TREATMENT EFFECT HETEROGENEITY AND STATISTICAL DECISION-MAKING IN THE PRESENCE OF INTERFERENCE

TREATMENT EFFECT HETEROGENEITY AND STATISTICAL DECISION-MAKING IN THE PRESENCE OF INTERFERENCE

By JULIUS OWUSU

A Dissertation Submitted to the School of Graduate Studies in Partial Fulfilment of the Requirements for the Degree Doctor of Philosophy

McMaster University DOCTOR OF PHILOSOPHY (2023) Hamilton, Ontario (Economics)

TITLE: Treatment Effect Heterogeneity and Statistical Decision-making in the Presence of Interference

AUTHOR: Julius Owusu

B.A. (University of Ghana)

M.A. (Brock University)

SUPERVISOR: Young Ki Shin

NUMBER OF PAGES: 130

Abstract

This dissertation consists of three chapters that generally focus on the design of welfare-maximizing treatment assignment rules in heterogeneous populations with interactions. In the first two chapters, I focus on an important pre-step in the design of treatment assignment rules: inference for heterogeneous treatment effects in populations with interactions. In the final chapter, I and my co-authors study treatment assignment rules in the presence of social interaction in heterogeneous populations.

In chapter one, I argue that statistical inference of heterogeneous treatment effects (HTEs) across predefined subgroups is complicated when economic units interact because treatment effects may vary by pretreatment variables, post-treatment exposure variables (that measure the exposure to other units' treatment statuses), or both. It invalidates the standard hypothesis testing technique used to infer HTEs. To address the problem, I develop statistical methods (asymptotic and bootstrap) to infer HTEs and disentangle the drivers of treatment effects heterogeneity in populations where units interact. Specifically, I incorporate clustered interference into the potential outcomes model and propose kernel-based test statistics for the null hypotheses of (a) no HTEs by treatment assignment (or post-treatment exposure variables) for all pretreatment variables values; and (b) no HTEs by pretreatment variables for all treatment assignment vectors. To disentangle the source of heterogeneity in treatment effects, I recommend a multiple-testing algorithm. In addition, I prove the asymptotic properties of the proposed test statistics via a modern poissonization technique.

As a robust alternative to the inferential methods I propose in chapter one, in chapter two, I design randomization tests of heterogeneous treatment effects (HTEs) when units interact on a single network. My modeling strategy allows network interference into the potential outcomes framework using the concept of network exposure mapping. I consider three null hypotheses that represent different notions of homogeneous treatment effects, but due to nuisance parameters and the multiplicity of potential outcomes, the hypotheses are not sharp. To address the issue of multiple potential outcomes, I propose a conditional randomization inference method that expands on existing methods. Additionally, I consider two techniques that overcome the nuisance parameter issue. I show that my conditional randomization inference method, combined with either of the proposed techniques for handling nuisance parameters, produces asymptotically valid p-values.

Chapter three is based on a joint paper with Young Ki Shin and Seungjin Han. We study treatment assignment rules in the presence of social interaction in heterogeneous populations. We construct an analytical framework under the anonymous interaction assumption, where the decision problem becomes choosing a treatment fraction. We propose a multinomial empirical success (MES) rule

that includes the empirical success rule of Manski (2004) as a special case. We investigate the non-asymptotic bounds of the expected utility based on the MES rule.

Acknowledgements

I am grateful to my dissertation supervisor, Young Ki Shin, for his guidance, inspiration, and generosity. During my doctoral studies, Young Ki Shin has provided me with meaningful advice that has intellectually pushed and nourished me. I feel fortunate to be one of his mentees because of his extraordinary commitment. Also, my profound gratitude goes to my other committee members, Jeffrey Racine and Michael Veall, for their guidance and inspiration throughout my doctoral studies. I am deeply indebted to all other faculty members in the Department of Economics at McMaster University, especially Arthur Sweetman, Alok Johri, and Pau S. Pujolas, for their invaluable advice throughout my years at McMaster. In addition, I want to thank all the Professors in the Department of Economics at Brock University, especially Robert Dimand, Tomson Ogwang, and Jean-Francois Lamarche, for their generosity during my master's studies.

This dissertation has also benefited from discussions with Rabuil Islam, Alex Sam, Zvev Todorov, Moyosore Sogaolu, Bright Komla Agbewu, Chris Muris, Irene Botosaru, Sukjin Han, David Pacini, Pietro Spini, Stefan Hubner, and Sami Stouli. I acknowledge the financial support from the Ontario Graduate Scholarship and the Productivity Partnership.

I am also thankful to my fellow graduate students and friends, particularly, Camille Simardone, Sergei Filiasov, Alamgir Farzana, Karen Ugarte Bravo, and Kwabena Nkansah-Amankra, for the stimulating intellectual discussions and motivation they provided. I regret I cannot mention all of you here. Finally, I am thankful to my parents, Phillip (late) and Susan, my brother Jerry and sisters Monica and Pina, and the rest of my family, whose unconditional love and undying support allowed me to complete this dissertation. I dedicate this dissertation in memory of my late father, Phillip Tephay Owusu, who inspired me and supported my education.

Contents

List of Tables

List of Figures

Declaration of Academic Achievement

I am the sole author of chapters one and two of this dissertation. Chapter three is based on joint work with Young Ki Shin and Seungjin Han.

Chapter 1

A Nonparametric Test of Heterogeneous Treatment Effects under Interference

1.1 Introduction

The literature on causal inference mainly focuses on the identification and estimation of aggregate treatment effects. Such aggregate effect metrics provide a measure of social welfare but fail to reveal the variations in treatment effects that are crucial for designing welfare-maximizing treatment assignment rules. For instance, a job search assistance program may have positive aggregate effects on income and welfare, yet create large income and welfare disparities in society because the effects may vary across persons. This observation has motivated a growing inquiry into the estimation and inference of heterogeneous treatment effects (HTEs). To infer HTEs, the classical approach involves the estimation and comparison of conditional average treatment effects (CATEs) of predefined subgroups in a population. The formal comparison of CATEs requires some hypothesis testing procedure. Crump et al. (2006), Ding et al. (2016), Wager and Athey (2018), and Sant'Anna (2021) are among the few existing HTEs testing papers.

This chapter proposes a pair of nonparametric tests to infer HTEs across subgroups while accounting for interference among economic units. According to Cox (1958), *interference* exists when the treatment of one unit affects the response of another unit. It may be due to physical, virtual, or social ties among the members of a population. Several mechanisms could explain how interference occurs, but regardless of the medium, it complicates the inference of HTEs across subgroups. Due to interference, treatment effects may vary by pretreatment variables, post-treatment exposure variables (that measure the exposure to other units' treatment statuses), or both. For instance, suppose we find that the effectiveness of a Covid-19 vaccine varies by city in Canada. Note that the effectiveness of a Covid-19 vaccine for an individual depends on the vaccination rate among her physical contacts. Therefore, it is crucial to verify if the observed variation is explained by the natural differences in the populations across cities (e.g., genetic variation) or if it is due to variations in the vaccination rate across cities. In this example, the Canadian city is a pretreatment variable, and the vaccination rate is a post-treatment exposure variable. This example demonstrates that the existing HTEs tests (developed under the assumption of *no interference* in the papers cited above) will most likely lead to erroneous decisions when some form of interference is present.

I propose kernel-based test statistics for the null hypothesis of (a) constant treatment effects (CTEs) by treatment assignment for all pretreatment variables values; and (b) CTEs by the pretreatment variables for all treatment assignment vectors. Then, I recommend the Holm (1979) multiple testing algorithm to jointly test the null hypotheses and disentangle the source of heterogeneity in the treatment effects. The proposed test statistics are sums of the weighted L_1-norm of the differences in consistent kernel estimators of CATEs that characterize the null hypotheses. Although, it is possible to construct the test statistics with the general L_p-norm for $p \geq 1$, I use the L_1-norm because it is the

natural distance measure and also because it eases the complexity of the asymptotic theory (Lee and Whang (2009)).

The motivation for this chapter is three-fold and closely tied to the null hypotheses mentioned above. First, the test of the null hypothesis of CTEs by treatment assignment for all pretreatment variables values in isolation informs policymakers whether to scale a program or not. If we fail to reject the null hypothesis of CTEs across treatment assignments for all values of the pretreatment variables, it implies that treatment spillover effects are absent and a program can be scaled without any negative or positive externalities. Second, the null hypothesis of CTEs by the pretreatment variables for all treatment assignment vectors in isolation helps to detect HTEs across subgroups defined by pretreatment variables. Knowledge of a program's treatment effect variation across subgroups can guide its extension to other populations (external validation). Third, jointly testing both null hypotheses helps to disentangle the source of variation in treatment effects. It is the leading motivation of this chapter because finding the drivers of the variations in treatment effects in an interconnected human society is a crucial step in designing welfare-maximizing treatment assignment rules.

The main theoretical results show that the proposed test statistics have an asymptotically standard normal null distribution. I prove these results using a modern poissonization technique due Giné et al. (2003). To quote Lee and Whang (2009, p. 309), "the poissonization technique introduces additional randomness by assuming that the sample size is a Poisson random variable." It enables the use of techniques that exploit the independent increments and infinite divisibility properties of Poisson processes. Also, I show that the test statistics have asymptotically valid sizes and are consistent against fixed and sequences of local alternatives. Moreover, to better approximate the null distributions of the proposed test statistics in small-sample settings, I propose and recommend bootstrap methods of inference.

In summary, I contribute to the literature in different ways. This chapter is the first to propose a pair of nonparametric tests for HTEs that (i) can accommodate several forms of clustered interference and (ii) can disentangle the sources of heterogeneity in treatment effects under clustered interference. To the best of my knowledge, this study is also the first to provide a bootstrap approach for testing HTEs.

The organization of the rest of the chapter is as follows. I review existing related work in the second part of this section. Section 1.2 describes the setup, the testing problem, and the test statistics. In Section 1.3, I present the main asymptotic properties of the test statistics. I introduce the bootstrap methods in Section 1.4. Monte Carlo simulation design and results are in Section 1.5. My concluding remarks are in Section 1.6. All proofs, useful theorems, lemmas, and detailed results

from simulation experiments are in the Appendix.

1.1.1 Related Literature

The nascent literature on the estimation and inference of HTEs continues to grow and spans multiple fields. The chapter falls under the arm of the literature that studies the inference of HTEs using tests based on average treatment effects (ATEs) of subgroups defined by pretreatment variables. Bitler, Gelbach, and Hoynes (2006) provide an in-depth critique of this approach to testing for HTEs. They argue that heterogeneity of CATEs across subgroups often does not imply individual treatment effect variation unless one assumes constant subgroup treatment effect (CSTE). Nonetheless, this is the approach of this current chapter because regardless of the CSTE assumption, variations in ATEs of subgroups are crucial in the design of treatment assignment rules where pretreatment variables are used to set eligibility conditions. In econometrics, tests to detect variation in ATEs across subgroups have been studied by Crump et al. (2006) and Lee and Shaikh (2014). Both studies abstract from interference and propose nonparametric tests to infer HTEs across predefined subgroups. In contrast, I allow clustered interference and use kernel-based estimators to construct the test statistics. To my knowledge, this is the first research to use kernel-based test statistics to test for HTEs.

On the theoretical side, Lee, Song, and Whang (2013) and Chang, Lee, and Whang (2015) establish asymptotic null distributions for the L_p-type functions of kernel-based CATE estimators using the poissonization technique of Giné, Mason, and Zaitsev (2003). Allowing for clustered interference in the potential outcomes model requires a modification of the estimator of CATE to fit the current framework. In addition, the proposed test statistics in this chapter are different, and the theoretical results are mostly nontrivial extensions of those in the articles mentioned above.

Finally, the nonparametric bootstrap algorithms I propose are similar to existing algorithms in the literature. For instance, Li, Maasoumi, and Racine (2009) use a bootstrap algorithm akin to that described in Subsection 1.4.1 to test for the equality of two density or conditional density functions. Also, Racine (1997) employs a residual bootstrap algorithm similar to that described in Subsection 1.4.2 to test for the significance of pretreatment variables in regression models. Despite the structural similarities, the bootstrap algorithms in this chapter are new constructs that have been adapted to fit the current framework.

1.2 Framework

1.2.1 Setup

Let N denote the sample size of a random sample drawn from a large population. The units $i \in \{1, \ldots, N\}$ interact on an undirected network. In addition, assume the network is clustered with C large (but finite), identical[1] and non-overlapping clusters having sample sizes N_c, $c = 1, \ldots, C$. In other words, only units in the same cluster interact. However, the within group network is unobservable. For example, clusters can be spatial like schools, villages, segregated labor markets, etc, or virtual like Facebook groups, WhatsApp groups, etc. Therefore, the type of interference I consider is the same as imposing the *partial interference* assumption in Sobel (2006). Furthermore, let $T \in \{0, 1\}$ be the binary treatment indicator assigned at the unit level. This implies that different clusters may have different treatment assignment vectors[2]. The treatment vector of the sample in cluster c is denoted as $\mathbf{T}_c$. Finally, we observe $Y \in \mathbb{R}$, the outcome variable, and a vector $X \in \mathcal{X} \subset \mathbb{R}^d$ of pretreatment variables. For notational simplicity, let $\mathcal{S}$ denote the support of (Y, X).

