L(x;Z)\right] \to_p E\left[\sum_{i_1=1}^{n_1} S_{1i_1}g(Y_{0i_0},Y_{1i_1})\mid X_{0i_0}=\mathbf{x}\right],
$$

which implies

$$
\left(E\left[\sum_{l=1}^{n_1} S_{1l}g(Y_{0j},Y_{1l})\mid X_{0j}\in L(x;Z)\right]\right)^2 \to_p \left(E\left[\sum_{i_1=1}^{n_1} S_{1i_1}g(Y_{0i_0},Y_{1i_1})\mid X_{0i_0}=\mathbf{x}\right]\right)^2. \tag{C.20}
$$

Moreover,

$$
E\left[\left(\sum_{l=1}^{n_1} S_{1l}g(Y_{0j},Y_{1l})\right)^2\right] = E\left[\sum_{l=1}^{n_1} S_{1l}^2 g(Y_{0j},Y_{1l})^2\right] + E\left[\sum_{(l_1,l_2)\in C_2^{n_1}} S_{1l_1}S_{1l_2}g(Y_{0j},Y_{1l_1})g(Y_{0j},Y_{1l_2})\right].
$$

Again by Lipschitz continuity,

$$\left| E\left[\sum_{l=1}^{n_1} S_{1l}^2 g(Y_{0j},Y_{1l})^2 \mid X_{0j}\in L(x;Z)\right] - E\left[\sum_{i_1=1}^{n_1} S_{1i_1}^2 g(Y_{0i_0},Y_{1i_1})^2 \mid X_{0i_0}=\mathbf{x}\right]\right| \le \sum_{l=1}^{n_1} S_{1l}^2 C_1 \operatorname{diam}(L_i),$$

$$\left| E\left[\sum_{(l_1,l_2)\in C_2^{n_1}} S_{1l_1}S_{1l_2} g(Y_{0j},Y_{1l_1}) g(Y_{0j},Y_{1l_2}) \mid X_{0j}\in L(x;Z)\right] + \right.$$

$$\left. - E\left[\sum_{(i_1,i'_1)\in C_2^{n_1}} S_{1i_1}S_{1i'_1} g(Y_{0i_0},Y_{1i_1}) g(Y_{0i_0},Y_{1i'_1}) \mid X_{0i_0}=\mathbf{x}\right]\right|$$

$$= \left| E\left[\sum_{(l_1,l_2)\in C_2^{n_1}} S_{1l_1}S_{1l_2} g(Y_{0j},Y_{1l_1})\left(g(Y_{0j},Y_{1l_2}) - g(Y_{0i_0},Y_{1l_2})\right) \mid X_{0j}\in L(x;Z), X_{0i_0}=\mathbf{x}\right] + \right.$$

$$+ E\left[\sum_{(i_1,i'_1)\in C_2^{n_1}} S_{1i_1}S_{1i'_1}\left(g(Y_{0j},Y_{1i_1}) - g(Y_{0i_0},Y_{1i_1})\right) g(Y_{0i_0},Y_{1i'_1}) \mid X_{0j}\in L(x;Z), X_{0i_0}=\mathbf{x}\right]$$

$$\le \sum_{l=1}^{n_1} S_{1l}S_{1l_2} C_2 \operatorname{diam}(L_i)$$

for some constant C_1, C_2. The second equation is because

$$E\left[\sum_{(l_1,l_2)\in C_2^{n_1}} S_{1l_1}S_{1l_2} g(Y_{0j},Y_{1l_1}) g(Y_{0i_0},Y_{1l_2}) \mid X_{0j}\in L(x;Z), X_{0i_0}=\mathbf{x}\right]$$

$$= E\left[\sum_{(i_1,i'_1)\in C_2^{n_1}} S_{1i_1}S_{1i'_1} g(Y_{0j},Y_{1i'_1}) g(Y_{0i_0},Y_{1i_1}) \mid X_{0j}\in L(x;Z), X_{0i_0}=\mathbf{x}\right].$$

Thus, we have

$$E\left[\sum_{l=1}^{n_1} S_{1l}^2 g(Y_{0j},Y_{1l})^2 \mid X_{0j}\in L(x;Z)\right] \to_p E\left[\sum_{i_1=1}^{n_1} S_{1i_1}^2 g(Y_{0i_0},Y_{1i_1})^2 \mid X_{0i_0}=\mathbf{x}\right], \quad \text{and}$$

$$E\left[\sum_{(l_1,l_2)\in C_2^{n_1}} S_{1l_1}S_{1l_2} g(Y_{0j},Y_{1l_1}) g(Y_{0j},Y_{1l_2}) \mid X_{0j}\in L(x;Z)\right]$$

$$\to_p E\left[\sum_{(i_1,i'_1)\in C_2^{n_1}} S_{1i_1}S_{1i'_1} g(Y_{0i_0},Y_{1i_1}) g(Y_{0i_0},Y_{1i'_1}) \mid X_{0i_0}=\mathbf{x}\right],$$

which implies

$$E\left[\left(\sum_{l=1}^{n_1} S_{1l}g(Y_{0j},Y_{1l})\right)^2 \mid X_{0j}\in L(x;Z)\right] \to_p E\left[\left(\sum_{i_1=1}^{n_1} S_{1i_1}g(Y_{0i_0},Y_{1i_1})\right)^2 \mid X_{0i_0}=\mathbf{x}\right]. \tag{C.21}$$

By (C.20) and (C.21),

$$\operatorname{Var}\left[\sum_{l=1}^{n_1} S_{1l}g(Y_{0j},Y_{1l}) \mid X_{0j}\in L_0(x;Z)\right] \to_p \operatorname{Var}\left[\sum_{i_1=1}^{n_1} S_{1i_1}g(Y_{0i_0},Y_{1i_1}) \mid X_{0i_0}=\mathbf{x}\right].$$

Next,

$$\begin{aligned}
\operatorname{Var}[T] &= \operatorname{Var}\left[\sum_{j=1}^{n_0}\sum_{l=1}^{n_1} S_{0j}S_{1l}g(Y_{0j},Y_{1l})\right] \\
&= \operatorname{Var}\left[E\left(\sum_{j=1}^{n_0}\sum_{l=1}^{n_1} S_{0j}S_{1l}g(Y_{0j},Y_{1l}) \mid X_{0j}\right)\right] + E\left[\operatorname{Var}\left(\sum_{j=1}^{n_0}\sum_{l=1}^{n_1} S_{0j}S_{1l}g(Y_{0j},Y_{1l}) \mid X_{0j}\right)\right] \\
&= \operatorname{Var}\left[\sum_{j=1}^{n_0} E\left(S_{0j}\mid X_{0j}\right)E\left(\sum_{l=1}^{n_1} S_{1l}g(Y_{0j},Y_{1l}) \mid X_{0j}\right)\right] + \\
&\quad + E\left[\operatorname{Var}\left(\sum_{\left\{j:X_{0j}\in L_0(x;Z)\right\}} \frac{1}{|L_0(x;Z)|}\sum_{l=1}^{n_1} S_{1l}g(Y_{0j},Y_{1l}) \mid X_{0j}\in L_0(x;Z)\right)\right] \\
&= \operatorname{Var}\left[\sum_{\left\{j:X_{0j}\in L_0(x;Z)\right\}} \frac{1}{|L_0(x;Z)|}E\left(\sum_{l=1}^{n_1} S_{1l}g(Y_{0j},Y_{1l}) \mid X_{0j}\in L_0(x;Z)\right)\right] + \\
&\quad + \sum_{\left\{j:X_{0j}\in L_0(x;Z)\right\}} \frac{1}{|L_0(x;Z)|^2}E\left[\operatorname{Var}\left(\sum_{l=1}^{n_1} S_{1l}g(Y_{0j},Y_{1l}) \mid X_{0j}\in L_0(x;Z)\right)\right]
\end{aligned} \tag{C.22}$$

so that

$$\begin{aligned}
&|L_0(x;Z)|\operatorname{Var}[T] \\
&\quad \to_p \operatorname{Var}\left[E\left[\sum_{i_1=1}^{n_1} S_{1i_1}g(Y_{0i_0},Y_{1i_1}) \mid X_{0i_0}=\mathbf{x}\right]\right] + E\left[\operatorname{Var}\left[\sum_{i_1=1}^{n_1} S_{1i_1}g(Y_{0i_0},Y_{1i_1}) \mid X_{0i_0}=\mathbf{x}\right]\right] \\
&\quad = \operatorname{Var}\left[\sum_{i_1=1}^{n_1} S_{1i_1}g(Y_{0i_0},Y_{1i_1}) \mid X_{0i_0}=\mathbf{x}\right].
\end{aligned}$$

Thus

$$\begin{aligned} k\,\mathrm{Var}[T] &\leq |L_0(x;Z)|\,\mathrm{Var}[T] \\ &\to_p \mathrm{Var}\left[\sum_{i_1=1}^{n_1} S_{1i_1} g(Y_{0i_0}, Y_{1i_1}) \mid X_{0i_0} = \mathbf{x}\right]. \end{aligned}$$

And we conclude that

$$\frac{\mathrm{Var}[\mathring{T}(\mathbf{x};Z_1,\ldots,Z_s)]}{\mathrm{Var}[T(\mathbf{x};Z_1,\ldots,Z_s)]} \gtrsim k \frac{s\,\mathrm{Var}\left[E\left[T(\mathbf{x};Z_1,\ldots,Z_s) \mid Z_{01}\right]\right]}{\mathrm{Var}\left[\sum_{l=1}^{n_1} S_{1l} g(Y_{0j}, Y_{1l}) \mid X_{0j} = \mathbf{x}\right]} \gtrsim \nu(s),$$

where $\nu(s) = C_{f,d}/log(s)^d$ for some constant $C_{f,d}$ from Wager and Athey (2018) Theorem 5.

Appendix D

Appendix for Chapter 5

D.1 Theory for ADLM

D.1.1 Proof of Proposition 13

Proof. Without loss of generality, we assume that Z is mean zero and unit variance. For $s \in S$, taking $\mathscr{N}(s)$ to be the partially connected neighborhood of s, for $s_a, s_b \in \mathscr{N}(s)$, we have

$$\begin{pmatrix} Z(s) \\ Z(s_a) \\ Z(s_b) \end{pmatrix} \sim N\left(\mathbf{0}, \begin{pmatrix} \Sigma_{AA} & \Sigma_{AB} \\ \Sigma_{BA} & \Sigma_{BB} \end{pmatrix}\right),$$

where $\Sigma_{AA} = \mathrm{Var}(Z(s)) = 1$, $\Sigma_{AB} = \Sigma_{BA}^{\top} = [\mathrm{Cov}(Z(s), Z(s_a))\ \mathrm{Cov}(Z(s), Z(s_b))] = [\rho(s, s_a)\ \rho(s, s_b)]$ and $\Sigma_{BB} = \mathrm{Var}\left[\begin{pmatrix} Z(s_a) \\ Z(s_b) \end{pmatrix}\right] = \begin{pmatrix} 1 & \rho(s_a, s_b) \\ \rho(s_a, s_b) & 1 \end{pmatrix}$.