Applying the potential outcomes model of Neyman (1923) and Rubin (1974), let $Y_i(\mathbf{t}_c)$ represent the potential outcome of unit i in cluster c when the cluster treatment assignment vector $\mathbf{T}_c = \mathbf{t}_c$. This indicates that the potential outcomes of a unit are indexed by her cluster treatment assignment vector. Effectively, the number of potential outcomes of a unit in cluster c equals 2^{N_c} (i.e., the number of all possible cluster treatment assignment vectors), which goes to infinity at an exponential rate as the cluster sample size increases. Note that in the classical case of no interference, there are only two potential outcomes for each unit. Hence, allowing for (clustered) interference aggravates the missing data problem of causal inference, and as such a salient element of causal inference in the presence of (clustered) interference is a mapping $\pi(\cdot)$ which summarizes how the (cluster) treatment vectors affect the treatment response outcome. This is called *exposure mapping* in Aronow et al. (2017). Formally, I define exposure mapping as

$$\pi : \{0, 1\}^{N_c} \mapsto \mathbf{\Pi} \tag{1.1}$$

that maps the cluster treatment vector $\mathbf{T}_c$ into an exposure variable $\Pi := \pi(\mathbf{T}_c) \in \mathbf{\Pi} \subset \mathbb{R}$. Now, given an exposure mapping, I assume that $Y_i(\mathbf{t}_c) = Y_i(t_i, \pi)$ is the potential outcome for unit i if the cluster treatment vector $\mathbf{t}_c$ is such that $T_i = t_i$ and $\pi(\mathbf{t}_c) = \pi$. Notice that this assumption

[1] Identical here implies that there are no cluster-specific shocks which affects the outcome of interest. The sizes of the clusters can be different.

[2] A common experimental design to obtain different treatment vectors in different clusters is the two-stage random saturation design in Baird et al. (2018).

asserts that *heterogeneity in treatment effects across treatment assignment vectors is analogous to heterogeneity in treatment effects across the exposure variable's values (henceforth exposure values)*. It is also worth mentioning that the definition of exposure mapping in (1.1) requires no knowledge of the links between economic units in the population (i.e., the adjacency matrix). This is convenient because, in many applied settings in social science, it is difficult to obtain information on links between economic units due to privacy concerns. For instance, Colpitts (2002) reveals that a targeting program designed to use individual-level and network information to assign unemployed workers to different labor market job activation programs in Canada was abandoned due to data security concerns. The following example provides a plausible concrete specification of exposure mapping defined in (1.1).

Example 1.2.1 (Distributional clustered interaction) *Manski (2013) explains that distributional clustered interaction occurs if the outcome of unit i does not depend on the sample size and it is invariant to permutations of the treatments received by other units in the same cluster. That is, distributional clustered interaction implies that* $\pi(\mathbf{t}_c) \in \{0, 1/N_c, 2/N_c, \ldots, 1\}$ *represents the treatment ratio in cluster c. Hence, for any two cluster treatment vectors* $\mathbf{t}_c \neq \mathbf{t}'_c$ *such that* $\mathbf{t}_c = (t_i, \mathbf{t}_c)$ *and* $\mathbf{t}'_c = (t_i, \mathbf{t}'_c)$, $Y_i(\mathbf{t}_c) = Y_i(\mathbf{t}'_c)$ *if* $\pi(\mathbf{t}_c) = \pi(\mathbf{t}'_c)$. *See Manski (2013) for other specifications under different scenarios.*

Now, I formalize the assumptions that describe the network structure as follows.

Assumption 1.2.1 (Treatment-invariant neighborhoods) *A unit's cluster and treatment status are independent.*

In other words, Assumption 1.2.1 means that the network is a fixed characteristic of the population, and units do not self-select into clusters after treatment assignment.

Assumption 1.2.2 (Clustered Interference) *Let* $\pi(\cdot)$ *be an exposure mapping function, i.e.,* $\pi : \{0, 1\}^{N_c} \mapsto \mathbf{\Pi}$, *with* $\mathbf{\Pi} = \{\pi_1, \ldots, \pi_K : K < C\}$ *being a discrete set of finite elements,* $\forall c = 1, \ldots, C$, $\forall i = 1, \ldots, N_c$, *and* $\forall \mathbf{t}, \mathbf{t}' \in \{0, 1\}^N$ *such that* $t_i = t'_i$ *and* $\pi(\mathbf{t}_c) = \pi(\mathbf{t}'_c)$ *then* $Y_i(\mathbf{t}) = Y_i(\mathbf{t}')$

Assumption 1.2.2 implicitly restricts the network structure, i.e., it allows intra-cluster interference but no inter-cluster interference. Also, this assumption restricts the range of the exposure variable to a discrete finite set. It implies that the number of potential outcomes reduces significantly to $2 \cdot K$. This assumption ensures that there are "enough" units for each treatment, exposure value, and covariates value to allow the precise estimation of the CATEs. Finally, assumption 1.2.2 implies that a unit's exposure to the treatment of other units in her cluster are in a "reduced form" and channels through

which interference occurs in clusters are not distinguishable (Aronow et al. (2017)). Exploring extensions that allow for other types of interference is a possible avenue for future research.

Without loss of generality, I assume that the exposure mapping $\pi(\cdot)$ is the cluster treatment ratio, i.e., $\pi(\mathbf{t}_c) = (\sum_{i=1}^{N_c} t_i)/N_c$. The results in this chapter hold for other definitions of $\pi(\cdot)$.

Next, realized outcomes can be written in terms of potential outcomes as:

$$Y_i := \sum_{k=1}^{K} \Big(Y_i(0, \pi_k) + [Y_i(1, \pi_k) - Y_i(0, \pi_k)] \cdot T_i \Big) \cdot \mathbb{1}(\Pi_i = \pi_k), \ \ \forall i. \tag{1.2}$$

where $\mathbb{1}(\cdot)$ is the indicator function.

I assume that independent and identically distributed (i.i.d) copies $\{(Y_i, T_i, X_i, \Pi_i) : i = 1, \ldots, N\}$ of (Y, T, X, Π) are available [3].

To proceed, for $\pi \in \mathbf{\Pi}$ and $x \in \mathcal{X}$, I define the CATE as $\tau(x; \pi) := \mathbb{E}[Y(1, \pi)|X = x] - \mathbb{E}[Y(0, \pi)|X = x]$. The following assumptions are necessary for the identification of the CATEs.

Assumption 1.2.3 (Unconfoundedness) *For all $\pi \in \mathbf{\Pi}$,*

$$(T, \Pi) \perp\!\!\!\perp (Y(0, \pi), Y(1, \pi))|X \tag{1.3}$$

Assumption 1.2.4 (Overlap) *For all $\pi \in \mathbf{\Pi}$, $x \in \mathcal{X}$ and for some $\eta > 0$*

$$\eta < \mathrm{P}(T = 1, \Pi = \pi|X = x) < 1 - \eta \tag{1.4}$$

Assumptions 1.2.3 and 1.2.4 are the extensions of the ignorability assumptions imposed on the treatment assignment mechanism under no interference, see Imbens and Rubin (2015). Assumption 1.2.3 is a modification of the usual unconfoundedness assumption. This assumption asserts that conditional on pretreatment variables, self-selection of *effective treatment*, (T, Π) is ruled out. Note that this assumption holds when data is from a randomized experiment. Assumption 1.2.4 is a modified version of the usual overlap or probabilistic assignment condition. It ensures a balance between treated and control units in each subgroup. It is crucial because of the Hájek-type estimators of CATE I employ in this chapter. Together, these assumptions are critical to identifying the CATEs, i.e., if Assumption 1.2.3 and 1.2.4 holds, then

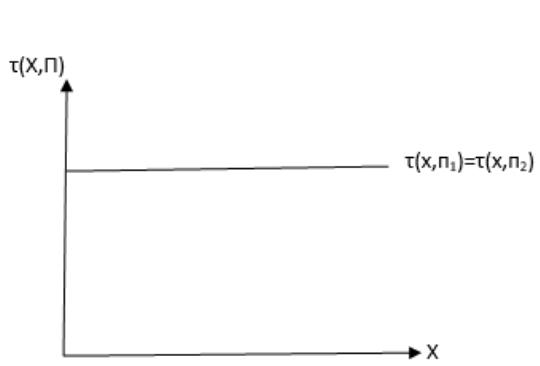
$$\begin{aligned} \tau(x; \pi) :=& \mathbb{E}[Y(1, \pi)|X = x] - \mathbb{E}[Y(0, \pi)|X = x] \\ =& \mathbb{E}[Y|T = 1, \Pi = \pi, X = x] - \mathbb{E}[Y|T = 0, \Pi = \pi, X = x] \end{aligned} \tag{1.5}$$

[3] A necessary requirement of obtaining i.i.d observations is that the population must consist of a large number of clusters. In particular, using the random saturation design, the sample size N must be less or equal to the number of clusters C for us to obtain an i.i.d data. It implies a sampling design where we draw at most one unit from each cluster after implementation of the experiment. This experimental-cum-sampling design merits further investigation as it produces i.i.d data and simplifies causal analysis under partial interference.

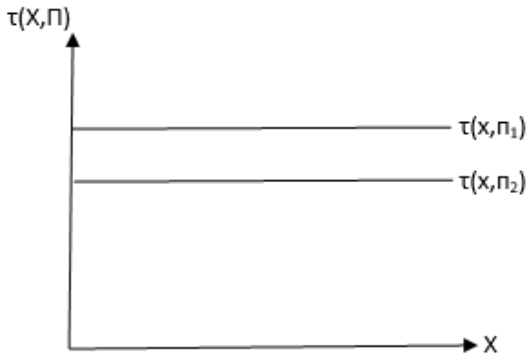
1.2.2 The Testing Problem

In this subsection, I provide a formal description of the testing problem. As mentioned above, in the presence of interference, treatment effects may vary by either pretreatment variables or post-treatment exposure variables. Figure 1.1 shows some possible cases in a simple setting where $\mathbf{\Pi} = \{\pi_1, \pi_2\}$ and CATE is linear in a continuous X. Panels (a) shows CTEs by both classes of variables. In contrast, Panels (b) and (c) show scenarios where there are HTEs by one class of variable and CTEs by the other class. Panel (d) depicts the case where there is HTEs by both classes of variables. These facts highlight that testing for HTEs in the presence of interference requires testing for heterogeneity across the two classes of variables and the need for methods to disentangle the source of the effect heterogeneity.

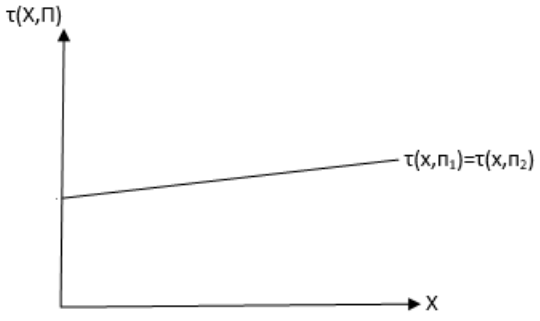
Figure 1.1: Treatment Effects Variation by a Continuous pretreatment Variable and a Binary Post-treatment Exposure Variable



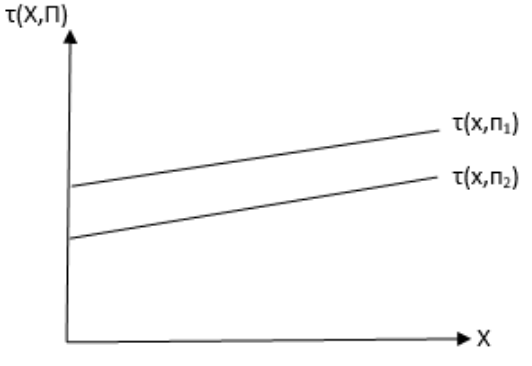
(a) CTEs by pretreatment variables and exposure variable.



(b) CTEs by pretreatment variables and HTEs by exposure variable.



(c) HTEs by pretreatment variables and CTEs by exposure variable.



(d) HTEs by pretreatment variables and exposure variable.

To test for heterogeneity across one class of variables requires controlling for the other class of variables. Formally, I consider the following null hypotheses. The first one is the null hypothesis of constant treatment effects (CTEs) by treatment ratio (exposure variable) while controlling for pretreatment variables:

$$H_0^{\Pi}: \ \forall x \in \mathcal{X}, \ \forall \pi, \pi' \in \mathbf{\Pi}, \ \tau(x;\pi) = \tau(x;\pi'), \tag{1.6}$$

against the alternative hypothesis of HTEs by treatment ratio:

$$H_1^{\Pi}: \ \exists x \in \mathcal{X}, \ \exists \pi, \pi' \in \mathbf{\Pi}, \ \tau(x;\pi) \neq \tau(x;\pi'). \tag{1.7}$$

Null hypothesis (1.6) is important in answering a number of questions in program evaluation. For example, testing (1.6) helps to determine whether a program can be scaled or not; rejecting the null hypothesis implies treatment spillover effects exist and program effects may vary by the scale of a program. Although (1.6) resembles Hypothesis 2 in Athey et al. (2018), they are not the same. The null hypothesis (1.6) is a restriction on the treatment effect, in other words, this null says that there are no indirect or spillover treatment effects, whereas, Hypothesis 2 in Athey et al. (2018) is a restriction on the outcome that there is no spillovers or interference. Failure to reject Hypothesis 2 in Athey et al. (2018) implies that we fail to reject (1.6), however, the converse is not true.

The second hypothesis of interest concerns the null hypothesis of CTEs by pretreatment variables while controlling for the post-treatment exposure variable:

$$H_0^{X}: \ \forall \pi \in \mathbf{\Pi}, \forall x, x' \in \mathcal{X}, \ \tau(x;\pi) = \tau(x';\pi), \tag{1.8}$$

against the alternative hypothesis of HTEs by pretreatment variables:

$$H_1^{X}: \ \exists \pi \in \mathbf{\Pi}, \exists\, x, x' \in \mathcal{X}, \ \tau(x;\pi) \neq \tau(x';\pi). \tag{1.9}$$

Tests of the null hypothesis (1.8) and their importance exist in the literature under the no interference assumption; see Crump et al. (2006). However, to my knowledge, this chapter provides the first of its kind to allow some form of interference. Testing (1.8) is a critical step in extending existing programs to new populations; rejecting this null hypothesis implies that treatment effects vary across subgroups defined by the pretreatment variables. Therefore, a policymaker should not expect the same aggregate effects if the program is extended to a new population that has a different distribution of pretreatment variables.

It is worth discussing the importance of distinguishing between the exposure variable (i.e., Π)

from the pretreatment variables in this framework. First, note that Π is a causal variable and, as such, is used to index the potential outcomes. In contrast, the pretreatment variables are invariant to treatment assignment i.e., they are a priori variables. Therefore, for identification, this distinction is necessary in the potential outcome framework. Secondly, distinguishing between the pretreatment variables and the treatment exposure variable is imperative to determine the drivers of the heterogeneity in treatment effects. It has interesting implications for policymakers. For instance, if the heterogeneity is solely due to the treatment exposure variable, then policymakers may have complete control over the distribution of treatment effects in the population via the treatment assignment. See Han, Owusu, and Shin (2022) for a statistical treatment assignment rule for homogeneous and heterogeneous populations in the presence of social interaction.

To sum, I emphasize that hypotheses (1.6) and (1.8) are critical to infer and disentangle treatment effects heterogeneity in the presence of clustered interference. Specifically, under the assumption of constant treatment effects within subgroups defined by pretreatment variables, rejecting both null hypotheses *simultaneously* implies HTEs.

1.2.3 Test Statistics

I describe the test statistics for the null hypotheses (1.6) and (1.8). The test statistic for the null hypothesis (1.6) is

$$\hat{T}_1 := \int_{\mathcal{X}} \sum_{k=1}^{K} \sum_{j=1}^{K} \left\{ \sqrt{N} |\hat{\tau}(x;\pi_k) - \hat{\tau}(x;\pi_j)| \right\} \frac{\hat{w}(x,\pi_k,\pi_j)}{2} dx, \tag{1.10}$$

where $\hat{\tau}(x;\pi)$ is a uniform consistent estimator of $\tau(x;\pi)$ and $\hat{w}(x,\pi,\pi')$ is the uniform consistent estimator of the inverse standard error of $\hat{\tau}(x;\pi) - \hat{\tau}(x;\pi')$ which is defined as $w(x,\pi,\pi') := 1/\sqrt{\rho_2(x,\pi) + \rho_2(x,\pi')}$. Here, $\rho_2(x,\pi)$ represents the standard error of $\hat{\tau}(x;\pi)$. Estimation of $\tau(x;\pi)$ and $w(x,\pi,\pi')$ can be done in several ways, but I propose a kernel estimation technique. The kernel estimator of $\tau(x;\pi)$ is defined as

$$\hat{\tau}(x;\pi) := \frac{1}{Nh^d} \sum_{i=1}^{N} Y_i \cdot \mathbb{1}(\Pi_i = \pi) \hat{\phi}(T_i, x, \pi) K\left(\frac{x - X_i}{h}\right),$$

where

$$\hat{\phi}(T_i, x, \pi) := \frac{T_i}{\hat{P}_1(x;\pi)} - \frac{(1 - T_i)}{\hat{P}_0(x;\pi)},$$

with

$$\hat{P}_t(x;\pi) := \frac{1}{Nh^d}\sum_{i=1}^{N}\mathbb{1}(\Pi_i=\pi)\mathbb{1}(T_i=t)K\left(\frac{x-X_i}{h}\right), \quad t=0,1.$$

On the other hand, the kernel estimator of $w(x,\pi,\pi')$ is defined as

$$\hat{w}(x,\pi,\pi') := \frac{1}{\sqrt{\hat{\rho}_2(x,\pi)+\hat{\rho}_2(x,\pi')}},$$

where $\hat{\rho}_2(x,\pi)$ is the kernel estimator of $\rho_2(x,\pi)$ defined as

$$\hat{\rho}_2(x,\pi) := (\hat{\mu}_1(x,\pi)-\hat{\mu}_2(x,\pi))\cdot\int K(\xi)^2 d\xi,$$

with

$$\hat{\mu}_1(x,\pi) := \frac{1}{Nh^d}\sum_{t\in\{0,1\}}\sum_{i=1}^{N}\frac{Y_i^2\mathbb{1}(T_i=t)\mathbb{1}(\Pi=\pi)K(\frac{x-X_i}{h})}{\hat{P}_t^2(x;\pi)},$$

and

$$\hat{\mu}_2(x,\pi) := \frac{1}{N^2h^{2d}}\sum_{t\in\{0,1\}}\sum_{j=1}^{N}\sum_{i=1}^{N}\frac{Y_jY_i\mathbb{1}(T_j=t)\mathbb{1}(T_i=t)\mathbb{1}(\Pi=\pi)K(\frac{x-X_j}{h})K(\frac{x-X_i}{h})}{\hat{P}_t^3(x;\pi)}.$$

Note that d is the dimension of X, $K(\cdot)$ is a d-dimensional kernel function and h is the bandwidth.

Remark 1.2.1 *Rewriting $\hat{T}_1$ in the form $\hat{T}_1 = 2^{-1}\sum_{k=1}^{K}\sum_{j=1}^{K}\int_{\mathcal{X}}\left\{\sqrt{N}|\hat{\tau}(x;\pi_k)-\hat{\tau}(x;\pi_j)|\right\}\cdot\hat{w}(x,\pi_k,\pi_j)dx$, the resemblance between $\hat{T}_1$ and the test statistic $\hat{D} := \int_{\mathcal{X}}\sqrt{N}|\hat{\tau}(x)|\hat{w}(x)dx$[4] by Chang, Lee, and Whang (2015) is obvious. They show that a studentized version of $\hat{D}$ converges to the standard normal distribution under the null hypothesis of non-positive treatment effects. Compared to $\hat{D}$, the proposed test statistic is the sum of dependent random variables $2^{-1}\int_{\mathcal{X}}\left\{\sqrt{N}|\hat{\tau}(x;\pi_k)-\hat{\tau}(x;\pi_j)|\right\}\hat{w}(x,\pi_k,\pi_j)dx$, for all $\pi_j,\pi_k\in\mathbf{\Pi}$, hence it is not a straightforward extension to prove the asymptotic normality and other asymptotic properties of $\hat{T}_1$ using the asymptotic results of $\hat{D}$.*

Remark 1.2.2 *Failure to reject the null hypothesis H_0^{Π} using $\hat{T}_1$ or a suitable studentized version implies either CTEs by treatment ratio and HTEs by pretreatment variables or CTEs by both treatment ratio and pretreatment variables. Hence in isolation, failure to reject the null hypothesis* (1.6) *using $\hat{T}_1$ does not help us disentangle the source of heterogeneity in the treatment effects.*

[4] $\hat{\tau}(x)$ and $\hat{w}(x)$ are the CATE and inverse standard error estimators defined similar to $\hat{\tau}(x;\pi_k)$ and $\hat{w}(x,\pi_k,\pi_j)$ respectively.

In a similar fashion, the test statistic for the null hypothesis (1.8) is defined as

$$\hat{T}_2 := \int_{\mathcal{X}} \int_{\mathcal{X}} \sum_{k=1}^{K} \left\{ \sqrt{N} |\hat{\tau}(x; \pi_k) - \hat{\tau}(x'; \pi_k)| \right\} \frac{\hat{w}(x, x', \pi_k)}{2} dx dx', \tag{1.11}$$

where $\hat{w}(x, x', \pi)$ for all $\pi \in \mathbf{\Pi}$, and $x, x' \in \mathcal{X}^2$, is a consistent kernel estimator of the inverse standard error of $\hat{\tau}(x; \pi) - \hat{\tau}(x'; \pi)$. Define $\rho_0(q) := \int_{-1}^{1} K(\xi + q)K(\xi)d\xi / \int_{-1}^{1} K(\xi)^2 d\xi$, then the estimator of the inverse standard error of $\hat{\tau}(x; \pi) - \hat{\tau}(x'; \pi)$ is $\hat{w}(x, x', \pi) := 1/\sqrt{2\hat{\rho}_2(x, \pi) \cdot (1 - \rho_0((x' - x)/h^d))}$.

Remark 1.2.3 *Failure to reject the null hypothesis* (1.8) *using* $\hat{T}_2$ *or its studentized version implies either CTEs by the pretreatment variables and HTEs by the treatment ratio or CTEs by both treatment ratio and pretreatment variables. Therefore, in isolation, failure to reject the null hypothesis* (1.8) *using* $\hat{T}_2$ *also does not help in disentangling the source of heterogeneity in the treatment effects.*

Now, if we test the null hypotheses (1.6) and (1.8) simultaneously using a multiple testing procedure (MTP), we can disentangle the source of treatment effects heterogeneity. If we fail to reject both null hypotheses, it implies CTEs by both variable classes. If we reject the null hypothesis (1.6) but fail to reject the second null hypothesis (1.8), it suggests HTEs by treatment ratio and CTEs by pretreatment variables. In contrast, if we fail to reject the null hypothesis (1.6) but reject the second null hypothesis (1.8), this implies CTEs by treatment ratio and HTEs by pretreatment variables. Finally, if we reject both null hypotheses then it suggests HTEs by both variable classes. Hence, implementing the tests using a multiple testing procedure is imperative to disentangling the source of treatment effects heterogeneity.

To control the probability of rejecting at least one null hypothesis, given that they are both true, often called the family-wise error rate (FWER), I recommend a step-wise multiple testing procedure based on Holm's algorithm (Holm (1979)). There are several similar adjustments for multiple testing, but I opt for Holm's procedure because it accounts for the dependency between the two test statistics and it is computational simple.

Let the $p_{d_1} \leq p_{d_2}$ be the ordered p-values, with corresponding null hypotheses H_{0,d_1} and H_{0,d_2}. Then the Holm step-down algorithm is as follows:

Algorithm 1.2.1 (Holm Procedure)

1. *If* $p_{d_1} > \alpha/2$ *fail to reject both* H_{0,d_1} *and* H_{0,d_2} *and stop. If* $p_{d_1} \leq \alpha/2$ *reject* H_{0,d_1} *and test* H_{0,d_2} *at level* α.

2. *If* $p_{d_1} \leq \alpha/2$ *but* $p_{d_2} > \alpha$, *fail to reject* H_{0,d_2} *and stop. If* $p_{d_1} \leq \alpha/2$ *and* $p_{d_2} \leq \alpha$, *reject* H_{0,d_2}.

1.3 Main Asymptotic Results

This section discusses the asymptotic properties of the proposed test statistics $\hat{T}_1$ and $\hat{T}_2$ when the null hypotheses are true and false. I use an asymptotic regime where the number of clusters grow large. I show that appropriate studentized versions of the test statistics $\hat{T}_1$ and $\hat{T}_2$ have asymptotic standard normal null distributions. To begin, I state the assumptions required to develop the asymptotic theory.

Assumption 1.3.1 *(a) The joint distribution of* $(Y, X) \in \mathcal{Y} \times \mathcal{X}$ *is absolutely continuous with respect to the Lebesgue measure; (b) the probability density function* f *of* X *is continuously differentiable almost everywhere; (c)* $\rho_2(\cdot, \pi)$ *is strictly positive and continuous almost everywhere on* $\mathcal{W}_X, \forall \pi \in \mathbf{\Pi}$*, where* $\mathcal{W}_X$ *is a compact subset of* $\mathcal{X}$*; (d)* K *is a product kernel function, i.e.,* $K(u) = \Pi_{j=1}^{d} K_j(u_j), u = (u_1, \ldots, u_d)$*, with each* $K_j : \mathbb{R} \mapsto \mathbb{R}$*,* $j = 1, \ldots, d$*, satisfying that* K_j *is an* s*-order kernel function with support* $\{u \in \mathbb{R} : |u| \leq 0.5\}$*, symmetric around zero, bounded, and is of bounded variation, and integrates to 1, where* s *is an integer that satisfies* $s > 1.5d$*; (e) as functions of* x*,* $\mathbb{E}[Y|X = x, T = t, \Pi = \pi]$*,* $f(x)$*,* $p_t(x, \pi)$ *for* $t = 0, 1$ *are* s*-times continuously differentiable almost everywhere for each* $\pi \in \mathbf{\Pi}$ *with uniformly bounded derivatives; (f)* $\sup_{x \in \mathcal{W}_X} \mathbb{E}[|Y|^3 | X = x, T = t, \Pi = \pi] < \infty$ *for* $t = 0, 1$ *and* $\pi \in \mathbf{\Pi}$*; (g) the bandwidth satisfies* $Nh^{2s} \to 0$*,* $Nh^{3d} \to \infty$ *and* $(Nh^{2d})^{1/2}/logN \to \infty$*, where* $s > 1.5d$*; (h)* $\sup_{(x,x',\pi) \in \mathcal{W}_X^2 \times \mathbf{\Pi}} |\hat{w}(x, x'\pi) - w(x, x'\pi)| = \sup_{(x,\pi,\pi') \in \mathcal{W}_X \times \mathbf{\Pi}^2} |\hat{w}(x, \pi, \pi') - w(x, \pi, \pi')| = o_p(h^{d/2})$.