For $z \in \mathbb{R}$, the covariance of $Z(s_a)$ and $Z(s_b)$ conditional on $Z(s) = z$ is

$$\begin{aligned} \mathrm{Var}\left[\begin{pmatrix} Z(s_a) \\ Z(s_b) \end{pmatrix} \middle| Z(s) = z\right] &= \Sigma_{BB} - \Sigma_{BA}\Sigma_{AA}^{-1}\Sigma_{AB} \\ &= \Sigma_{BB} - \Sigma_{BA}\Sigma_{AB} \\ &= \begin{pmatrix} 1 & \rho(s_a, s_b) \\ \rho(s_a, s_b) & 1 \end{pmatrix} - \begin{pmatrix} \rho(s, s_a)^2 & \rho(s, s_a)\rho(s, s_b) \\ \rho(s, s_a)\rho(s, s_b) & \rho(s, s_b)^2 \end{pmatrix}. \end{aligned}$$

Under the assumption that Z is obtained by integrating Gaussian white noise against a separable Gaussian kernel, the off-diagonal entries are

$$\rho(s_a, s_b) - \rho(s, s_a)\rho(s, s_b) = \rho^{||s_a - s_b||^2} - \rho^{||s_a - s||^2}\rho^{||s_b - s||^2} = \rho^{||s_a - s_b||^2} - \rho^{||s_a - s||^2 + ||s_b - s||^2},$$

where ρ is the correlation between two adjacent voxels and $|| \cdot ||$ denotes the Euclidean norm. Thus, if $||s_a - s_b||^2 = ||s_a - s||^2 + ||s_b - s||^2$, then $\mathrm{Cov}(Z(s_a), Z(s_b)|Z(s)) = 0$. Now $(Z(s_a), Z(s_b))$ follows a bivariate Gaussian distribution so it follows that $Z(s_a)$ and $Z(s_b)$ is independent conditional on $Z(s_5)$. When $s_a = s \pm e_{d_1}$ and $s_b = s \pm e_{d_2}$, where d_1 and d_2 denote different lattice directions, $||s_a - s_b||^2 = ||e_{d_1} \pm e_{d_2}||^2$ which equals $||s_a - s||^2 + ||s_b - s||^2 = ||e_{d_1}||^2 + ||e_{d_2}||^2$ and so the result follows. □

D.1.2 Theoretical derivation of the probability density function of ADLM method

Under the condition of 13. For each $s \in S$ assume that $\mathcal{N}_{PC}(s)$ is the partially connected neighbourhood defined in (5.2). Then

$$P[Z(s) > u | Z(t) < Z(s), \forall t \in \mathcal{N}_{PC}(s)] = \frac{\int_u^\infty \left(\prod_{d=1}^D Q(\rho_d, z)(z)\right) \phi(z) dz}{\int_{-\infty}^\infty \left(\prod_{d=1}^D Q(\rho_d, z)(z)\right) \phi(z) dz}$$

Proof. First, by the law of iterated expectations,

$$\begin{aligned} P\left[\{Z(s) > u\} \cap_{t \in \mathcal{N}_{PC}(s)} \{Z(t) < Z(s)\}\right] &= \int_{-\infty}^\infty P[\{z > u\} \cap \{Z(t) < z, \forall t \in \mathcal{N}_{PC}(s) | Z(s) = z\}] \phi(z) dz \\ &= \int_{-\infty}^\infty \mathbb{1}_{\{z > u\}} \cdot P[\{Z(t) < z, \forall t \in \mathcal{N}_{PC}(s) | Z(s) = z\}] \phi(z) dz \\ &= \int_u^\infty P[\{Z(t) < z, \forall t \in \mathcal{N}_{PC}(s) | Z(s) = z\}] \phi(z) dz. \end{aligned}$$

Next, by applying Proposition 13,

$$P[\{Z(t) < z, \forall t \in \mathcal{N}_{PC}(s) | Z(s) = z\}] = P\left[\bigcap_{d=1}^D (Z(t) < z,\ \forall t = s \pm v_d e_d | Z(s) = z)\right] = \prod_{d=1}^D Q(\rho_d, z),$$

where

$$Q(\rho_d, z) = P\{Z(t) < z,\ \forall t = s \pm v_d e_d | Z(s) = z\}.$$

Thus we have

$$P\left[\{Z(s) > u\} \cap_{t \in \mathscr{N}_{PC}(s)} \{Z(t) < Z(s)\}\right] = \int_u^\infty \left(\prod_{d=1}^{D} Q(\rho_d, z)\right) \phi(z) dz.$$

By Bayes Rule,

$$\begin{aligned}
P[Z(s) > u | Z(t) < Z(s), \forall t \in \mathscr{N}_{PC}(s)] &= \frac{P[\{Z(s) > u\} \cap \{Z(t) < Z(s), \forall t \in \mathscr{N}_{PC}(s)\}]}{P[Z(t) < Z(s), \forall t \in \mathscr{N}_{PC}(s)]} \\
&= \frac{\int_u^\infty P\left[(Z(t) < z, \forall t \in \mathscr{N}_{PC}(s) | Z(x) = z)\right] \phi(z) dz}{\int_{-\infty}^\infty P\left[(Z(t) < z, \forall t \in \mathscr{N}_{PC}(s) | Z(x) = z)\right] \phi(z) dz} \\
&= \frac{\int_u^\infty \left(\prod_{d=1}^{D} Q(\rho_d, z)(z)\right) \phi(z) dz}{\int_{-\infty}^\infty \left(\prod_{d=1}^{D} Q(\rho_d, z)(z)\right) \phi(z) dz}.
\end{aligned}$$

□

Thus, the peak height density is

$$f_{\mathrm{DLM}} = \frac{\prod_{d=1}^{D} Q(\rho_d, z) \phi(z)}{\int_{-\infty}^\infty \left(\prod_{d=1}^{D} Q(\rho_d, z)\right) \phi(z) dz}.$$

Follow the proof from [87],

$$Q(\rho_d, z) = 1 - 2\Phi(h_d z^+) + \frac{1}{\pi} \int_0^{\alpha_d} \exp(-\frac{1}{2} h_d^2 z^2 / \sin^2 \theta) d\theta.$$

where

$$h_d = \sqrt{\frac{1-\rho_d}{1+\rho_d}}, \alpha_d = \sin^{-1}\left(\sqrt{(1-\rho_d^2)/2}\right), z^+ = \max(z, 0).$$

D.2 The neighbourhood covariance matrix in the fully connected setting

D.2.1 Theoretical derivation of the neighbourhood covariance for the integral convolution field

To prove (5.10), it suffices to establish following claim:

Claim: Let Z be a D dimensional isotropic Gaussian random field on a discrete lattice S with mean zero and unit variance. Assume that Z is derived by integrating continuous white noise against a Gaussian kernel. Let $s_0, t_0 \in S$, we first define a special indexing of $\mathcal{N}_{FC}(s_0)$ and $\mathcal{N}_{FC}(t_0)$, the elements of the fully connected neighbourhood of s_0 and t_0 defined in (5.3). We index the neighbours in the order of $0,1,2$ in each direction. A 2D example of $\mathcal{N}_{FC}(s_0)$ for $s_0 = (1,1)$ is given as

$$\begin{array}{ccc} (0,0) & (0,1) & (0,2) \\ (1,0) & (1,1) & (1,2) \\ (2,0) & (2,1) & (2,2). \end{array}$$

Let $s = (s_1, \ldots, s_D)^\top$ and $t = (t_1, \ldots, t_D)^\top$ be the vector that is one element inside of $\mathcal{N}_{FC}(s_0)$ and $\mathcal{N}_{FC}(t_0)$. Then

$$\operatorname{Cov}(Z(s), Z(t)) = [A \otimes A \otimes \ldots \otimes A]_{m,n}, \tag{D.1}$$

where

$$A = \begin{pmatrix} 1 & \rho & \rho^4 \\ \rho & 1 & \rho \\ \rho^4 & \rho & 1 \end{pmatrix},$$

ρ is the correlation between adjacent voxels, and $m = \sum_{i=1}^{D} 3^{i-1} s_i + 1$ and $n = \sum_{i=1}^{D} 3^{i-1} t_i + 1$.

Proof. Because of the form of Z, $\operatorname{Cov}(Z(s), Z(t)) = \rho^{||s-t||^2}$. We have that $[A \otimes A \otimes \ldots \otimes A]_{m,n} = \prod_{i=1}^{D} A_{s_i t_i}$.

If $||s_i - t_i|| = 0, A_{s_it_i} = 1 = \rho^0$; if $||s_i - t_i|| = 1, A_{s_it_i} = 1 = \rho$ and if $||s_i - t_i|| = 2, A_{s_it_i} = \rho^4$. Thus,

$$A_{s_it_i} = \rho^{||s_i - t_i||^2},$$

and hence

$$\prod_{i=1}^{D} A_{s_it_i} = \rho^{\Sigma_{i=1}^{D} ||s_i - t_i||^2} = \rho^{||s-t||^2}.$$

As such both sides of (D.1) match and the result follows.

This proof relies on a special indexing of $\mathbf{Z}$. However, this is in practice not a restriction as a re-indexing of $\mathbf{Z}$ can be treated as a linear transformation, i.e. $\mathbf{Z} = T\mathbf{Z}_o$, where $\mathbf{Z}_o$ is the vector with the special indexing. In that case Σ then can be calculated by

$$\Sigma = T\text{Cov}(\mathbf{Z}_o)T^\top.$$

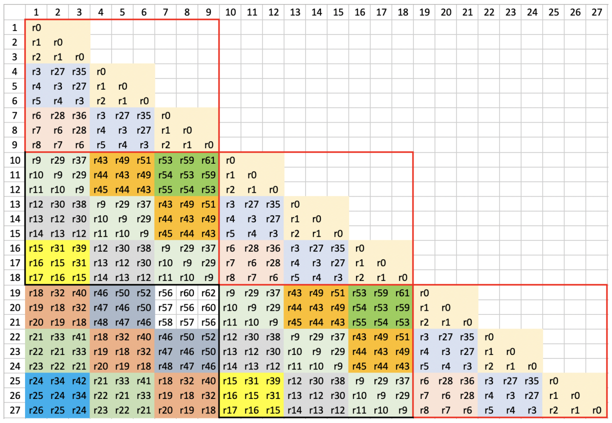
□

D.2.2 Neighbourhood covariance matrix for a 3D stationary Gaussian random field

Figure D.1 shows the theoretical covariance function for a 3D stationary field. The construction and indexing of this 27×27 matrix follow the logic from Section D.2.1. The numbers $r0, r1, ..., r61$ denote all 62 distinct values of covariance between two voxels. We use the same color to denote all 3×3 matrices with the same values. One useful conclusion from this figure is this covariance matrix is a block Toeplitz matrix with 9 blocks, and each block is still a block Toeplitz matrix with 9 sub-blocks. In addition, each sub-block is a Toeplitz matrix.