Even though these are standard regularity conditions in the literature (see Lee, Song, and Whang (2013), and Chang, Lee, and Whang (2015, p. 315)), it is worthwhile to comment on them for completeness. Assumptions 1.3.1(a) and (b) are unnecessary for the asymptotic results. They are convenient assumptions imposing continuity on X and Y that help to present my main results. The results in this chapter generalize to non-continuous X and Y. Assumption 1.3.1(c) ensures that the inverse standard error weight function is continuous and well-defined within a compact subset of $\mathcal{X}$. Assumption 1.3.1(d) imposes conditions on the kernel function. Assumption 1.3.1(e) and (f) imposes restrictions on the underlying true data-generating process to ensure smooth and finite moments. Assumption 1.3.1(g) imposes standard restrictions on the choice of bandwidth. And finally, Assumption 1.3.1(h) ensures that the estimated weight functions are uniformly consistent.

1.3.1 Asymptotic Null Distribution and Properties of the Test Statistics

I first provide a studentized version of the test statistics $\hat{T}_1$ and then derive its asymptotic null distribution. To begin, note that there is a non-vanishing bias term of the kernel estimator $\hat{\tau}(x; \pi)$ for

each $\pi \in \mathbf{\Pi}$ and $x \in \mathcal{X}$ that affects the test statistics. Three common approaches to addressing this issue are (i) explicit bias correction, (ii) the use of higher-order kernels, and (iii) the use of smaller bandwidths (under-smoothing). See Racine (1997) for more details. I settle on the first remedy because it is the theoretically sound approach. The asymptotic bias[5] of $\hat{T}_1$ is

$$a_1 := h^{\frac{-d}{2}} \cdot \mathbb{E}|\mathbb{Z}_1| \cdot \frac{K^2}{2} \cdot \int_{\mathcal{X}} dx, \tag{1.12}$$

where $\mathbb{Z}_1$ is the standard normal distribution. The asymptotic bias only depends on the true data generating process (DGP) through the bounds of X. Hence, we can compute it exactly without estimation.

The other parameter required to studentized $\hat{T}_1$ is the asymptotic standard error of $\hat{T}_1$. Let $\hat{\Gamma}(x; \pi, \pi') = \hat{\tau}(x; \pi) - \hat{\tau}(x; \pi')$, then define the asymptotic variance[6] of $\hat{T}_1$ as

$$\sigma_1^2 := \frac{1}{4} \int_1 \mathrm{Cov}\left(\left| \sqrt{1 - \rho(x,t,\pi_i,\pi_j,\pi_k,\pi_l)^2}\mathbb{Z}_1 + \rho(x,t,\pi_i,\pi_j,\pi_k,\pi_l)\mathbb{Z}_2 \right|, |\mathbb{Z}_2| \right) dxdt, \tag{1.13}$$

where $\int_1 = \int_{\mathbb{R}^d} \int_{T_0} \sum_{i=1}^K \sum_{j=1}^K \sum_{k=1}^K \sum_{l=1}^K$, $T_0 = [-1, 1]^d$ and $\rho(x, t, \pi_i, \pi_j, \pi_k, \pi_l)$ is the *unknown* correlation between $\hat{\Gamma}(x; \pi_i, \pi_j)/\sqrt{\hat{\rho}_2(x, \pi_i, \pi_j)}$ and $\hat{\Gamma}(x; \pi_k, \pi_l)/\sqrt{\hat{\rho}_2(x, \pi_k, \pi_l)}$, (i.e., $\rho(x, t, \pi_i, \pi_j, \pi_k, \pi_l) = \mathrm{Corr}(\hat{\Gamma}(x; \pi_i, \pi_j)/\sqrt{\hat{\rho}_2(x, \pi_i, \pi_j)}, \hat{\Gamma}(x; \pi_k, \pi_l)/\sqrt{\hat{\rho}_2(x, \pi_k, \pi_l)})$). $\mathbb{Z}_1$ and $\mathbb{Z}_2$ are mutually independent standard normal random variables. A kernel estimator of $\rho(x, t, \pi_i, \pi_j, \pi_k, \pi_l)$ is defined as

$$\hat{\rho}(x,t,\pi_i,\pi_j,\pi_k,\pi_l) = \begin{cases} \frac{\int K(\xi)K(\xi+t)d\xi}{\int K(\xi)^2 d\xi} & i = k \,\&\, j = l \\ \frac{-\hat{\rho}_2(x,\pi_j)}{\sqrt{\hat{\rho}_2(x,\pi_i,\pi_j)}\sqrt{\hat{\rho}_2(x,\pi_k,\pi_l)}} \cdot \frac{\int K(\xi)K(\xi+t)d\xi}{\int K(\xi)^2 d\xi} & j = k \,\&\, i \neq l \\ \frac{-\hat{\rho}_2(x,\pi_i)}{\sqrt{\hat{\rho}_2(x,\pi_i,\pi_j)}\sqrt{\hat{\rho}_2(x,\pi_k,\pi_l)}} \cdot \frac{\int K(\xi)K(\xi+t)d\xi}{\int K(\xi)^2 d\xi} & j \neq k \,\&\, i = l \\ \frac{\hat{\rho}_2(x,\pi_j)}{\sqrt{\hat{\rho}_2(x,\pi_i,\pi_j)}\sqrt{\hat{\rho}_2(x,\pi_k,\pi_l)}} \cdot \frac{\int K(\xi)K(\xi+t)d\xi}{\int K(\xi)^2 d\xi} & j = l \,\&\, i \neq k \\ \frac{\hat{\rho}_2(x,\pi_i)}{\sqrt{\hat{\rho}_2(x,\pi_i,\pi_j)}\sqrt{\hat{\rho}_2(x,\pi_k,\pi_l)}} \cdot \frac{\int K(\xi)K(\xi+t)d\xi}{\int K(\xi)^2 d\xi} & j \neq l \,\&\, i = k \\ 0 & i \neq k \,\&\, j \neq l, \end{cases} \tag{1.14}$$

where $\hat{\rho}_2(x, \pi, \pi') := \hat{\rho}_2(x, \pi) + \hat{\rho}_2(x, \pi')$. Plug $\hat{\rho}(x, t, \pi_i, \pi_j, \pi_k, \pi_l)$ into the right hand side of (1.13), and obtain the asymptotic variance estimator

$$\hat{\sigma}_1^2 := \frac{1}{4} \int_1 \mathrm{Cov}\left(\left| \sqrt{1 - \hat{\rho}(x,t,\pi_i,\pi_j,\pi_k,\pi_l)^2}\mathbb{Z}_1 + \hat{\rho}(x,t,\pi_i,\pi_j,\pi_k,\pi_l)\mathbb{Z}_2 \right|, |\mathbb{Z}_2| \right) dxdt. \tag{1.15}$$

[5] See Appendix 1.7.3 for the derivation.

[6] See Appendix 1.7.3 for the derivation.

Given the expressions for the estimated asymptotic variance and bias in (1.12) and (1.15), define the studentized version of the test statistic $\hat{T}_1$ as

$$\hat{S}_1 := \frac{\hat{T}_1 - a_1}{\hat{\sigma}_1}.$$

Similarly, define the studentized version of the test statistic $\hat{T}_2$ as

$$\hat{S}_2 := \frac{\hat{T}_2 - a_2}{\sigma_2},$$

where

$$a_2 := h^{\frac{-d}{2}} \cdot \mathbb{E}|\mathbb{Z}_1| \cdot \frac{K}{2} \cdot \int_{\mathcal{X}} \int_{\mathcal{X}} dx dx',$$

and

$$\sigma_2^2 := \frac{h^{2d}}{4} \int_2 \text{Cov}\left(\left| \sqrt{1 - \rho_1^*(x,q,r,s,\pi_k)^2}\mathbb{Z}_1 + \rho_1^*(x,q,r,s,\pi_k)\mathbb{Z}_2 \right|, |\mathbb{Z}_2| \right) drdsdqdx, \tag{1.16}$$

with $\int_2 := \int_{\mathcal{X}} \int_{T_0} \int_{T_0} \sum_{k=1}^{K}$, and

$$\rho_1^*(x,q,r,s,\pi) := \frac{\rho_1(x,r,\pi) - \rho_1(x,s,\pi) - \rho_1(x,q,r,\pi) + \rho_1(x,q,s,\pi)}{2\sqrt{(\rho_2(x,\pi) - \rho_1(x,q,\pi))(\rho_2(x,\pi) - \rho_1(x,r,s,\pi))}}$$

which reduces to

$$\rho_1^*(x,q,r,s,\pi) := \frac{\int K(\xi + r)K(\xi) - K(\xi + s)K(\xi) - K(\xi + q)K(\xi + r) + K(\xi + q)K(\xi + s)d\xi}{2 \cdot \sqrt{(\int K^2(\xi) - K(\xi + q)K(\xi)d\xi)(\int K^2(\xi) - K(\xi + q)K(\xi + s)d\xi)}}$$

since

$$\rho_1(x,r,s,\pi) := \left\{ \sum_{t \in \{0,1\}} \frac{\mathbb{E}[Y^2|X = x, T = t, \Pi = \pi] - (\mathbb{E}[Y|X = x, T = t, \Pi = \pi])^2}{P_t(x,\pi)} \right\} \cdot \int K(\xi + s)K(\xi + r)d\xi.$$

Theorem 1.3.1 *Let Assumptions 1.2.1 - 1.3.1 hold, then under the*
(i) null hypothesis (1.6) $\hat{S}_1$ *converges to the standard normal distribution, i.e.,* $\hat{S}_1 \to N(0,1)$ *;*
(ii) null hypothesis (1.8) $\hat{S}_2$ *converges to the standard normal distribution, i.e.,* $\hat{S}_2 \to N(0,1)$.

Theorem 1.3.1 implies that we can compute the critical values of the tests from the standard normal distribution. This theorem forms an integral part of most of the asymptotic properties I discuss

below. First, I show that the tests have asymptotically valid sizes in the following theorem.

Theorem 1.3.2 *Let Assumptions 1.2.1 - 1.3.1 hold, then under the*
(i) null hypothesis (1.6)

$$\lim_{N\to\infty} \Pr(\hat{S}_1 > z_{1-\alpha}) = \alpha;$$

(ii) null hypothesis (1.8)

$$\lim_{N\to\infty} \Pr(\hat{S}_2 > z_{1-\alpha}) = \alpha$$

Theorem 1.3.2 shows that the test statistics $\hat{S}_1$ and $\hat{S}_2$ have correct sizes asymptotically under H_0^{Π} and H_0^{X} respectively. Hence, the following decision rule suffices: For $j = 1, 2$ reject the null hypothesis if $\hat{S}_j > z_{1-\alpha}$, where $z_{1-\alpha}, \alpha \in [0, 1]$ is the $(1-\alpha)^{th}$ quantile (critical value) obtained from the standard normal distribution.

1.3.2 Power Properties of the Test statistics

In this subsection, I investigate the power properties of the test statistics against a fixed and a sequence of local alternatives. First, I establish that $\hat{S}_1$ and $\hat{S}_2$ are consistent against the following fixed alternatives

$$H_1^{\Pi} : \int_{\mathcal{X}} \sum_{k=1}^{K} \sum_{j=1}^{K} \left\{ \sqrt{N} |\tau(x;\pi_k) - \tau(x;\pi_j)| \right\} \frac{w(x,\pi_k,\pi_j)}{2} dx > 0, \tag{1.17}$$

$$H_1^{X} : \int_{\mathcal{X}} \int_{\mathcal{X}} \sum_{k=1}^{K} \left\{ \sqrt{N} |\tau(x;\pi_k) - \tau(x';\pi_k)| \right\} \frac{w(x,x',\pi_k)}{2} dx dx' > 0 \tag{1.18}$$

respectively.

Theorem 1.3.3 *Let Assumptions 1.2.1 - 1.3.1 hold, then*
(i) under the fixed alternative hypothesis (1.17)

$$\lim_{N\to\infty} \Pr(\hat{S}_1 > z_{1-\alpha}) = 1, \quad \textit{and}$$

(ii) under the fixed alternative hypothesis (1.18)*,*

$$\lim_{N\to\infty} \Pr(\hat{S}_2 > z_{1-\alpha}) = 1.$$

Theorem 1.3.3 establishes that the proposed test statistics are powerful against a fixed alternative. Next, I show that the proposed tests $\hat{S}_1$ and $\hat{S}_2$ can detect a sequence of local alternatives converging to the null hypotheses at the rate $N^{-1/2}h^{-d/4}$. Specifically, consider the following sequences of local alternatives

$$H_a^{\Pi} : \tau(x,\pi) - \tau(x,\pi') = N^{-1/2}h^{-d/4}\delta_1(x,\pi,\pi') \ \forall x \in \mathcal{X}, \pi, \pi' \in \Pi \tag{1.19}$$

$$H_a^{X} : \tau(x,\pi) - \tau(x',\pi) = N^{-1/2}h^{-d/4}\delta_2(x,x',\pi) \ \forall x, x' \in \mathcal{X}, \pi \in \Pi, \tag{1.20}$$

converging to the null hypotheses H_a^{Π} and H_a^{X} respectively where for $j = 1, 2,\ \delta_j(\cdot,\cdot,\cdot)$ is a real bounded function satisfying
$\int_{\mathcal{X}} \sum_{\pi\in\Pi} \sum_{\pi'\in\Pi} |\delta_1(x,\pi,\pi')| w(x,\pi,\pi')dx > 0$, and $\int_{\mathcal{X}} \int_{\mathcal{X}} \sum_{\pi\in\Pi} |\delta_2(x,x',\pi)| w(x,x',\pi)dxdx' > 0$.

Theorem 1.3.4 *Let Assumptions 1.2.1 - 1.3.1 hold, then*
(i) under the sequences of alternative hypotheses (1.19),

$$\lim_{N\to\infty} \Pr(\hat{S}_1 > z_{1-\alpha}) = 1 - \Phi\left(z_{1-\alpha} - \frac{1}{2\sqrt{2\pi}\sigma_1} \int_{\mathcal{X}} \sum_{k=1}^{K} \sum_{j=1}^{K} \delta^2(x,\pi_k,\pi_j)dx\right),$$

and;
(ii) under the sequences of alternative hypotheses (1.20),

$$\lim_{N\to\infty} \Pr(\hat{S}_2 > z_{1-\alpha}) = 1 - \Phi\left(z_{1-\alpha} - \frac{1}{2\sqrt{2\pi}\sigma_2} \int_{\mathcal{X}} \int_{\mathcal{X}} \sum_{j=1}^{K} \delta^2(x,x',\pi_k)dxdx'\right),$$

where Φ denotes the cumulative distribution function (CDF) of the standard normal distribution.

Theorem 1.3.4 demonstrates that the test statistics have statistical power greater than zero against local alternatives that converge to the null hypothesis at the rate $N^{-1/2}h^{-d/4}$.

1.4 Bootstrap Approach

I introduce bootstrap resampling methods to obtain the null distributions of $\hat{S}_1$ and $\hat{S}_2$ in this section. Although Theorem 1.3.1 shows that $\hat{S}_1$ and $\hat{S}_2$ have asymptotic standard normal null distributions, the simulation results in Appendix 1.7.1 suggest that these limiting distributions are poor approximations of the finite sample null distributions when the sample size is small. Moreover, decisions informed by the asymptotic inference approach may be sensitive to bandwidth choice. This is because the null distributions of test statistics do not depend on the bandwidth, yet the values

of the test statistics are directly affected by the bandwidth (Racine, 1997, p. 369). On the other hand, the main demerit of bootstrap resampling techniques is their high computing cost. Nevertheless, for small sample sizes, the gains in accurate decision-making could offset these costs. As a result, I recommend the following bootstrap procedures for practitioners in settings where the sample size is small (or when there exist clusters with small sample sizes).

1.4.1 The Bootstrap Resampling Algorithm for $\hat{S}_1$

Let $W_i = (X_i, T_i, Y_i)$ denote the vector of variables for the i^{th} unit. Therefore, the pooled sample across all treatment ratios can be written as $\{W_i\}_{i=1}^{\sum_{k=1}^{K} N_k}$. The bootstrapping algorithm to generate the null distribution of $\hat{S}_1$ is as follows:

1. For $k = 1 \dots K$, randomly draw N_k observations from the pooled sample $\{W_i\}_{i=1}^{\sum_{k=1}^{K} N_k}$ with replacement and denote the resulting bootstrapped sample combined with a new variable $\Pi = \pi_k$ as $W^*(\pi_k) := \{W_i^*, \pi_k\}_{i=1}^{N_k}$.
2. Compute the test statistic $\hat{S}_1^* = \hat{T}_1^* - a_1^*$ using the pooled bootstrapped data $\{W^*(\pi_k)\}_{k=1}^{K}$, where the definition of $\hat{T}_1^*$ and a_1^* are the same as $\hat{T}_1$ and a_1 respectively. Comparing $\hat{S}_1^*$ to its asymptotic counterpart, note that I omit the standard error term in the denominator. Simulation results not reported show that omitting the scaling term has negligible impact on empirical power and size. On the other hand, it reduces computation time.
3. Repeat 1 and 2 a large number of times (say B_1 times) and use the empirical distribution of the B_1 bootstrapped test statistics $\{\hat{S}_{1,b}^*\}_{b=1}^{B_1}$ to approximate the null distribution of $\hat{T}_1 - a_1$.
4. Compute the empirical p-value as $\hat{p}* = B_1^{-1} \sum_{b=1}^{B_1} \mathbb{1}(\hat{S}_{1,b}^* > \hat{S}_1^o)$ where $\hat{S}_1^o$ is the test statistic computed via the original data.

1.4.2 The Bootstrap Resampling Algorithm for $\hat{S}_2$

Akin to $\hat{S}_1$, I propose the following bootstrap algorithm for the null distribution of $\hat{S}_2$:

1. Estimate a "restricted" conditional mean $\mathbb{E}(Y|X = \bar{x}_i, \Pi = \pi_i, T = t_i)$. Let the resulting fitted values be $\hat{M}(\bar{x}_i, \pi_i, t_i), i = 1 \dots N$. This restricted conditional mean does not vary by the pretreatment variables X, because they are held constant at their average value $\bar{x}$.
2. Obtain the residuals $\hat{\varepsilon}_i = Y_i - \hat{M}(\bar{x}_i, \pi_i, t_i)$, $i = 1 \dots N$. Demean the residuals using the sample average residual. Since we compute these residuals using the restricted conditional means, they are residuals obtained under the null hypothesis H_0^X.
3. Draw a random sample of size N from the centered residuals with replacement and name it the bootstrap residual sample $\{\hat{\varepsilon}_i^*\}_{i=1}^{N}$.

4. Now generate a dependent variable $Y_i^* = \hat{M}(\bar{x}_i, \pi_i, t_i) + \hat{\varepsilon}_i^*$, $i = 1 \dots N$. With these dependent variables (under H_0^X), create a null bootstrap sample $(Y_i^*, X_i, \Pi_i, T_i), i = 1 \dots N$, where $(X_i, \Pi_i, T_i), i = 1 \dots N$ are from the original sample.
5. Compute the test statistic $\hat{S}_2^* = \hat{T}_2^* - a_2^*$ using the null bootstrap sample where I define $\hat{T}_2^*$ and a_2^* the same way as $\hat{T}_2$ and a_2 respectively. Here also, I omit the denominator standard error term based on the same argument used for $\hat{S}_1^*$
6. Repeat 3–5 many times (say B_2 times) and use the empirical distribution of the B_2 bootstrapped test statistics $\{\hat{S}_{2,b}^*\}_{b=1}^{B_2}$ to approximate the null distribution of $\hat{T}_1 - a_2$.
7. Compute the empirical p-value as $\hat{p}* = B_2^{-1} \sum_{b=1}^{B_2} \mathbb{1}(\hat{S}_{2,b}^* > \hat{S}_2^o)$ where $\hat{S}_2^o$ is the test statistic computed via the original data.

In Table 1.8, I compare the empirical sizes of the bootstrap-based test statistics with their asymptotic counterparts when the sample size is small, specifically when $N = 200\,(N_k = 50)$. The result shows that the empirical size computed using the bootstrapping algorithms is closer to the nominal probabilities than their asymptotic counterparts. It confirms that the limiting distributions are poor approximations of the finite sample null distributions when the sample size is small.

1.5 Monte Carlo Simulation

I present the design and results of two Monte Carlo experiments in this section. Other simulation designs and results are deferred to Appendix 1.7.1. First, I numerically examine the empirical size and power properties of the proposed test statistics. And secondly, I design a simulation exercise to compare the proposed test statistics to their parametric counterparts.

1.5.1 Empirical Size and Statistical power

I design Monte Carlo experiments to examine the empirical rejection probabilities of the proposed test statistics. All the rejection probabilities are computed using 1000 replications. In each experiment, I consider a single pretreatment variable X[7] drawn from the uniform $[0, 1]$ distribution. Each cluster is assigned one of four treatment indicator variables $T(k), k = 1 \dots, 4$. $T(k)$ follows a binomial distribution with probability π_k where $\{\pi_1, \pi_2, \pi_3, \pi_4\} = \{0.3, 0.4, 0.5, 0.6\} = \mathbf{\Pi}$. The realized outcome Y is constructed by

$$Y = (\tau(X, \Pi) + U_1) \times T + U_0 \times (1 - T),$$

[7] I extend the experiment to the case with multivariate pretreatment variables in Appendix 1.7.2.

where U_1 and U_0 are independent normals with a mean of zero and variance of 0.1. The general specification of $\tau(x, \pi)$ is

$$\tau(x, \pi) = \beta_0 x + \beta_1 \pi + \beta_2 \cdot (x\pi).$$

In the experiments, I consider a uniform grid in the interval between the 10^{th} and 90^{th} percentiles of X. This way, I avoid the boundary bias issue associated with the kernel estimators. Using the uniform grid, I compute the integrals in the test statistics with the composite trapezoid rule.

I use the following kernel function that satisfies Assumption 1.3.1(d):

$$K(u) = 1.5(1 - (2u)^2) \cdot \mathbb{1}\{|u| \leq 0.5\}, \tag{1.21}$$

and a bandwidth

$$h = C_h \hat{s}_X N^{-2/7}, \tag{1.22}$$

where $\hat{s}_X$ is the sample standard deviation of the X and C_h is a constant. Finally, the sample size is $N = 1200$ with equal units in for each π (i.e., 300 units for each treatment ratio).

To obtain the empirical rejection probabilities of $\hat{S}_1$, set $\beta_0 = 1$ and $\beta_1 = 0$. Therefore, $\tau(x, \pi) = x + \beta_2 \cdot (x\pi)$, and when $\beta_2 = 0$, the null hypothesis (1.6) of CTEs by treatment ratio is true. As β_2 deviates further from 0 in both directions, the null hypothesis deviates further away from the truth. I report the empirical rejection probabilities for values of β_2 in the range $[-1, 1]$, with 0.1 increments.

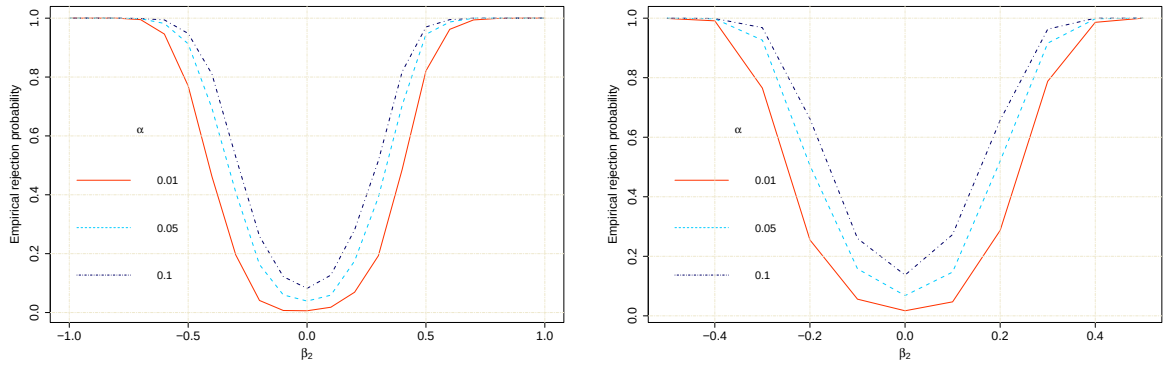
Similarly, to obtain the empirical rejection probabilities of $\hat{S}_2$, set $\beta_0 = 0$, and $\beta_1 = 1$. Therefore, $\tau(x, \pi) = \pi + \beta_2 \cdot (x\pi)$, and when $\beta_2 = 0$, the null hypothesis (1.8) of CTEs by pretreatment variables is true. Here also, I report the empirical rejection probabilities for each β_2 in the range $[-0.5, 0.5]$, with 0.1 increments.

Focusing on the asymptotic method of inference, I provide a summary of the empirical rejection probabilities of the test statistics in Figure 1.2 (and in Tables 1.4–1.5 in Appendix 1.7.1). In each panel, the three graphs represent the rejection probabilities at the 1%, 5%, and 10% nominal levels. The left panel of Figure 1.2 reports the empirical rejection probabilities of $\hat{S}_1$ and the right panel of Figure 1.2 reports those of $\hat{S}_2$. Note that when $\beta_2 = 0$, the empirical rejection probabilities are close to the nominal probabilities which confirms the theoretical results in Theorem 1.3.2. On the contrary, as β_2 deviates towards ± 1, the rejection probabilities approaches 1 which is in line with the consistency results in Theorem 1.3.3.