D.3 Introduction to the *pp* plot

The *pp* plots are used throughout Section 5.3 and 5.4. In this section, we will formally define this *pp* plot. *pp* plots are used to compare the peak *p*-values computed using the ADLM, MCDLM and continuous methods we discussed in Section 5.2 to the reference tail probability of the true peak



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27
1	r0																										
2	r1	r0																									
3	r2	r1	r0																								
4	r3	r27	r35	r0																							
5	r4	r3	r27	r1	r0																						
6	r5	r4	r3	r2	r1	r0																					
7	r6	r28	r36	r3	r27	r35	r0																				
8	r7	r6	r28	r4	r3	r27	r1	r0																			
9	r8	r7	r6	r5	r4	r3	r2	r1	r0																		
10	r9	r29	r37	r43	r49	r51	r53	r59	r61	r0																	
11	r10	r9	r29	r44	r43	r49	r54	r53	r59	r1	r0																
12	r11	r10	r9	r45	r44	r43	r55	r54	r53	r2	r1	r0															
13	r12	r30	r38	r9	r29	r37	r43	r49	r51	r3	r27	r35	r0														
14	r13	r12	r30	r10	r9	r29	r44	r43	r49	r4	r3	r27	r1	r0													
15	r14	r13	r12	r11	r10	r9	r45	r44	r43	r5	r4	r3	r2	r1	r0												
16	r15	r31	r39	r12	r30	r38	r9	r29	r37	r6	r28	r36	r3	r27	r35	r0											
17	r16	r15	r31	r13	r12	r30	r10	r9	r29	r7	r6	r28	r4	r3	r27	r1	r0										
18	r17	r16	r15	r14	r13	r12	r11	r10	r9	r8	r7	r6	r5	r4	r3	r2	r1	r0									
19	r18	r32	r40	r46	r50	r52	r56	r60	r62	r9	r29	r37	r43	r49	r51	r53	r59	r61	r0								
20	r19	r18	r32	r47	r46	r50	r57	r56	r60	r10	r9	r29	r44	r43	r49	r54	r53	r59	r1	r0							
21	r20	r19	r18	r48	r47	r46	r58	r57	r56	r11	r10	r9	r45	r44	r43	r55	r54	r53	r2	r1	r0						
22	r21	r33	r41	r18	r32	r40	r46	r50	r52	r12	r30	r38	r9	r29	r37	r43	r49	r51	r3	r27	r35	r0					
23	r22	r21	r33	r19	r18	r32	r47	r46	r50	r13	r12	r30	r10	r9	r29	r44	r43	r49	r4	r3	r27	r1	r0				
24	r23	r22	r21	r20	r19	r18	r48	r47	r46	r14	r13	r12	r11	r10	r9	r45	r44	r43	r5	r4	r3	r2	r1	r0			
25	r24	r34	r42	r21	r33	r41	r18	r32	r40	r15	r31	r39	r12	r30	r38	r9	r29	r37	r6	r28	r36	r3	r27	r35	r0		
26	r25	r24	r34	r22	r21	r33	r19	r18	r32	r16	r15	r31	r13	r12	r30	r10	r9	r29	r7	r6	r28	r4	r3	r27	r1	r0	
27	r26	r25	r24	r23	r22	r21	r20	r19	r18	r17	r16	r15	r14	r13	r12	r11	r10	r9	r8	r7	r6	r5	r4	r3	r2	r1	r0

Figure D.1. The theoretical covariance function for a 3D stationary field.

height distribution, i.e. the reference *p*-value. We obtain the latter from simulation. To do so, first we generate N i.i.d. random fields as described in each setting in Section 5.3. Let n be the number of obtained local maxima across all fields, where the local maxima are selected based on the criteria that their height values are larger than all their neighbours in the specified neighbourhood (we will consider both the fully connected and partially connected neighbourhoods). Let $g_1, \dots, g_n$ be the heights of the recorded local maxima. For each peak this allows us to compute an estimation of the reference *p*-value as

$$p_i = \frac{1}{n}\sum_{j=1}^{n} \mathbb{1}[g_j > g_i], \quad 1 \le i \le n, \tag{D.2}$$

where $\mathbb{1}[\cdot]$ denotes the indicator function, p_i is the *p*-value when observed value is g_i. As $n \to \infty$, p_i calculated by (D.2) converges to the true tail probability. Moreover for each peak, we calculate a *p*-value for each of the three approaches. Next we plot the reference *p*-values against the *p*-values obtained using each method. Since the reference distribution converges to the true peak height distribution as the number of instances converges to infinity, the closer these plots are to the identity function, the closer the approximation to the true distribution. We use these *pp* plots to compare the performance of the three approaches in all of our simulation studies.

The idea of this *pp* plot is similar to the one used in [80]. Although the two plotting mechanisms look different, the logic behind is the same, as we justify below.

Our two plotting mechanisms are

- Plot $p(i)$ vs. i/n and $q(i)$ vs. i/n ([80]),

- Plot $p(i)$ vs. $p(i)$ and $q(i)$ vs. $p(i)$ (pp plot).

Let $F(z)$ be true cdf of z, $G(z)$ be one of the other cdfs of z used for comparison purpose. Next, we define

$$p = 1 - F(z)$$

$$q = 1 - G(z)$$

Suppose that we now generate n p-values for both the true distribution and the distribution for comparison purpose, i.e., we generate $p_1, ..., p_n$ and $q_1, ..., q_n$ as in (D.2). Denote $p(i)$ and $q(i)$ as the order statistics of $p_1, ..., p_n$ and $q_1, ..., q_n$. Then under monotonicity,

$$\begin{aligned} \text{ecdf}(p(i)) &= \frac{1}{n}\sum_{i=1}^{n} I[p_i \le p(i)] = \frac{i}{n} \\ \text{ecdf}(q(i)) &= \frac{1}{n}\sum_{i=1}^{n} I[q_i \le q(i)] = \frac{i}{n} \end{aligned}$$

If z is distributed according to F, p-values are uniformly distributed on $[0,1]$. Thus, by the LLN,

$$\text{ecdf}(p) = \frac{1}{n}\sum_{i=1}^{n} I(p_i \le p) \rightarrow_p P(p_i \le p) = p$$

Thus, when p is from the true distribution and n is large enough, we expect the two plotting mechanisms provide similar plots. The comparison of two method is shown below in Figure D.2.

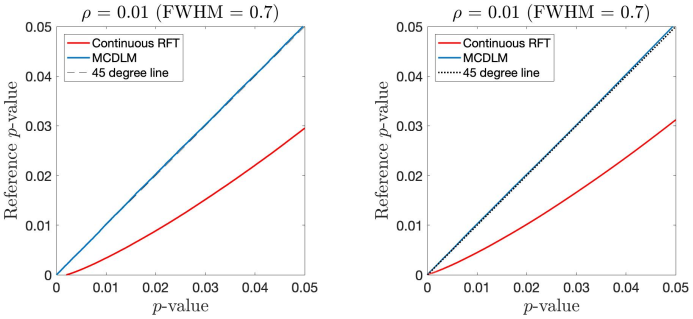


Figure D.2. Comparison of two plotting mechanism using the height distribution generated from 2D isotropic Gaussian field with spatial correlation $\rho = 0.01$. Left is from the plotting mechanism in [80] and right is from the plotting mechanism in this paper

.

D.4 Additional results

D.4.1 Partial connectivity case

In Section 5.2 we discussed some of the disadvantages of ADLM. In particular local separability (i.e. the result of Proposition 13) does not apply to diagonal neighbours and so ADLM cannot be used for peaks of the fully connected neighbourhood. However (when its assumptions hold) it can be used to infer on the the peak height of peaks in the partially connected neighbourhood defined in ((5.5)). We illustrate this here with in the same simulation setting of Section 5.3.1.

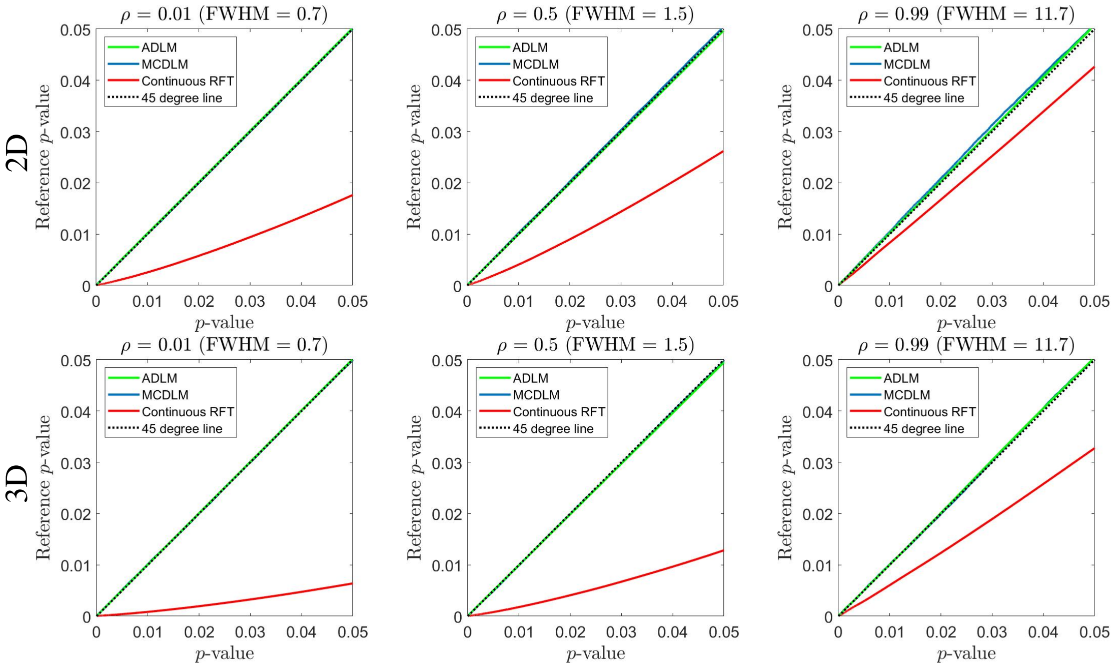


Figure D.3. Comparison of the peak height distribution for peaks in a partially connected neighbourhood calculated via the different methods for 2D and 3D isotropic Gaussian fields.

Figure D.3 shows the comparison between MCDLM, ADLM and the continuous RFT method in 2D and 3D. In all three scenarios, the p-value distribution of ADLM and MCDLM match and are close to a uniform distribution. As in the main text the continuous RFT approach is conservative.

D.4.2 Applying the neighborhood covariance function in (5.7)

In this section we perform the same simulations as in Section 5.3.1 but with the neighbourhood covariance of [101] (given in (5.7)) instead of the actual neighbourhood covariance (which we derived in

(5.9).