Similarly, using the bootstrap method, I summarize the empirical rejection probabilities in Figure 1.3 (and in Tables 1.6–1.7 in Appendix 1.7.1). The results are based on 399 bootstrap resamples. Compared to the asymptotic-based empirical rejection probabilities, the differences in the probabilities are negligible. It confirms that the bootstrap algorithm works and the test statistics

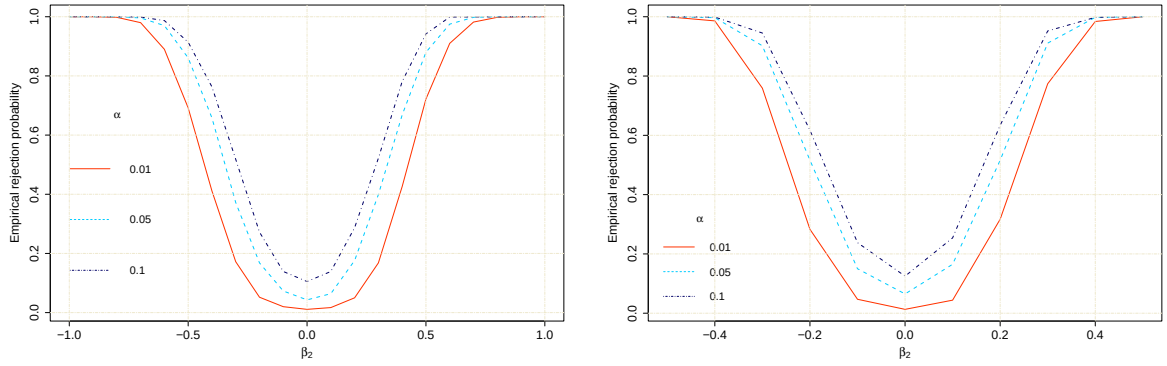
under the null hypotheses "converges in bootstrap distribution" to the standard normal distribution as the sample size increases.

Figure 1.2: Empirical Rejection Probabilities using Asymptotic Method.



(a) Power curve for $\hat{S}_1$ when β_2 lies between -1 and 1. (b) Power curve for $\hat{S}_2$ when β_2 lies between -0.5 and 0.5.

Figure 1.3: Empirical Rejection Probabilities using the Bootstrap Method.



(a) Power curve for $\hat{S}_1^*$ when β_2 lies between -1 and 1. (b) Power curve for $\hat{S}_2^*$ when β_2 lies between -0.5 and 0.5.

1.5.2 Parametric Testing and Misspecification

In this subsection, I design an experiment to show that parametric tests of the null hypotheses (1.6) and (1.8) may be misleading because parametric models are always misspecified to a certain degree. Here, I focus on asymptotic method of inference only. To begin, suppose the sample data at hand is

$\{Y_i, X_i, T_i, \Pi_i\}$ of size N=600, and $\Pi_i \in \{0.3, 0.6\}$. Without knowledge of the true DGP, I estimate the following linear regression model using the method of least squares:

$$Y_i = \beta_0 + \beta_1 T_i + \beta_2 X_i + \beta_3 \mathbb{1}(\Pi_i = 0.3) + \beta_4 T_i \cdot X_i + \beta_5 T_i \cdot \mathbb{1}(\Pi_i = 0.3) + \varepsilon_i. \tag{1.23}$$

The summary results in Tables 1.2 and 1.3 show that the parameters of interest β_4 and β_5 are insignificant when we use either clustered or Heteroscedasticity and Autocorrelation Consistent (HAC) standard errors. Hence, we conclude that treatment effects do not vary by treatment ratio Π, and the pretreatment variable X. Now, I test hypotheses (1.6) and (1.8) using the proposed nonparametric test statistics. I use the kernel function in (1.21) and the plug-in bandwidth selection method in (1.22). Table 1.1 summarizes the results of the two tests at different bandwidth choices (different Cs in the bandwidth formula in (1.22)). The results are unequivocal rejections of the null hypotheses of CTEs by the treatment ratio Π, and the pretreatment variable X.

Now note that the true DGP for the sample data is

$$Y = (\tau(X, \Pi) + U_1) \times T + U_0 \times (1 - T),$$

where U_1 and U_0 are independent normal random variables with a mean of zero and variance of 0.1. In addition, CATE has the following specification:

$$\tau(x, \pi) = 30 \cdot \cos(2\pi x) \cdot (\pi^2 - \pi)$$

and it varies in a highly non-linear way in both X and Π. Therefore, this is a confirmation that a misspecification of the functional form of the conditional mean function in parametric models results in erroneous inference of HTEs.

Table 1.1: Summary of Test Results for Simulated DGP based on Proposed Nonparametric Test

Bandwidth (h)	H_0^{Π}: CTEs across Π		H_0^{X}: CTEs across X	
	$\hat{S}_1$	p-value	$\hat{S}_2$	p-value
0.143	9.510	0.00	133.77	0.00
0.167	7.895	<0.01	96.425	0.00
0.191	6.717	<0.01	72.494	0.00
0.214	5.854	<0.01	56.562	0.00

Table 1.2: Summary of Test Results for Simulated DGP based on Parametric Tests using Clustered Standard Errors

	Estimates	Std. Error	t value	p-value
Constant	-0.020	0.004	-4.664	$3.840e^{-6}$ ***
T	-0.009	0.604	-0.016	0.988
X	0.031	0.008	3.930	$9.485e^{-5}$ ***
$\mathbb{1}(\Pi = 0.3)$	0.023	0.001	19.970	$< 2.2e^{16}$ ***
$T \cdot X$	-0.674	1.704	-0.396	0.693
$T \cdot \mathbb{1}(\Pi = 0.3)$	0.474	0.509	0.932	0.3517
Observations	600			
Note:	*p<0.1; **p<0.05; ***p<0.01. Std. Errors clustered by Π.			

1.6 Conclusion

The nonparametric tests I develop here allow for valid asymptotic and bootstrap inference for heterogeneous treatment effects in the presence of clustered interference. Importantly, the proposed tests help to disentangle the source of variation in the treatment effects. The test statistics are sums of weighted L_1-norm differences in consistent nonparametric kernel estimators of conditional average treatment effects. I show that the test statistics have correct sizes and (asymptotic) standard normal null distributions. Moreover, they are consistent under fixed and sequences of local alternatives.

In addition, I propose a bootstrap method for small sample sizes, and I show numerically that the bootstrap algorithm works for a given DGP. Monte Carlo simulation exercises corroborate the theoretical findings.

Table 1.3: Summary of Test Results for Simulated DGP based on Parametric Tests using Heteroscedasticity and Autocorrelation Consistent (HAC) Standard Errors

	Estimates	Std. Error	t value	p-value
Constant	-0.020	0.021	-0.960	0.338
T	-0.009	1.103	-0.009	0.993
X	0.031	0.018	1.687	0.092 *
$\mathbb{1}(\Pi = 0.3)$	0.023	0.041	0.557	0.578
$T \cdot X$	-0.674	1.129	-0.597	0.551
$T \cdot \mathbb{1}(\Pi = 0.3)$	0.474	2.115	0.224	0.823
Observations	600			
Note:	*p<0.1; **p<0.05; ***p<0.01.			

Bibliography

Aronow, P. M., C. Samii, et al. (2017). Estimating average causal effects under general interference, with application to a social network experiment. *The Annals of Applied Statistics 11*(4), 1912–1947.

Athey, S., D. Eckles, and G. W. Imbens (2018). Exact p-values for network interference. *Journal of the American Statistical Association 113*(521), 230–240.

Baird, S., J. A. Bohren, C. McIntosh, and B. Özler (2018). Optimal design of experiments in the presence of interference. *Review of Economics and Statistics 100*(5), 844–860.

Billingsley, P. (1968). Convergence of probability measures.

Bitler, M. P., J. B. Gelbach, and H. W. Hoynes (2006). What mean impacts miss: Distributional effects of welfare reform experiments. *American Economic Review 96*(4), 988–1012.

Chang, M., S. Lee, and Y.-J. Whang (2015). Nonparametric tests of conditional treatment effects with an application to single-sex schooling on academic achievements. *The Econometrics Journal 18*(3), 307–346.

Colpitts, T. (2002). Targeting reemployment services in canada. *Targeting Employment Services. Kalamazoo, MI: WE Upjohn Institute for Employment Research*, 283–302.

Cox, D. R. (1958). Planning of experiments.

Crump, R. K., V. J. Hotz, G. Imbens, and O. A. Mitnik (2006). Nonparametric tests for treatment effect heterogeneity.

Ding, P., A. Feller, and L. Miratrix (2016). Randomization inference for treatment effect variation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 78*(3), 655–671.

Giné, E., D. M. Mason, and A. Y. Zaitsev (2003). The \bm {L} _\mathbf {1}-norm density estimator process. *The Annals of Probability 31*(2), 719–768.

Han, S., J. Owusu, and Y. Shin (2022). Statistical treatment rules under social interaction. *arXiv preprint arXiv:2209.09077*.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, 65–70.

Imbens, G. W. and D. B. Rubin (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Lee, S. and A. M. Shaikh (2014). Multiple testing and heterogeneous treatment effects: re-evaluating the effect of progresa on school enrollment. *Journal of Applied Econometrics 29*(4), 612–626.

Lee, S., K. Song, and Y.-J. Whang (2013). Testing functional inequalities. *Journal of Econometrics 172*(1), 14–32.

Lee, S. and Y.-J. Whang (2009). Nonparametric tests of conditional treatment effects.

Li, Q., E. Maasoumi, and J. S. Racine (2009). A nonparametric test for equality of distributions with mixed categorical and continuous data. *Journal of Econometrics 148*(2), 186–200.

Manski, C. F. (2013). Identification of treatment response with social interactions. *The Econometrics Journal 16*(1), S1–S23.

Neyman, J. (1923). Sur les applications de la théorie des probabilités aux experiences agricoles: Essai des principes. *Roczniki Nauk Rolniczych 10*, 1–51.

Racine, J. (1997). Consistent significance testing for nonparametric regression. *Journal of Business & Economic Statistics 15*(3), 369–378.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology 66*(5), 688.

Sant'Anna, P. H. (2021). Nonparametric tests for treatment effect heterogeneity with duration outcomes. *Journal of Business & Economic Statistics 39*(3), 816–832.

Shergin, V. (1993). Central limit theorem for finitely-dependent random variables. *Journal of Soviet Mathematics 67*(4), 3244–3248.

Sobel, M. E. (2006). What do randomized studies of housing mobility demonstrate? causal inference in the face of interference. *Journal of the American Statistical Association 101*(476), 1398–1407.

Sweeting, T. J. (1977). Speeds of convergence for the multidimensional central limit theorem. *The Annals of Probability*, 28–41.

Wager, S. and S. Athey (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association 113*(523), 1228–1242.

1.7 Appendix

1.7.1 Simulation Results

1.7.1.1 Asymptotic-based Inference

Table 1.4: Empirical Rejection Probabilities: $N = 1200$, Bandwidth=$C_h \hat{s}_X n^{-2/7}$ and $\Pi = (0.3, 0.4, 0.5, 0.6)$.

		Nominal probabilities		
Test statistic	β	0.01	0.05	0.10
$\hat{S}_1$	-1	1.000	1.000	1.000
	-0.9	1.000	1.000	1.000
	-0.8	0.995	0.998	1.000
	-0.7	0.954	0.988	0.993
	-0.6	0.849	0.946	0.981
	-0.5	0.602	0.795	0.875
	-0.4	0.324	0.554	0.680
	-0.3	0.123	0.312	0.440
	-0.2	0.032	0.121	0.200
	-0.1	0.017	0.061	0.123
	0.0	0.006	0.035	0.081
	0.1	0.015	0.061	0.122
	0.2	0.052	0.165	0.257
	0.3	0.149	0.322	0.447
	0.4	0.339	0.591	0.710
	0.5	0.663	0.852	0.915
	0.6	0.896	0.965	0.987
	0.7	0.980	0.993	0.995
	0.8	0.995	0.998	1.000
	0.9	0.999	1.000	1.000
	1	1.000	1.000	1.000

Table 1.5: Empirical Rejection Probabilities: $N = 1200$, Bandwidth= $C_h \hat{s}_X N^{-2/7}$ and $\Pi = (0.3, 0.4, 0.5, 0.6)$.

		Nominal probabilities		
Test statistic	β	0.01	0.05	0.10
$\hat{S}_2$	-0.5	0.999	1.000	1.000
	-0.4	0.991	0.998	0.999
	-0.3	0.765	0.926	0.968
	-0.2	0.255	0.505	0.662
	-0.1	0.056	0.158	0.260
	0.0	0.017	0.068	0.137
	0.1	0.047	0.147	0.273
	0.2	0.287	0.520	0.657
	0.3	0.788	0.916	0.963
	0.4	0.986	0.998	1.000
	0.5	1.000	1.000	1.000

1.7.1.2 Bootstrap-based Inference

Table 1.6: Empirical Rejection Probabilities: $N = 1200$, Bandwidth=$C_h \hat{s}_X n^{-2/7}$ and $\Pi = (0.3, 0.4, 0.5, 0.6)$.

		Nominal probabilities		
Test statistic	β	0.01	0.05	0.1
$\hat{S}_1^*$	-1	1.000	1.000	1.000
	-0.9	1.000	1.000	1.000
	-0.8	0.998	1.000	1.000
	-0.7	0.980	0.996	0.999
	-0.6	0.890	0.970	0.987
	-0.5	0.692	0.861	0.914
	-0.4	0.411	0.658	0.763
	-0.3	0.172	0.372	0.518
	-0.2	0.052	0.168	0.272
	-0.1	0.020	0.073	0.139
	0.0	0.011	0.043	0.105
	0.1	0.017	0.064	0.139
	0.2	0.050	0.176	0.287
	0.3	0.168	0.399	0.522
	0.4	0.426	0.668	0.781
	0.5	0.722	0.880	0.942
	0.6	0.910	0.975	0.999
	0.7	0.982	0.998	0.999
	0.8	0.998	1.000	1.000
	0.9	1.000	1.000	1.000
	1	1.000	1.000	1.000

Table 1.7: Empirical Rejection Probabilities: $N = 1200$, Bandwidth=$C_h \hat{s}_X N^{-2/7}$ and $\Pi = (0.3, 0.4, 0.5, 0.6)$.

		Nominal probabilities		
Test statistic	β	0.01	0.05	0.10
$\hat{S}_2^*$	-0.5	1.000	1.000	1.000
	-0.4	0.986	0.997	0.999
	-0.3	0.759	0.902	0.945
	-0.2	0.283	0.516	0.617
	-0.1	0.047	0.150	0.238
	0.0	0.013	0.065	0.125
	0.1	0.044	0.165	0.254
	0.2	0.317	0.516	0.634
	0.3	0.774	0.911	0.952
	0.4	0.984	0.998	0.998
	0.5	1.000	1.000	1.000

Table 1.8: Comparison of Empirical Size for the Bootstrap and Asymptotic-based Testing Approach when Sample Size is Small: $N = 200$, bandwidth=$3 \cdot \hat{s}_X N^{-2/7}$ and $\Pi = (0.3, 0.4, 0.5, 0.6)$.

Nominal probabilities	Test statistic for H_0^{Π}		Test statistic for H_0^{X}	
	Bootstrap-based	Asymptotic-based	Bootstrap-based	Asymptotic-based
0.01	0.008	0.350	0.027	0.033
0.05	0.053	0.498	0.085	0.143
0.1	0.102	0.579	0.152	0.278

1.7.2 Extension of the Monte Carlo Simulation Experiment to Multivariate Covariates

I extend the experiment in Section 1.5 to multivariate pretreatment variables. For both test statistics, each pretreatment variable X_d, $d \geq 1$ is drawn independently from the standard uniform distribution. I have two clusters and the cluster treatment indicator variable $T_k, k = 1, 2$ is drawn independently from a Bernoulli distribution with probability $\pi_k, k = 1, 2$ where $\mathbf{\Pi} = \{\pi_1, \pi_2\} = \{0.3, 0.4\}$. I use the Monte Carlo integration technique to compute integrals in the test statistics. I consider the following general functional forms of the CATE:

$$\tau(\mathbf{x}, \pi) = \beta_0 \sum_{l=1}^{d} x_l + \beta_1 \pi - \beta_2 \cdot \left(\pi \sum_{l=1}^{d} x_l \right). \tag{1.24}$$

Focusing on $\hat{S}_1$, fix $\beta_0 = 1$ and $\beta_1 = 0$ and consider two specifications of β_2: $\beta_2 = 0$ (i.e., the null hypothesis of CTEs by Π is true); and $\beta_2 = 1$, (which implies that the null hypothesis of CTEs by Π is false). Naturally, due to the curse of dimensionality associated with kernel estimation, one should expect a poor performance of the tests when the dimension of continuous variables increases. As as result, in this simulation exercise, I consider two cases: $d = 2$ and 3. In Table 1.9, I report the rejection probabilities of $\hat{S}_1$ under the two β_2 specifications, i.e., $\beta_2 = 0$ and $\beta_2 = 1$ which represents the empirical size and power respectively.

Table 1.9: Empirical Size and Power of $\hat{S}_1$ using Multivariate X. $N = 600$, Bandwidth=$5 \cdot \hat{s}_X N^{-2/7}$ and $\Pi = (0.3, 0.4)$.

Nominal probabilities	$d = 2$		$d = 3$	
	Size	Power	Size	Power
0.01	0.018	0.851	0.005	0.838
0.05	0.042	0.911	0.119	0.915
0.10	0.071	0.941	0.172	0.942

Next, I focus on $\hat{S}_2$. Using the CATE in (1.24), fix $\beta_0 = 0$ and $\beta_1 = 1$ and consider two specifications of β_2: $\beta_2 = 0$ (i.e., the null hypothesis of CTEs by X is true); and $\beta_2 = 0.5$, (which implies that the null hypothesis of CTEs by X is false). In Table 1.10, I report the rejection probabilities of $\hat{S}_2$ under the two β_2 specifications.

Table 1.10: Empirical Size and Power of $\hat{S}_2$ using Multivariate X. $N = 600$, Bandwidth=$5 \cdot \hat{s}_X N^{-2/7}$ and $\Pi = (0.3, 0.4)$.

Nominal probabilities	$d = 2$		$d = 3$	
	Size	Power	Size	Power
0.01	0.000	0.968	0.000	0.999
0.05	0.002	0.990	0.019	1.000
0.10	0.008	0.994	0.069	1.000

The empirical size and power calculations in Tables 1.9 and 1.10 shows that the proposed test statistics are consistent and valid for multivariate X.

1.7.3 Asymptotic Variance and Bias Derivations

1.7.3.1 Test statistics $\hat{T}_1$

Define

$$\begin{aligned}\hat{\Gamma}(x;\pi_k,\pi_j) :=& \hat{\tau}(x;\pi_k) - \hat{\tau}(x;\pi_j) \\ =& \frac{1}{Nh^d}\sum_{i=1}^{N} Y_i\left[\mathbb{1}(\Pi_i=\pi_k)\hat{\phi}(T_i,x,\pi_k) - \mathbb{1}(\Pi_i=\pi_j)\hat{\phi}(T_i,x,\pi_j)\right] K\left(\frac{x-X_i}{h}\right).\end{aligned}$$

Under the null hypothesis, the bias of $\hat{T}_1$ can be defined as

$$\begin{aligned}\text{Bias}(\hat{T}_1) :=& \mathbb{E}[\hat{T}_1] \\ =& \mathbb{E}\left[\int_{\mathcal{X}}\sum_{k=1}^{K}\sum_{j=1}^{K}\left\{\sqrt{N}|\hat{\Gamma}(x;\pi_k,\pi_j)|\right\}\frac{w(x,\pi_k,\pi_j)}{2}dx\right] \\ =& \int_{\mathcal{X}}\sum_{k=1}^{K}\sum_{j=1}^{K}\mathbb{E}\left[\sqrt{N}|\hat{\Gamma}(x;\pi_k,\pi_j)|\right]\frac{w(x,\pi_k,\pi_j)}{2}dx \\ =& \int_{\mathcal{X}}\sum_{k=1}^{K}\sum_{j=1}^{K}\left\{\mathbb{E}\left|\sqrt{N}(\hat{\Gamma}(x;\pi_k,\pi_j))\frac{\sqrt{N\text{Var}(\hat{\Gamma}(x;\pi_k,\pi_j))}}{\sqrt{N\text{Var}(\hat{\Gamma}(x;\pi_k,\pi_j))}}\right|\right\}\frac{w(x,\pi_k,\pi_j)}{2}dx \\ =& \int_{\mathcal{X}}\sum_{k=1}^{K}\sum_{j=1}^{K}\left\{\sqrt{N\text{Var}(\hat{\Gamma}(x;\pi_k,\pi_j))}\mathbb{E}\left|\frac{\sqrt{N}(\hat{\Gamma}(x;\pi_k,\pi_j))}{\sqrt{N\text{Var}(\hat{\Gamma}(x;\pi_k,\pi_j))}}\right|\right\}\frac{w(x,\pi_k,\pi_j)}{2}dx \\ \to& \int_{\mathcal{X}}\sum_{k=1}^{K}\sum_{j=1}^{K}\left\{\sqrt{N\text{Var}(\hat{\Gamma}(x;\pi_k,\pi_j))}\mathbb{E}\,|\mathbb{Z}_1|\right\}\frac{w(x,\pi_k,\pi_j)}{2}dx \\ =& \frac{\mathbb{E}|\mathbb{Z}_1|}{2}\cdot\int_{\mathcal{X}}\sum_{k=1}^{K}\sum_{j=1}^{K}\sqrt{N\text{Var}(\hat{\Gamma}(x;\pi_k,\pi_j))}w(x,\pi_k,\pi_j)dx \\ \to& h^{\frac{-d}{2}}\cdot\mathbb{E}|\mathbb{Z}_1|\cdot\frac{K^2}{2}\cdot\int_{\mathcal{X}}dx \\ =& a_1.\end{aligned}$$

On the other hand, the variance of $\hat{T}_1$ can be defined as

$$\sigma_1^2 := \frac{N}{4}\int_1 \text{Cov}(|\hat{\Gamma}(x;\pi_i,\pi_j)|,|\hat{\Gamma}(x';\pi_k,\pi_l))|)w(x,\pi_i,\pi_j)w(x',\pi_k,\pi_l)dxdx'$$

$$
\begin{aligned}
&= \frac{1}{4}\int_1 \mathrm{Cov}\left(\sqrt{NVar(\hat{\Gamma}(x;\pi_i,\pi_j))}\left|\frac{\sqrt{N}\cdot(\hat{\Gamma}(x;\pi_i,\pi_j))}{\sqrt{NVar(\hat{\Gamma}(x;\pi_i,\pi_j))}}\right|, \sqrt{NVar(\hat{\Gamma}(x';\pi_k,\pi_l)))}\left|\frac{\sqrt{N}\cdot(\hat{\Gamma}(x';\pi_k,\pi_l))}{\sqrt{NVar(\hat{\Gamma}(x';\pi_k,\pi_l)))}}\right| \right) \\
&\quad \cdot w(x,\pi_i,\pi_j)w(x',\pi_k,\pi_l)dxdx' \\
&= \frac{1}{4}\int_1 \mathrm{Cov}\left(\left|\frac{\sqrt{N}\cdot(\hat{\Gamma}(x;\pi_i,\pi_j))}{\sqrt{NVar(\hat{\Gamma}(x;\pi_i,\pi_j))}}\right|, \left|\frac{\sqrt{N}\cdot(\hat{\Gamma}(x';\pi_k,\pi_l))}{\sqrt{NVar(\hat{\Gamma}(x';\pi_k,\pi_l))}}\right| \right) \sqrt{NVar(\hat{\Gamma}(x;\pi_i,\pi_j))}\sqrt{NVar(\hat{\Gamma}(x';\pi_k,\pi_l))} \\
&\quad \cdot w(x,\pi_i,\pi_j)w(x',\pi_k,\pi_l)dxdx' \\
&\to \frac{h^{-d}}{4}\int_1 \mathrm{Cov}\left(\left|\frac{\sqrt{N}\cdot(\hat{\Gamma}(x;\pi_i,\pi_j))}{\sqrt{NVar(\hat{\Gamma}(x;\pi_i,\pi_j))}}\right|, \left|\frac{\sqrt{N}\cdot(\hat{\Gamma}(x';\pi_k,\pi_l))}{\sqrt{NVar(\hat{\Gamma}(x';\pi_k,\pi_l))}}\right| \right) dxdx' \\
&= \int_{\mathbb{R}^d}\int_{[-1,1]^d} \sum_{i=1}^K\sum_{j=1}^K\sum_{k=1}^K\sum_{l=1}^K \mathrm{Cov}\left(\left| \sqrt{1-\rho(x,t,\pi_i,\pi_j,\pi_k,\pi_l)^2}\mathbb{Z}_1 + \rho(x,t,\pi_i,\pi_j,\pi_k,\pi_l)\mathbb{Z}_2 \right|, |\mathbb{Z}_2| \right) dxdt,
\end{aligned}
$$

where $\int_1 := \int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\sum_{i=1}^K\sum_{j=1}^K\sum_{k=1}^K\sum_{l=1}^K$, and

$$
\begin{aligned}
\rho(x,t,\pi_i,\pi_j,\pi_k,\pi_l) &= \frac{\mathrm{Cov}(\hat{\Gamma}(x;\pi_i,\pi_j),\hat{\Gamma}(x;\pi_k,\pi_l))}{\sqrt{Var(\hat{\Gamma}(x;\pi_i,\pi_j))}\sqrt{Var(\hat{\Gamma}(x;\pi_k,\pi_l))}} \\
&= \begin{cases}
\frac{\int K(\xi)K(\xi+t)d\xi}{\int K(\xi)^2 d\xi} & i=k \,\&\, j=l \\
\frac{-Var(\hat{\tau}(x,\pi_j))}{\sqrt{Var(\hat{\Gamma}(x;\pi_i,\pi_j))}\sqrt{Var(\hat{\Gamma}(x;\pi_k,\pi_l))}}\cdot\frac{\int K(\xi)K(\xi+t)d\xi}{\int K(\xi)^2 d\xi} & j=k \,\&\, i\neq l \\
\frac{-Var(\hat{\tau}(x,\pi_i))}{\sqrt{Var(\hat{\Gamma}(x;\pi_i,\pi_j))}\sqrt{Var(\hat{\Gamma}(x;\pi_k,\pi_l))}}\cdot\frac{\int K(\xi)K(\xi+t)d\xi}{\int K(\xi)^2 d\xi} & j\neq k \,\&\, i=l \\
\frac{Var(\hat{\tau}(x,\pi_j))}{\sqrt{Var(\hat{\Gamma}(x;\pi_i,\pi_j))}\sqrt{Var(\hat{\Gamma}(x;\pi_k,\pi_l))}}\cdot\frac{\int K(\xi)K(\xi+t)d\xi}{\int K(\xi)^2 d\xi} & j=l \,\&\, i\neq k \\
\frac{Var(\hat{\tau}(x,\pi_i))}{\sqrt{Var(\hat{\Gamma}(x;\pi_i,\pi_j))}\sqrt{Var(\hat{\Gamma}(x;\pi_k,\pi_l))}}\cdot\frac{\int K(\xi)K(\xi+t)d\xi}{\int K(\xi)^2 d\xi} & j\neq l \,\&\, i=k \\
0 & i\neq k \,\&\, j\neq l.
\end{cases}
\end{aligned}
$$