The results are shown in Figure D.4. They are similar to those of Figure 5.4. However, as exemplified in the $\rho = 0.5$ case the MCDLM approach is incorrect when this covariance function is used. This is because it is not in fact the correct neighbourhood covariance.

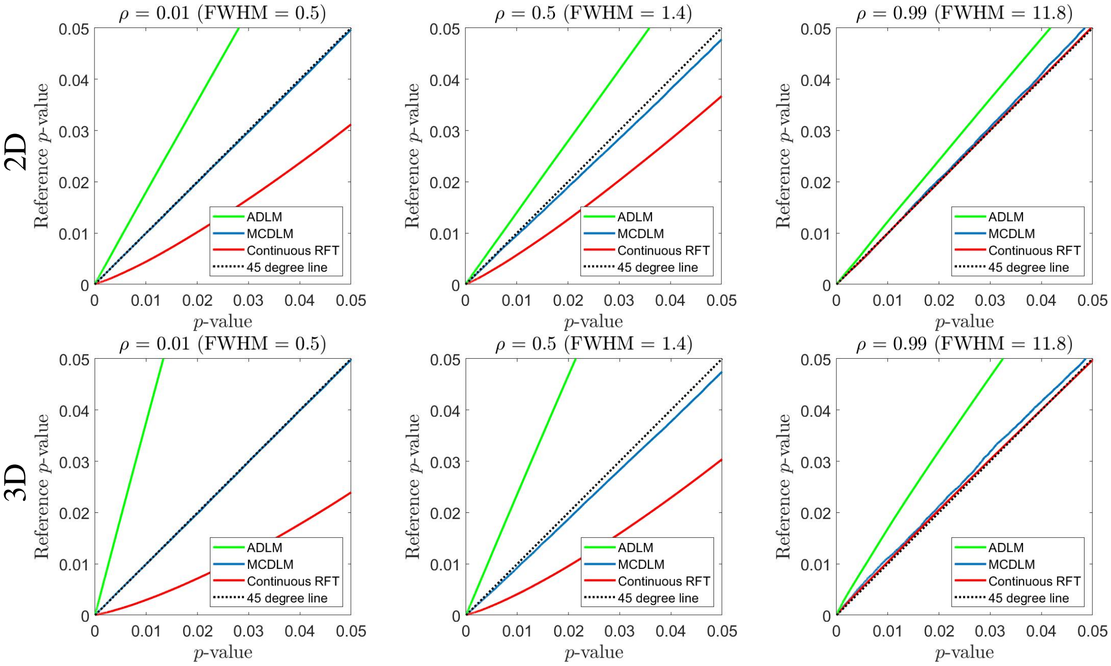


Figure D.4. Comparison of peak height distribution calculated from different methods for 2D and 3D isotropic Gaussian field using neighborhood covariance function in (5.7).

D.4.3 Applying the look-up table

Since our method is Monte Carlo based method, it is desirable to reduce the computation time in certain settings. We list the computation time under different scenarios in Appendix D.5. In order to cut the time of running the simulation each time for a different height threshold u and spatial correlation ρ, it is possible to use a look-up table to pre-record the simulation results for 10^5 possible local maxima height values at different values of ρ when the Gaussian random field is isotropic. To do so we vary ρ form 0.01 to 0.99 with increments of 0.01. In order to calculate a look up table we do the following.

1. Loop through the array of different values of ρ and obtain 10^5 local maxima for each ρ.

2. From the union obtained in step 1, sample 10^5 local maxima, which consist a set of local maxima,

u, that we want to evaluate the CDF, $F(\cdot)$, at.

3. Loop through the array of different values of ρ again. For each value of ρ, interpolate $F(\cdot)$ at each of u we get in step 2. Record all the $F(u;\rho)$ in a matrix of look-up table with row the ρ and column the u.

To evaluate the p-value at a given threshold u and correlation ρ, we interpolate $F(\cdot)$ at u and ρ through pre-recorded look-up table, and the p-value is calculated by $1 - F(u;\rho)$.

After generating a look-up table, we apply the cubic spline smoothing to smooth the noisy look-up table. The procedure of our smoothing is as follow:

1. Use the Cubic spline smoothing to smooth the matrix across ρ;
2. Use the Cubic spline smoothing to smooth the matrix we smoothed in step 1 across u.

In doing so, we aim to reduce the violation of monotonicity across u but also retain the smoothness. The smoothing parameters in Cubic spline smoothing is selected by 5-fold cross validation in each scenario separately.

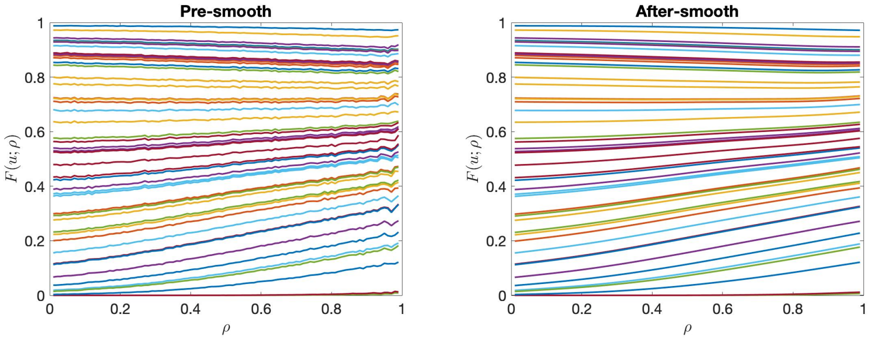


Figure D.5. This figure shows $F(u;\rho)$ of selected 50 samples across ρ from pre-smooth table (left) and after-smooth table (right). Same color is used for the same sample before and after-smoothing.

In Figure D.5, we select 50 columns of both pre-smooth and after-smooth look-up table generated from 3D Gaussian random field and plot the CDF across different ρ values. Different samples selected within one look-up table are denoted by different colors, while same color means same sample between two tables. The pre-smooth plot shows that the look-up table we generated is very noisy across ρ,

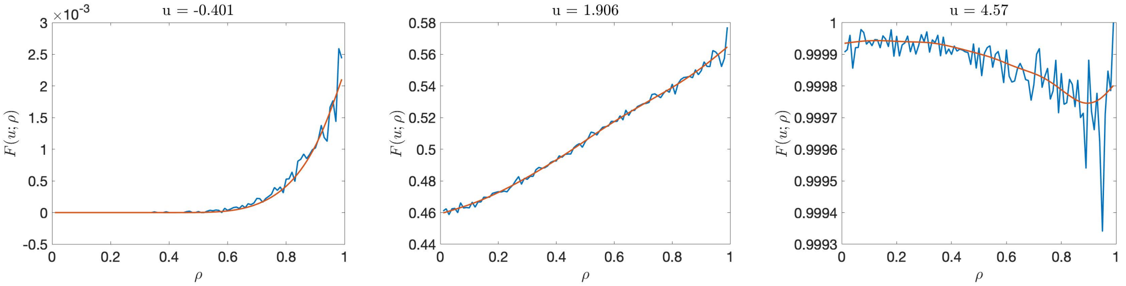


Figure D.6. From left to right are the 10, 50000 and 99990 column of the pre-smooth table (red) and after-smooth table (blue).

especially at some large values of ρ. This will cause big problems in computing p-values for different values of ρ. For example, if two are interested in calculating the p-values for a ρ close to 1, a slightly difference in picking the ρ will bring a significant change in p-values, and such ambiguity could finally result in an inconsistent interpretation of the scientific findings. From the after-smooth plot, we observe that the look-up table is smooth enough to provide consistent results.

In addition, we explore to what extent the smoothing works in figure D.6. In this figure, we select three columns of both pre-smooth table and after-smooth table and then compare. The u that we select are -0.401, 1.906 and 4.57, which are in both ends and middle of the support. From these three plots, the noise before the smoothing is obvious, but the cubic spline smoothing fits the curve perfectly, in the sense that it preserves the general shape of the curve yet removes the noise.

D.4.4 Applying Gaussianization transformation of the t-fields

As discussed in Section 5.3.2, to improve the computation efficiency for the height distribution of peaks of t-fields, we consider using Gaussianization transformation of the t-fields. In this section we perform the same simulations as in Section 5.3.2 but the simulated t-fields were Gaussianized by (5.15) and compared with the MCDLM for Gaussian field and continuous RFT approach.

The results are shown in Figure D.7 (2D) and Figure D.8 (3D). MCDLM works well when degrees of freedom is large and ρ is small. At high smoothness, MCDLM and continuous RFT work similarly. They are only correct when degrees of freedom is large since in this case the t-fields can approximate Gaussian fields.

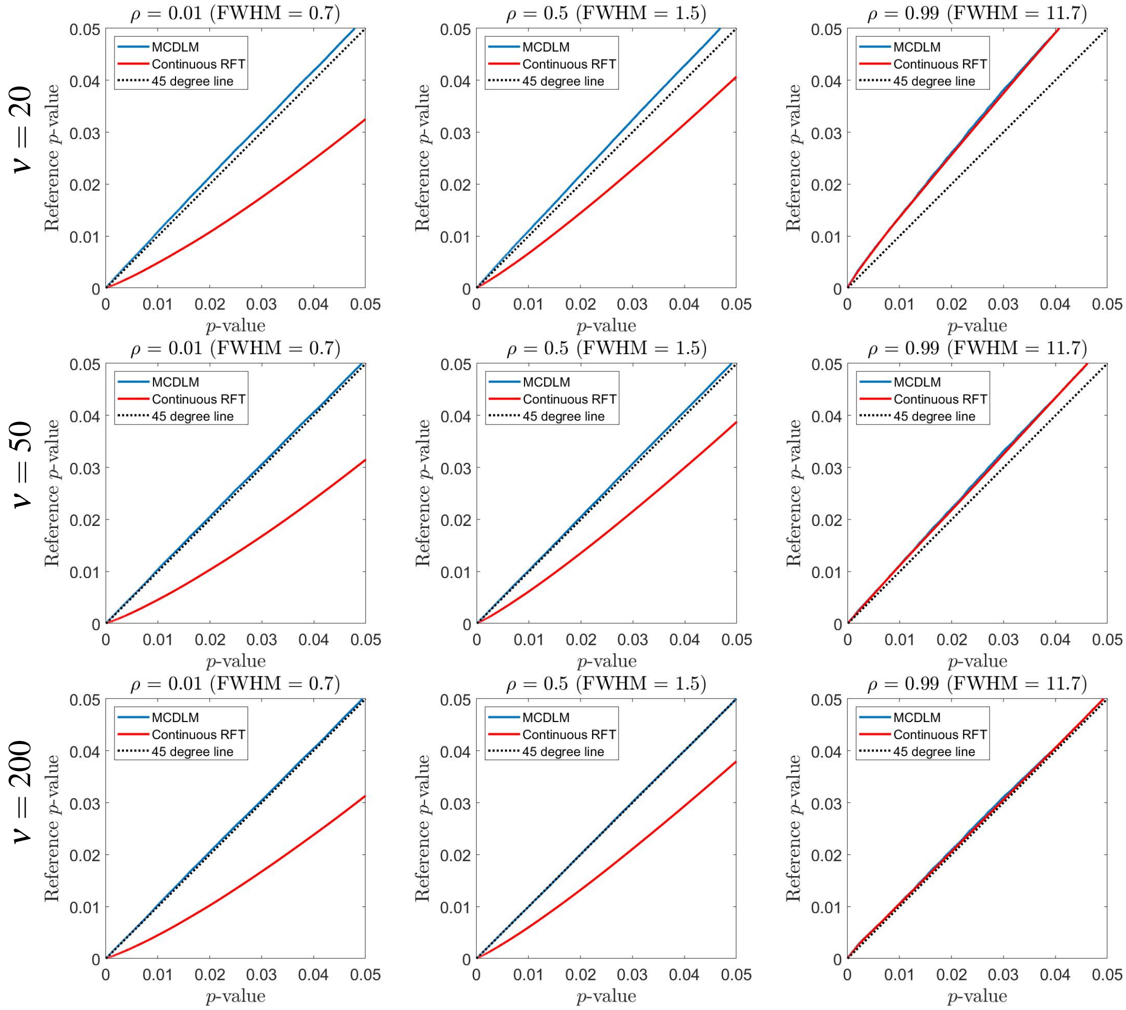


Figure D.7. Comparison methods for calculating the peak height distribution of a Gaussianized 2D t-field with ν degrees of freedom.

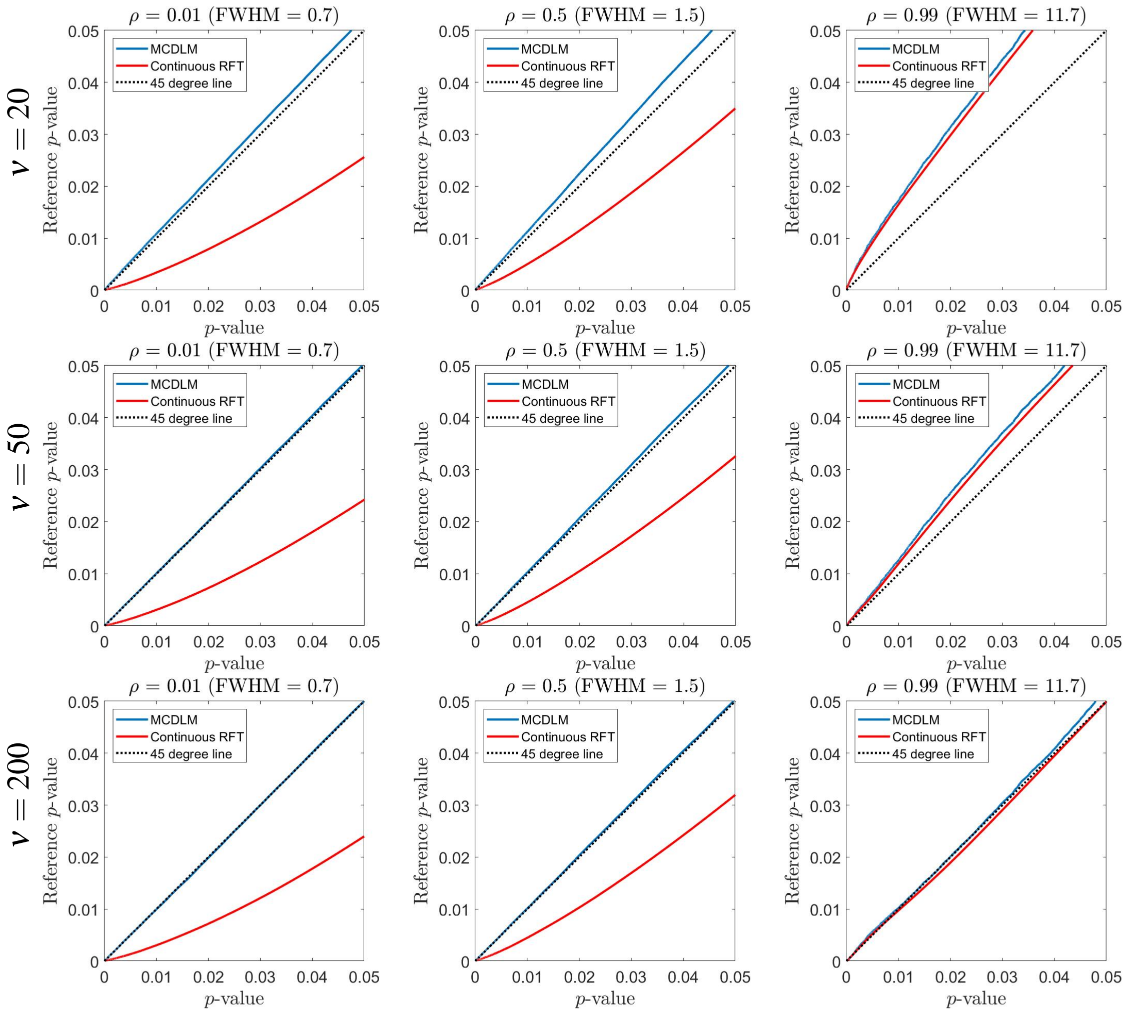


Figure D.8. Comparison methods for calculating the peak height distribution of a Gaussianized 3D t-field with ν degrees of freedom.

D.4.5 Additional simulation results for MCDLM based on estimated neighborhood covariance

In Section 5.4.3 we discuss that the performance of MCDLM is worsen as ρ increases to 0.99. Since we observe when $\rho = 0.5$ both 2D and 3D have good performance, we consider exploring 3 more scenarios in between, $\rho = 0.9, 0.93, 0.95$. In Figure D.9, when $\rho < 0.95$, or FWHM < 5.2, MCDLM with estimated covariance works well for both 2D and 3D cases with enough number of random fields to estimate the covariance function (nsim ≥ 50). By comparing the neighborhood covariance matrices (D.3) (theoretical covariance) and (D.4) (estimated covariance) for $\rho = 0.99$ in 2D case, the estimation is very close to the true covariance. However, the covariance matrix in this case is nearly singular, which causes problems when applied to our MCDLM method.

$$\begin{bmatrix}
1.0000 & 0.9900 & 0.9606 & 0.9900 & 0.9801 & 0.9510 & 0.9606 & 0.9510 & 0.9227 \\
0.9900 & 1.0000 & 0.9900 & 0.9801 & 0.9900 & 0.9801 & 0.9510 & 0.9606 & 0.9510 \\
0.9606 & 0.9900 & 1.0000 & 0.9510 & 0.9801 & 0.9900 & 0.9227 & 0.9510 & 0.9606 \\
0.9900 & 0.9801 & 0.9510 & 1.0000 & 0.9900 & 0.9606 & 0.9900 & 0.9801 & 0.9510 \\
0.9801 & 0.9900 & 0.9801 & 0.9900 & 1.0000 & 0.9900 & 0.9801 & 0.9900 & 0.9801 \\
0.9510 & 0.9801 & 0.9900 & 0.9606 & 0.9900 & 1.0000 & 0.9510 & 0.9801 & 0.9900 \\
0.9606 & 0.9510 & 0.9227 & 0.9900 & 0.9801 & 0.9510 & 1.0000 & 0.9900 & 0.9606 \\
0.9510 & 0.9606 & 0.9510 & 0.9801 & 0.9900 & 0.9801 & 0.9900 & 1.0000 & 0.9900 \\
0.9227 & 0.9510 & 0.9606 & 0.9510 & 0.9801 & 0.9900 & 0.9606 & 0.9900 & 1.0000
\end{bmatrix} \tag{D.3}$$

$$\begin{bmatrix}
1.0000 & 0.9906 & 0.9612 & 0.9903 & 0.9809 & 0.9516 & 0.9606 & 0.9513 & 0.9228 \\
0.9906 & 1.0000 & 0.9906 & 0.9811 & 0.9903 & 0.9809 & 0.9518 & 0.9606 & 0.9513 \\
0.9612 & 0.9906 & 1.0000 & 0.9521 & 0.9811 & 0.9903 & 0.9237 & 0.9518 & 0.9606 \\
0.9903 & 0.9811 & 0.9521 & 1.0000 & 0.9906 & 0.9612 & 0.9903 & 0.9809 & 0.9516 \\
0.9809 & 0.9903 & 0.9811 & 0.9906 & 1.0000 & 0.9906 & 0.9811 & 0.9903 & 0.9809 \\
0.9516 & 0.9809 & 0.9903 & 0.9612 & 0.9906 & 1.0000 & 0.9521 & 0.9811 & 0.9903 \\
0.9606 & 0.9518 & 0.9237 & 0.9903 & 0.9811 & 0.9521 & 1.0000 & 0.9906 & 0.9612 \\
0.9513 & 0.9606 & 0.9518 & 0.9809 & 0.9903 & 0.9811 & 0.9906 & 1.0000 & 0.9906 \\
0.9228 & 0.9513 & 0.9606 & 0.9516 & 0.9809 & 0.9903 & 0.9612 & 0.9906 & 1.0000
\end{bmatrix} \tag{D.4}$$

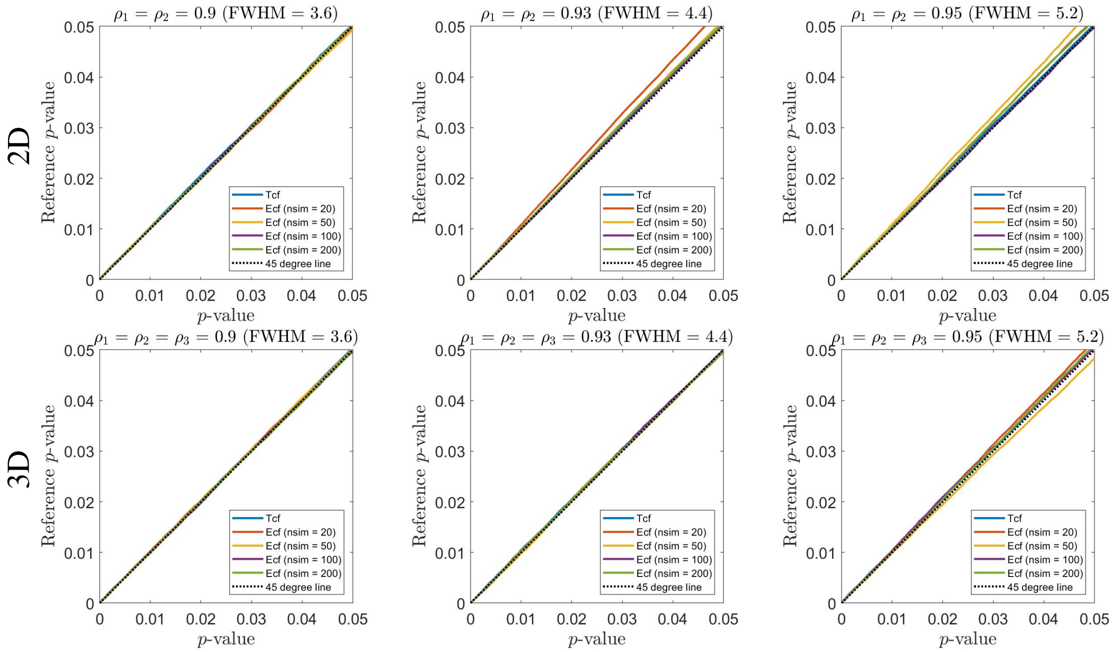


Figure D.9. Comparison of the peak height distribution calculated from using MCDLM with different neighborhood covariance for 2D and 3D isotropic Gaussian fields. The covariance functions used here are theoretical covariance function (Tcf) and empirically estimated covariance function (Ecf). The number of random fields used to estimate the covariance function is denoted using nsim. From left to right: $\rho = 0.9, 0.93, 0.95$.

D.5 Computation time table

The simulation time under different scenarios are shown in Table D.1. In the case of 2D Gaussian fields, applying MCDLM with full or partial connectivity is comparable in terms of running time, but the look up table is much faster (between 5 and 50 times faster), getting more efficient as the field correlation increases. In the case of 3D Gaussian fields, running times are 5 to 10 times larger than in 2D. However, the look up table method is just as fast as in 2D, making its use very worthwhile from a computational standpoint. In the case of 2D *t*-field, the running times increase about 10 to 15 times from degrees of freedom equals 20 to 200. In the case of 3D *t*-field, the running times are 5 to 10 times larger than in 2D, making the code stop at a pre-specified threshold when degrees of freedom increases to 200. The MCDLM with empirical covariance function has similar running times to MCDLM with theoretical covariance function, in the case of 2D Gaussian field with the number of fields to estimate the covariance small (200 or 1000). In 3D case, when the number of fields to estimate the covariance is 200, the running times are 2 times larger than the case of applying the theoretical covariance function. When the number of fields increases to 1000, the time to estimate the covariance further increases, leading to the running times 4-8 times larger than the case of applying the theoretical covariance function.

Table D.1. Running time (in seconds) of our MCDLM method under different scenarios. For $\rho = 0.01$ and $\rho = 0.05$, the code is targeted to obtain at least $n = 1e6$ peak height values and for $\rho = 0.99$, $n = 2e5$ peak height values. In some extreme cases, the code stops at a pre-specified threshold with the number of instances generated recorded in parentheses.

	$\rho = 0.01$ ($n = 1e6$)	$\rho = 0.5$ ($n = 1e6$)	$\rho = 0.99$ ($n = 2e5$)
2D Gaussian field			
Full connectivity (continuous covariance function)	9.81	13.83	106.41
Full connectivity (discrete covariance function)	9.95	14.76	112.24
Partial connectivity (discrete covariance function)	7.43	12.15	110.86
Full connectivity (look up table)	1.67	1.77	1.88
2D t-field			
$\nu = 20$	74.48	105.74	835.80
$\nu = 50$	217.63	307.56	1992.58
$\nu = 200$	1064.24	1356.55	1646.86 ($n = 37289$)
3D Gaussian field			
Full connectivity (continuous covariance function)	41.79	64.16	1395.14 ($n = 1e5$)
Full connectivity (discrete covariance function)	41.79	66.32	1377.69 ($n = 1e5$)
Partial connectivity (discrete covariance function)	13.87	29.18	1753.71
Full connectivity (look up table)	1.56	2.04	1.76
3D t-field			
$\nu = 20$	659.09	1043.29	2199.84 ($n = 11614$)
$\nu = 50$	1719.52	2755.58	2611.89 ($n = 11244$)
$\nu = 200$	5325.24 ($n = 74403$)	5159.63 ($n = 46409$)	11914.56 ($n = 402$)
2D isotropic Gaussian field (empirical covariance case)			
number of fields = 200	9.56	15.13	208.33
number of fields = 1000	9.91	14.19	145.04
number of fields = 10,000	14.56	18.05	119.83
3D isotropic Gaussian field (empirical covariance case)			
number of fields = 200	82.94	105.66	1412.21
number of fields = 1000	320.91	290.76	1722.92

Bibliography

[1] Omar M Abdeldayem, Areeg M Dabbish, Mahmoud M Habashy, Mohamed K Mostafa, Mohamed Elhefnawy, Lobna Amin, Eslam G Al-Sakkari, Ahmed Ragab, and Eldon R Rene. Viral outbreaks detection and surveillance using wastewater-based epidemiology, viral air sampling, and machine learning techniques: A comprehensive review and outlook. *Science of The Total Environment*, 803:149834, 2022.

[2] Robert J Adler. The geometry of random fields, vol. 62. *SIAM, Philadelphia*, 1981.

[3] Shelesh Agrawal, Laura Orschler, and Susanne Lackner. Long-term monitoring of sars-cov-2 rna in wastewater of the frankfurt metropolitan area in southern germany. *Scientific reports*, 11(1):1–7, 2021.

[4] Warish Ahmed, Nicola Angel, Janette Edson, Kyle Bibby, Aaron Bivins, Jake W O'Brien, Phil M Choi, Masaaki Kitajima, Stuart L Simpson, Jiaying Li, Ben Tscharke, Rory Verhagen, Smith Wendy JM, Julian Zaugg, Leanne Dierens, Philip Hugenholtz, Kevin V Thomas, and Jochen F Mueller. First confirmed detection of sars-cov-2 in untreated wastewater in australia: a proof of concept for the wastewater surveillance of covid-19 in the community. *Science of the Total Environment*, 728:138764, 2020.

[5] Yuehan Ai, Fan He, Emma Lancaster, and Jiyoung Lee. Application of machine learning for multi-community covid-19 outbreak predictions with wastewater surveillance. *Plos one*, 17(11):e0277154, 2022.

[6] Elizabeth J Atkinson and Terry M Therneau. An introduction to recursive partitioning using the rpart routines. *Rochester: Mayo Foundation*, 2000, 2000.

[7] Itay Bar-Or, Karin Yaniv, Marilou Shagan, Eden Ozer, Merav Weil, Victoria Indenbaum, Michal Elul, Oran Erster, Ella Mendelson, Batya Mannasse, Rachel Shirazi, Esti Kramarsky-Winter, Oded Nir, Hala Abu-Ali, Zeev Ronen, Ehud Rinott, Yair E Lewis, Eran Friedler, Eden Bitkover, Yossi Paitan, Yakir Berchenko, and Ariel Kushmaro. Regressing SARS-CoV-2 sewage measurements onto COVID-19 burden in the population: A Proof-of-Concept for quantitative environmental surveillance. *Front Public Health*, 9:561710, January 2022.

[8] Daniel Barich and Joan L Slonczewski. Wastewater virus detection complements clinical covid-19 testing to limit spread of infection at kenyon college. *medRxiv*, pages 2021–01, 2021.

[9] Michael P Battaglia, David C Hoaglin, and Martin R Frankel. Practical considerations in raking survey data. *Survey Practice*, 2(5), 2009.

[10] Gérard Biau, Luc Devroye, and Gäbor Lugosi. Consistency of random forests and other averaging classifiers. *Journal of Machine Learning Research*, 9(9), 2008.

[11] Aaron Bivins and Kyle Bibby. Wastewater surveillance during mass covid-19 vaccination on a college campus. *Environmental Science & Technology Letters*, 8(9):792–798, 2021.

[12] Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001.

[13] Leo Breiman. *Classification and regression trees*. Routledge, 2017.

[14] Leo Breiman, Jerome Friedman, Richard Olshen, and Charles Stone. Cart. *Classification and regression trees*, 1984.

[15] H. Buschmanm and S. LaFee. Introducing the uc san diego return to learn program. uc san diego new center. https://health.ucsd.edu/news/releases/Pages/2020-05-05-introducing-uc-san-diego-return-to-learn-program.aspx, May 2020.

[16] Yongtao Cao and Roland Francis. On forecasting the community-level covid-19 cases from the concentration of sars-cov-2 in wastewater. *Science of The Total Environment*, 786:147451, 2021.

[17] Nicole Bohme Carnegie, Rui Wang, Vladimir Novitsky, and Victor De Gruttola. Linkage of viral sequences among hiv-infected village residents in botswana: estimation of linkage rates in the presence of missing data. *PLoS computational biology*, 10(1):e1003430, 2014.

[18] Nitesh V Chawla. Data mining for imbalanced datasets: An overview. *Data mining and knowledge discovery handbook*, pages 875–886, 2010.

[19] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[20] Tian Chen, Jeanne Kowalski, Rui Chen, Pan Wu, Hui Zhang, Changyong Feng, and Xin M Tu. Rank-preserving regression: a more robust rank regression model against outliers. *Statistics in Medicine*, 35(19):3333–3346, 2016.

[21] Dan Cheng and Armin Schwartzman. Distribution of the height of local maxima of gaussian random fields. *Extremes*, 18(2):213–240, 2015.

[22] Dan Cheng and Armin Schwartzman. On the explicit height distribution and expected number of local maxima of isotropic gaussian random fields. *arXiv preprint arXiv:1503.01328*, 2015.

[23] Dan Cheng and Armin Schwartzman. Multiple testing of local maxima for detection of peaks in random fields. *Annals of statistics*, 45(2):529, 2017.

[24] Dan Cheng and Armin Schwartzman. On critical points of gaussian random fields under diffeomorphic transformations. *Statistics & Probability Letters*, 158:108672, 2020.

[25] Justin Chumbley, Keith Worsley, Guillaume Flandin, and Karl Friston. Topological FDR for neuroimaging. *Neuroimage*, 49(4):3057–3064, 2010.

[26] Justin R Chumbley and Karl J Friston. False discovery rate revisited: FDR and topological inference using gaussian random fields. *Neuroimage*, 44(1):62–70, 2009.

[27] William G Cochran. *Sampling techniques*. John Wiley & Sons, 2007.

[28] Saskia Comess, Hannah Wang, Susan Holmes, and Claire Donnat. Statistical modeling for practical pooled testing during the covid-19 pandemic. *Statistical Science*, 37(2):229–250, 2022.

[29] Pascal Coorevits, Mats Sundgren, Gunnar O Klein, Anne Bahr, Brecht Claerhout, Christel Daniel, Martin Dugas, Danielle Dupont, Andreas Schmidt, Peter Singleton, Georges De Moor, and Dipak Kalra. Electronic health records: new opportunities for clinical research. *Journal of internal medicine*, 274(6):547–560, 2013.

[30] Martin R Cowie, Juuso I Blomster, Lesley H Curtis, Sylvie Duclaux, Ian Ford, Fleur Fritz, Samantha Goldman, Salim Janmohamed, Jörg Kreuzer, Mark Leenay, Alexander Michel, Seleen Ong, Jill P Pell, Southworth Mary R, Stough Wendy G, Martin Thoenes, Faiez Zannad, and Andrew Zalewski. Electronic health records to facilitate clinical research. *Clinical Research in Cardiology*, 106:1–9, 2017.

[31] Sharoda Dasgupta, Anne Marie France, Mary-Grace Brandt, Jennifer Reuer, Tianchi Zhang, Nivedha Panneer, Angela L Hernandez, and Alexandra M Oster. Estimating effects of hiv sequencing data completeness on transmission network patterns and detection of growing hiv transmission clusters, 4 2019.

[32] Christian G Daughton. Wastewater surveillance for population-wide covid-19: The present and future. *Science of the Total Environment*, 736:139631, 2020.

[33] Samuel Davenport and Thomas E Nichols. Selective peak inference: Unbiased estimation of raw and standardized effect size at local maxima. *Neuroimage*, 209:116375, 2020.

[34] Andreas Deckert, Till Bärnighausen, and Nicholas NA Kyei. Simulation of pooled-sample analysis strategies for covid-19 mass testing. *Bulletin of the World Health Organization*, 98(9):590, 2020.

[35] George W Divine, H James Norton, Anna E Barón, and Elizabeth Juarez-Colunga. The wilcoxon–mann–whitney procedure fails as a test of medians. *The American Statistician*, 72(3):278–286, 2018.

[36] Anders Eklund, Hans Knutsson, and Thomas E Nichols. Cluster failure revisited: Impact of first level design and physiological noise on cluster false positive rates. *Human brain mapping*, 40(7):2017–2032, 2019.

[37] Anders Eklund, Thomas E Nichols, and Hans Knutsson. Cluster failure: Why fmri inferences for spatial extent have inflated false-positive rates. *Proceedings of the national academy of sciences*, 113(28):7900–7905, 2016.

[38] Abdulrahman M El-Sayed, Peter Scarborough, Lars Seemann, and Sandro Galea. Social network analysis and agent-based modeling in social epidemiology. *Epidemiologic Perspectives & Innovations*, 9(1):1, 2012.

[39] Anthony S. Fauci, Robert R. Redfield, George Sigounas, Michael D. Weahkee, and Brett P. Giroir. Ending the HIV Epidemic: A Plan for the United States. *JAMA*, 321(9):844–845, 03 2019.

[40] Tom Fawcett. An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874, 2006.

[41] Mehrdad Fazli and Heman Shakeri. Leveraging wastewater monitoring for covid-19 forecasting in the us: a deep learning study. *arXiv preprint arXiv:2212.08798*, 2022.

[42] Centers for Disease Control and Prevention (CDC). National health and nutrition examination survey data. *Hyattsville, MD: US Department of Health and Human Services, Centers for Disease Control and Prevention*, 2020, 2010.

[43] KJ Friston, RE Passingham, JG Nutt, JD Heather, GV Sawle, and RSJ Frackowiak. Localisation in PET images: direct fitting of the intercommissural (AC—PC) line. *Journal of Cerebral Blood Flow & Metabolism*, 9(5):690–695, 1989.

[44] Joseph L Gastwirth. The estimation of the lorenz curve and gini index. *The review of economics and statistics*, pages 306–316, 1972.

[45] Pritha Guha, Apratim Guha, and Tathagata Bandyopadhyay. Application of pooled testing in estimating the prevalence of covid-19. *Health Services and Outcomes Research Methodology*, 22(2):163–191, 2022.

[46] Sasha Harris-Lovett, Kara L. Nelson, Paloma Beamer, Heather N. Bischel, Aaron Bivins, Andrea Bruder, Caitlyn Butler, Todd D. Camenisch, Susan K. De Long, Smruthi Karthikeyan, David A. Larsen, Katherine Meierdiercks, Paula J. Mouser, Sheree Pagsuyoin, Sarah M. Prasek, Tyler S. Radniecki, Jeffrey L. Ram, D. Keith Roper, Hannah Safford, Samendra P. Sherchan, William Shuster, Thibault Stalder, Robert T. Wheeler, and Katrina Smith Korfmacher. Wastewater surveillance for sars-cov-2 on college campuses: Initial efforts, lessons learned, and research needs. *International Journal of Environmental Research and Public Health*, 18(9), 2021.

[47] Olga E Hart and Rolf U Halden. Computational analysis of sars-cov-2/covid-19 surveillance by wastewater-based epidemiology locally and globally: Feasibility, economy, opportunities and challenges. *Science of the Total Environment*, 730:138875, 2020.

[48] Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

[49] Daniel G Horvitz and Donovan J Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, 47(260):663–685, 1952.

[50] Guangming Jiang, Jiangping Wu, Jennifer Weidhaas, Xuan Li, Yan Chen, Jochen Mueller, Jiaying Li, Manish Kumar, Xu Zhou, Sudipti Arora, Eiji Haramoto, Samendra Sherchan, Gorka Orive, Unax Lertxundi, Ryo Honda, Masaaki Kitajima, and Greg Jackson. Artificial neural network-based estimation of covid-19 case numbers and effective reproduction rate using wastewater-based epidemiology. *Water Research*, 218:118451, 2022.

[51] Smruthi Karthikeyan, Joshua I. Levy, Peter De Hoff, Greg Humphrey, Amanda Birmingham, Kristen Jepsen, Sawyer Farmer, Helena M. Tubb, Tommy Valles, Caitlin E. Tribelhorn, Rebecca Tsai, Stefan Aigner, Shashank Sathe, Niema Moshiri, Benjamin Henson, Adam M. Mark, Abbas Hakim, Nathan A. Baer, Tom Barber, Pedro Belda-Ferre, Marisol Chacón, Willi Cheung, Evelyn S. Cresini, Emily R. Eisner, Alma L. Lastrella, Elijah S. Lawrence, Clarisse A. Marotz, Toan T. Ngo, Tyler Ostrander, Ashley Plascencia, Rodolfo A. Salido, Phoebe Seaver, Elizabeth W. Smoot, Daniel McDonald, Robert M. Neuhard, Angela L. Scioscia, Alysson M. Satterlund, Elizabeth H. Simmons, Dismas B. Abelman, David Brenner, Judith C. Bruner, Anne Buckley, Michael Ellison, Jeffrey Gattas, Steven L. Gonias, Matt Hale, Faith Hawkins, Lydia Ikeda, Hemlata Jhaveri, Ted Johnson, Vince Kellen, Brendan Kremer, Gary Matthews, Ronald W. McLawhon, Pierre Ouillet, Daniel Park, Allorah Pradenas, Sharon Reed, Lindsay Riggs, Alison Sanders, Bradley Sollenberger, Angela Song, Benjamin White, Terri Winbush, Christine M. Aceves, Catelyn Anderson, Karthik Gangavarapu, Emory Hufbauer, Ezra Kurzban, Justin Lee, Nathaniel L. Matteson, Edyth Parker, Sarah A. Perkins, Karthik S. Ramesh, Refugio Robles-Sikisaka, Madison A. Schwab, Emily Spencer, Shirlee Wohl, Laura Nicholson, Ian H. McHardy, David P. Dimmock, Charlotte A. Hobbs, Omid Bakhtar, Aaron Harding, Art Mendoza, Alexandre Bolze, David Becker, Elizabeth T. Cirulli, Magnus Isaksson, Kelly M. Schiabor Barrett, Nicole L. Washington, John D. Malone, Ashleigh Murphy Schafer, Nikos Gurfield, Sarah Stous, Rebecca Fielding-Miller, Richard S. Garfein, Tommi Gaines, Cheryl Anderson, Natasha K. Martin, Robert Schooley, Brett Austin, Duncan R. MacCannell, Stephen F. Kingsmore, William Lee, Seema Shah, Eric McDonald, Alexander T. Yu, Mark Zeller, Kathleen M. Fisch, Christopher Longhurst, Patty Maysent, David Pride, Pradeep K. Khosla, Louise C. Laurent, Gene W. Yeo, Kristian G. Andersen, and Rob Knight. Wastewater sequencing reveals early cryptic sars-cov-2 variant transmission. *Nature*, 609(7925):101–108, Sep 2022.

[52] Smruthi Karthikeyan, Andrew Nguyen, Daniel McDonald, Yijian Zong, Nancy Ronquillo, Junting Ren, Jingjing Zou, Sawyer Farmer, Greg Humphrey, Diana Henderson, Tara Javidi, Karen Messer, Cheryl Anderson, Robert Schooley, Natasha K. Martin, and Rob Knight. Rapid, large-scale wastewater surveillance and automated reporting system enable early detection of nearly 85% of covid-19 cases on a university campus. *mSystems*, 6(4):10.1128/msystems.00793–21, 2021.

[53] Smruthi Karthikeyan, Nancy Ronquillo, Pedro Belda-Ferre, Destiny Alvarado, Tara Javidi, Christopher A Longhurst, and Rob Knight. High-throughput wastewater sars-cov-2 detection enables forecasting of community infection dynamics in san diego county. *Msystems*, 6(2):e00045–21, 2021.

[54] Mary Ann Knovich, Dora Il'yasova, Anastasia Ivanova, and István Molnár. The association between serum copper and anaemia in the adult second national health and nutrition examination

survey (nhanes ii) population. *British journal of nutrition*, 99(6):1226–1229, 2008.

[55] Jeanne Kowalski and Xin M Tu. *Modern applied U-statistics*, volume 714. John Wiley & Sons, 2008.

[56] Nad̆a Krivoňáková, Andrea Šoltýsová, Michal Tamáš, Zdenko Takáč, Ján Krahulec, Andrej Ficek, Miroslav Gál, Marián Gall, Miroslav Fehér, Anna Krivjanská, Ivana Horáková, Noemi Belišová, Paula Bímová, Andrea Butor Škulcová, and Tomáš Mackul̆ak. Mathematical modeling based on rt-qpcr analysis of sars-cov-2 in wastewater as a tool for epidemiology. *Scientific Reports*, 11(1):19456, Sep 2021.

[57] William H Kruskal and W Allen Wallis. Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association*, 47(260):583–621, 1952.

[58] David A Larsen and Krista R Wigginton. Tracking covid-19 with wastewater. *Nature Biotechnology*, 38(10):1151–1153, 2020.

[59] Eva Leidman, Lindsey M Duca, John D Omura, Krista Proia, James W Stephens, and Erin K Sauber-Schatz. Covid-19 trends among persons aged 0–24 years—united states, march 1–december 12, 2020. *Morbidity and Mortality Weekly Report*, 70(3):88, 2021.

[60] Xuan Li, Jagadeeshkumar Kulandaivelu, Shuxin Zhang, Jiahua Shi, Muttucumaru Sivakumar, Jochen Mueller, Stephen Luby, Warish Ahmed, Lachlan Coin, and Guangming Jiang. Data-driven estimation of covid-19 community prevalence through wastewater-based epidemiology. *Science of The Total Environment*, 789:147947, 2021.

[61] Tuo Lin, Tian Chen, Jinyuan Liu, and Xin M Tu. Extending the mann-whitney-wilcoxon rank sum test to survey data for comparing mean ranks. *Statistics in Medicine*, 40(7):1705–1717, 2021.

[62] J. Liu, Xinlian Zhang, T. Chen, T. Wu, T. Lin, L. Jiang, S. Lang, L. Liu, L. Natarajan, J.X. Tu, T. Kosciolek, J. Morton, T.T. Nguyen, B. Schnabl, R. Knight, C. Feng, Y. Zhong, and X.M. Tu. A semiparametric model for between-subject attributes: Applications to beta-diversity of microbiome data. *Biometrics*, 78(3):950–962, 2022.

[63] Jinyuan Liu, Tuo Lin, Tian Chen, Xinlian Zhang, and Xin M Tu. On semiparametric efficiency of an emerging class of regression models for between-subject attributes. *arXiv preprint arXiv:2205.08036*, 2022.

[64] Thomas Lumley and Alastair J Scott. Two-sample rank tests under complex sampling. *Biometrika*, 100(4):831–842, 2013.

[65] Lerato E Magosi, Yinfeng Zhang, Tanya Golubchik, Victor DeGruttola, Eric Tchetgen Tchetgen, Vladimir Novitsky, Janet Moore, Pam Bachanas, Tebogo Segolodi, Refeletswe Lebelonyane, Molly Pretorius Holme, Sikhulile Moyo, Joseph Makhema, Shahin Lockman, Christophe Fraser, Myron Max Essex, and Marc Lipsitch. Deep-sequence phylogenetics to quantify patterns of hiv transmission in the context of a universal testing and treatment trial – bcpp/ya tsie trial. *eLife*,

11:e72657, mar 2022.

[66] Henry B Mann and Donald R Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, pages 50–60, 1947.

[67] Ryan Seamus McGee, Julian R Homburger, Hannah E Williams, Carl T Bergstrom, and Alicia Y Zhou. Model-driven mitigation measures for reopening schools during the covid-19 pandemic. *Proceedings of the National Academy of Sciences*, 118(39):e2108909118, 2021.

[68] Gertjan Medema, Leo Heijnen, Goffe Elsinga, Ronald Italiaander, and Anke Brouwer. Presence of sars-coronavirus-2 rna in sewage and correlation with reported covid-19 prevalence in the early stage of the epidemic in the netherlands. *Environmental Science & Technology Letters*, 7(7):511–516, 2020.

[69] VP Nosko. Local structure of gaussian random fields in vicinity of high-level shines. *Doklady Akademii Nauk SSSR*, 189(4):714, 1969.

[70] Vlad Novitsky, Melissa Zahralban-Steele, Sikhulile Moyo, Tapiwa Nkhisang, Dorcas Maruapula, Mary Fran McLane, Jean Leidner, Kara Bennett, PANGEA Consortium, Kathleen E Wirth, Tendani Gaolathe, Etienne Kadima, Unoda Chakalisa, Molly Pretorius Holme, Shahin Lockman, Mompati Mmalane, Joseph Makhema, Simani Gaseitsiwe, Victor DeGruttola, and M Essex. Mapping of HIV-1C Transmission Networks Reveals Extensive Spread of Viral Lineages Across Villages in Botswana Treatment-as-Prevention Trial. *The Journal of Infectious Diseases*, 222(10):1670–1680, 06 2020.

[71] Ralph G O'Brien and John Castelloe. Exploiting the link between the wilcoxon-mann-whitney test and a simple odds statistic. In *Proceedings of the Thirty-first Annual SAS Users Group International Conference*, pages 209–31. Citeseer, 2006.

[72] A David Paltiel, Amy Zheng, and Rochelle P Walensky. Assessment of sars-cov-2 screening strategies to permit the safe reopening of college campuses in the united states. *JAMA network open*, 3(7):e2016818–e2016818, 2020.

[73] Nikita Patel and Saurabh Upadhyay. Study of various decision tree pruning methods with their empirical comparison in weka. *International journal of computer applications*, 60(12), 2012.

[74] Jordan Peccia, Alessandro Zulli, Doug E. Brackney, Nathan D. Grubaugh, Edward H. Kaplan, Arnau Casanovas-Massana, Albert I. Ko, Amyn A. Malik, Dennis Wang, Mike Wang, Joshua L. Warren, Daniel M. Weinberger, Wyatt Arnold, and Saad B. Omer. Measurement of sars-cov-2 rna in wastewater tracks community infection dynamics. *Nature Biotechnology*, 38(10):1164–1167, Oct 2020.

[75] Laura Elena Raileanu and Kilian Stoffel. Theoretical comparison between the gini index and information gain criteria. *Annals of Mathematics and Artificial Intelligence*, 41:77–93, 2004.

[76] J Jeffery Reeves, Hannah M Hollandsworth, Francesca J Torriani, Randy Taplitz, Shira Abeles,

Ming Tai-Seale, Marlene Millen, Brian J Clay, and Christopher A Longhurst. Rapid response to covid-19: health informatics support for outbreak management in an academic health system. *Journal of the American Medical Informatics Association*, 27(6):853–859, 2020.

[77] J Jeffery Reeves, Christopher A Longhurst, Stacie J San Miguel, Reina Juarez, Joseph Behymer, Kevin M Ramotar, Patricia Maysent, Angela L Scioscia, and Marlene Millen. Bringing student health and well-being onto a health system ehr: the benefits of integration in the covid-19 era. *Journal of American College Health*, 70(7):1968–1974, 2022.

[78] Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974.

[79] Armin Schwartzman, Yulia Gavrilov, and Robert J Adler. Multiple testing of local maxima for detection of peaks in 1D. *Annals of statistics*, 39(6):3290, 2011.

[80] Armin Schwartzman and Fabian Telschow. Peak p-values and false discovery rate inference in neuroimaging. *NeuroImage*, 197:402–413, 2019.

[81] Erwan Scornet, Gérard Biau, and Jean-Philippe Vert. Consistency of random forests. 2015.

[82] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[83] Benjamin Shickel, Patrick James Tighe, Azra Bihorac, and Parisa Rashidi. Deep ehr: a survey of recent advances in deep learning techniques for electronic health record (ehr) analysis. *IEEE journal of biomedical and health informatics*, 22(5):1589–1604, 2017.

[84] Yan-Yan Song and LU Ying. Decision tree methods: applications for classification and prediction. *Shanghai archives of psychiatry*, 27(2):130, 2015.

[85] Name Statistical Analysis System Institute. Sas/stat user's guide (release 9.2), 2008.

[86] Wan Tang, Hua He, and Xin M Tu. *Applied categorical and count data analysis*. CRC Press, 2012.

[87] Jonathan E Taylor, KJ Worsley, and F Gosselin. Maxima of discretely sampled random fields, with an application to 'bubbles'. *Biometrika*, 94(1):1–18, 2007.

[88] Alaa Tharwat. Classification assessment methods. *Applied Computing and Informatics*, 17(1):168–192, 2020.

[89] T Therneau, B Atkinson, and B Ripley. Recursive partitioning and regression trees. r package version 4.1-15, 2019.

[90] Steven K Thompson. *Sampling*, volume 755. John Wiley & Sons, 2012.

[91] Times T.N.Y. Tracking coronavirus cases at us colleges and universities. *New York Times*, 25, 2020.

[92] Fabian Tomaschek, Peter Hendrix, and R Harald Baayen. Strategies for addressing collinearity in multivariate linguistic data. *Journal of Phonetics*, 71:249–267, 2018.

[93] Anastasios A Tsiatis. *Semiparametric theory and missing data*. Springer, 2006.

[94] Juan A. Vallejo, Soraya Rumbo-Feal, Kelly Conde-Pérez, Ángel López-Oriona, Javier Tarrío, Rubén Reif, Susana Ladra, Bruno K. Rodiño-Janeiro, Mohammed Nasser, Ángeles Cid, María C Veiga, Antón Acevedo, Carlos Lamora, Germán Bou, Ricardo Cao, and Margarita Poza. Highly predictive regression model of active cases of covid-19 in a population by screening wastewater viral load. *medRxiv*, 2020.

[95] Tyler Vu, Tuo Lin, Jingjing Zou, Vladimir Novitsky, Xin Tu, and Victor De Gruttola. Estimating viral genetic linkage rates in the presence of missing data. *arXiv preprint arXiv:2203.12779*, 2022.

[96] Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.

[97] Stefan Wager and Guenther Walther. Adaptive concentration of regression trees, with application to random forests. *arXiv preprint arXiv:1503.06388*, 2015.

[98] Rochelle P Walensky, Henry T Walke, and Anthony S Fauci. Sars-cov-2 variants of concern in the united states—challenges and opportunities. *Jama*, 325(11):1037–1038, 2021.

[99] Frank Wilcoxon. Probability tables for individual comparisons by ranking methods. *Biometrics*, 3(3):119–122, 1947.

[100] Christopher Winship and Larry Radbill. Sampling weights and regression analysis. *Sociological Methods & Research*, 23(2):230–257, 1994.

[101] Keith J Worsley. An improved theoretical P value for SPMs based on discrete local maxima. *Neuroimage*, 28(4):1056–1062, 2005.

[102] Keith J Worsley, Alan C Evans, Sean Marrett, and P Neelin. A three-dimensional statistical analysis for CBF activation studies in human brain. *Journal of Cerebral Blood Flow & Metabolism*, 12(6):900–918, 1992.

[103] Keith J Worsley, Sean Marrett, Peter Neelin, Alain C Vandal, Karl J Friston, and Alan C Evans. A unified statistical approach for determining significant signals in images of cerebral activation. *Human brain mapping*, 4(1):58–73, 1996.

[104] Fuqing Wu, Amy Xiao, Jianbo Zhang, Katya Moniz, Noriko Endo, Federica Armas, Richard Bonneau, Megan A. Brown, Mary Bushman, Peter R. Chai, Claire Duvallet, Timothy B. Erickson, Katelyn Foppe, Newsha Ghaeli, Xiaoqiong Gu, William P. Hanage, Katherine H. Huang, Wei Lin Lee, Mariana Matus, Kyle A. McElroy, Jonathan Nagler, Steven F. Rhode, Mauricio Santillana, Joshua A. Tucker, Stefan Wuertz, Shijie Zhao, Janelle Thompson, and Eric J. Alm. Sars-cov-2 rna concentrations in wastewater foreshadow dynamics and clinical presentation of new covid-19 cases.

Science of The Total Environment, 805:150121, 2022.

[105] Jionglin Wu, Jason Roy, and Walter F Stewart. Prediction modeling using ehr data: challenges, strategies, and a comparison of machine learning approaches. *Medical care*, pages S106–S113, 2010.

[106] Q Yu, W Tang, J Kowalski, and XM Tu. Multivariate u-statistics: a tutorial with applications. *Wiley Interdisciplinary Reviews: Computational Statistics*, 3(5):457–471, 2011.