$\rho_2(x,\pi) := Var(\hat{\tau}(x,\pi))$ is estimated nonparametrically by $\hat{\rho}_2(x,\pi)$. Plugging in the $\hat{\rho}_2(x,\pi)$'s into the formula for $\rho(x,t,\pi_i,\pi_j,\pi_k,\pi_l)$, we obtain the plug-in estimator $\hat{\rho}(x,t,\pi_i,\pi_j,\pi_k,\pi_l)$. Hence a consistent estimator of the asymptotic variance is

$$
\hat{\sigma}_1^2 := \tfrac{1}{4}\int_{\mathbb{R}^d}\int_{[-1,1]^d}\sum_{i=1}^K\sum_{j=1}^K\sum_{k=1}^K\sum_{l=1}^K \mathrm{Cov}\left(\left| \sqrt{1-\hat{\rho}(x,t,\pi_i,\pi_j,\pi_k,\pi_l)^2}\mathbb{Z}_1 + \hat{\rho}(x,t,\pi_i,\pi_j,\pi_k,\pi_l)\mathbb{Z}_2 \right|, |\mathbb{Z}_2| \right) dxdt.
$$

1.7.3.2 Test statistics $\hat{T}_2$

Define

$$\hat{\Gamma}(x; x', \pi) := \hat{\tau}(x; \pi_k) - \hat{\tau}(x'; \pi_k)$$

Now under the null hypothesis, the bias of $\hat{T}_2$ can be defined as

$$\begin{aligned}
\text{Bias}(\hat{T}_2) :=& \mathbb{E}[\hat{T}_2] \\
=& \mathbb{E}\left[\int_{\mathcal{X}}\int_{\mathcal{X}}\sum_{k=1}^{K}\left\{\sqrt{N}|\hat{\Gamma}(x,x',\pi_k)|\right\}\frac{w(x,x',\pi_k)}{2}dxdx'\right] \\
=& \int_{\mathcal{X}}\int_{\mathcal{X}}\sum_{k=1}^{K}\mathbb{E}\left[\sqrt{N}|\hat{\Gamma}(x,x',\pi_k)|\right]\frac{w(x,x',\pi_k)}{2}dx \\
=& \int_{\mathcal{X}}\int_{\mathcal{X}}\sum_{k=1}^{K}\left\{\mathbb{E}\left|\sqrt{N}(\hat{\Gamma}(x,x',\pi_k))\frac{\sqrt{N\text{Var}(\hat{\Gamma}(x,x',\pi_k))}}{\sqrt{N\text{Var}(\hat{\Gamma}(x,x',\pi_k))}}\right|\right\}\frac{w(x,x',\pi_k)}{2}dxdx' \\
=& \int_{\mathcal{X}}\int_{\mathcal{X}}\sum_{k=1}^{K}\left\{\sqrt{N\text{Var}(\hat{\Gamma}(x,x',\pi_k))}\mathbb{E}\left|\frac{\sqrt{N}(\hat{\Gamma}(x,x',\pi_k))}{\sqrt{N\text{Var}(\hat{\Gamma}(x,x',\pi_k))}}\right|\right\}\frac{w(x,x',\pi_k)}{2}dxdx' \\
\to& \int_{\mathcal{X}}\int_{\mathcal{X}}\sum_{k=1}^{K}\left\{\sqrt{N\text{Var}(\hat{\Gamma}(x,x',\pi_k))}\mathbb{E}\,|\mathbb{Z}_1|\right\}\frac{w(x,x',\pi_k)}{2}dxdx' \\
=& \frac{\mathbb{E}|\mathbb{Z}_1|}{2}\cdot\int_{\mathcal{X}}\int_{\mathcal{X}}\sum_{k=1}^{K}\sqrt{N\text{Var}(\hat{\Gamma}(x,x',\pi_k))}w(x,x',\pi_k)dxdx' \\
\to& h^{\frac{-d}{2}}\cdot\mathbb{E}|\mathbb{Z}_1|\cdot\frac{K}{2}\cdot\int_{\mathcal{X}}\int_{\mathcal{X}}dxdx' = a_2.
\end{aligned}$$

Also, the variance of $\hat{T}_2$ can be defined as

$$\begin{aligned}
\sigma_2^2 :=& \frac{N}{4}\int_2 \text{Cov}(|\hat{\Gamma}(x,x',\pi_k)|, |\hat{\Gamma}(x'',x''',\pi_j))|)w(x,x',\pi_k)w(x'',x''',\pi_j)dxdx'dx''dx''' \\
=& \frac{1}{4}\int_2 \text{Cov}\left(\sqrt{NVar(\hat{\Gamma}(x,x',\pi_k))}\left|\frac{\sqrt{N}\cdot(\hat{\Gamma}(x,x',\pi_k))}{\sqrt{NVar(\hat{\Gamma}(x,x',\pi_k))}}\right|, \sqrt{NVar(\hat{\Gamma}(x'',x''',\pi_j))}\left|\frac{\sqrt{N}\cdot(\hat{\Gamma}(x'',x''',\pi_j))}{\sqrt{NVar(\hat{\Gamma}(x'',x''',\pi_j))}}\right|\right) \\
& \cdot w(x,x',\pi_k)w(x'',x''',\pi_j)dxdx'dx''dx''' \\
=& \frac{1}{4}\int_2 \text{Cov}\left(\left|\frac{\sqrt{N}\cdot(\hat{\Gamma}(x,x',\pi_k))}{\sqrt{NVar(\hat{\Gamma}(x,x',\pi_k))}}\right|, \left|\frac{\sqrt{N}\cdot(\hat{\Gamma}(x'',x''',\pi_j))}{\sqrt{NVar(\hat{\Gamma}(x'',x''',\pi_j))}}\right|\right)\sqrt{NVar(\hat{\Gamma}(x,x',\pi_k))}\sqrt{NVar(\hat{\Gamma}(x'',x''',\pi_j))}
\end{aligned}$$

$$\cdot w(x,x',\pi_k)w(x'',x''',\pi_j)dxdx'dx''dx'''$$

$$\to \frac{h^{-d}}{4}\int_2 \mathrm{Cov}\left(\left|\frac{\sqrt{N}\cdot(\hat{\Gamma}(x,x',\pi_k))}{\sqrt{NVar(\hat{\Gamma}(x,x',\pi_k))}}\right|,\left|\frac{\sqrt{N}\cdot(\hat{\Gamma}(x'',x''',\pi_j))}{\sqrt{NVar(\hat{\Gamma}(x'',x''',\pi_j))}}\right|\right)dxdx'dx''dx'''$$

$$= \frac{h^{2d}}{4}\int_B\int_{T_0}\int_{T_0}\int_{T_0}\sum_{k=1}^{K}\mathrm{Cov}\left(\left|\sqrt{1-\rho_1^*(x,q,r,s,\pi_k)^2}\mathbb{Z}_1+\rho_1^*(x,q,r,s,\pi_k)\mathbb{Z}_2\right|,|\mathbb{Z}_2|\right)drdsdqdx,$$

where $\int_2 := \int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\sum_{k=1}^{K}\sum_{j=1}^{K}$ and

$$\rho_1^*(x,q,r,s,\pi_k) := \frac{\rho_1(x,r,\pi_k)-\rho_1(x,s,\pi_k)-\rho_1(x,q,r,\pi_k)+\rho_1(x,q,s,\pi_k)}{2\sqrt{(\rho_2(x,\pi_k)-\rho_1(x,q,\pi_k))(\rho_2(x,\pi_k)-\rho_1(x,r,s,\pi_k))}}$$

which reduces to

$$\rho_1^*(x,q,r,s,\pi_k) := \frac{\int K(\xi+r)K(\xi)-K(\xi+s)K(\xi)-K(\xi+q)K(\xi+r)+K(\xi+q)K(\xi+s)d\xi}{2\cdot\sqrt{(\int K^2(\xi)-K(\xi+q)K(\xi)d\xi)(\int K^2(\xi)-K(\xi+q)K(\xi+s)d\xi)}}.$$

since

$$\rho_1(x,r,s,\pi) := \left\{\sum_{t\in\{0,1\}}\frac{\mathbb{E}[Y^2|X=x,T=t,\Pi=\pi]-(\mathbb{E}[Y|X=x,T=t,\Pi=\pi])^2}{P_t(x,\pi)}\right\}\cdot\int K(\xi+s)K(\xi+r)d\xi$$

Therefore,

$$\sigma_2^2 = \frac{h^{2d}}{4}\cdot K\cdot V$$

where V is a constant which depends on the kernel function and the support of X.

1.7.4 Proofs of Lemmas and Theorems

In this section, I provide a detailed proof of the asymptotic normality proof of $\hat{S}_1$. The asymptotic normality proof of $\hat{S}_2$ is similar therefore I layout a sketch proof at the end of this section

1.7.4.1 Proof of the asymptotic normality $\hat{S}_1$

Uniform asymptotic approximation of $\hat{T}_1$

Write

$$\hat{\tau}(x,\pi) = \tau(x,\pi)+(\tau_{N0}(x,\pi)-\mathbb{E}(\tau_{N0}(x,\pi)))+(\mathbb{E}(\tau_{N0}(x,\pi))-\tau(x,\pi))+R_N(x,\pi),$$

where

$$\tau_{N0}(x,\pi) := \frac{1}{Nh^d}\sum_{i=1}^{N} Y_i \mathbb{1}(\Pi_i = \pi)\left[\frac{T_i}{P_1(x;\pi)} - \frac{(1-T_i)}{P_0(x;\pi)}\right]\cdot K\left(\frac{x-X_i}{h}\right),$$

$$\begin{aligned} R_N(x,\pi) :=& \frac{1}{Nh^d}\sum_{i=1}^{N} Y_i \mathbb{1}(\Pi_i = \pi)\left[\frac{T_i}{P_1(x;\pi)} - \frac{(1-T_i)}{P_0(x;\pi)}\right] \\ &\times\left(T_i\frac{P_1(x,\pi)-\hat{P}_1(x,\pi)}{\hat{P}_1(x,\pi)} + (1-T_i)\frac{P_0(x,\pi)-\hat{P}_0(x,\pi)}{\hat{P}_0(x,\pi)}\right)\cdot K\left(\frac{x-X_i}{h}\right). \end{aligned}$$

Therefore,

$$\begin{aligned} \hat{\Gamma}(x;\pi_k,\pi_j) =& \tau(x,\pi_k) - \tau(x,\pi_j) + (\tau_{N0}(x,\pi_k) - \tau_{N0}(x,\pi_j) - \mathbb{E}(\tau_{N0}(x,\pi_k) - \tau_{N0}(x,\pi_j))) \\ &+ (\mathbb{E}(\tau_{N0}(x,\pi_k) - \tau_{N0}(x,\pi_j)) - \tau(x,\pi_k) - \tau(x,\pi_j))) + R_N(x,\pi_k) - R_N(x,\pi_j). \end{aligned}$$

Now, define

$$\begin{aligned} \zeta_{_N}(x,\pi) =& \mathbb{E}[Y|X=x,\Pi=\pi,T=1] - \mathbb{E}[Y|X=x,\Pi=\pi,T=0] \\ &- \mathbb{E}[Y|X=x,\Pi=\pi,T=1]\frac{1}{Nh^dP_1(x,\pi)}\sum_{i=1}^{N} T_i\mathbb{1}(\Pi=\pi)K\left(\frac{x-X_i}{h}\right) \\ &+ \mathbb{E}[Y|X=x,\Pi=\pi,T=0]\frac{1}{Nh^dP_0(x,\pi)}\sum_{i=1}^{N}(1-T_i)\mathbb{1}(\Pi=\pi)K\left(\frac{x-X_i}{h}\right). \end{aligned}$$

The following lemma shows that $R_N(x,\pi)$ can be approximated by $\zeta_{_N}(x,\pi)$ uniformly over x at a rate faster than $N^{-1/2}$.

Lemma 1.7.1 *Under the regularity conditions, we find that for $\pi_k, \pi_j \in \mathbf{\Pi}$,*

$$\sup_{x\in\mathcal{X}} |(R_N(x,\pi_k) - R_N(x,\pi_j)) - (\zeta_{_N}(x,\pi_k) - \zeta_{_N}(x,\pi_j))| = o_p(N^{-1/2}).$$

From, Lemma B.1 in Chang, Lee, and Whang (2015), for $\pi_k \in \Pi$, we have

$$\sup_{x\in\mathcal{X}} |R_N(x,\pi_k) - \zeta_{_N}(x,\pi_k))| = o_p(N^{-1/2}).$$

Now,

$$\begin{aligned}\sup_{x\in\mathcal{X}}|(R_N(x,\pi_k)-R_N(x,\pi_j))-(\zeta_N(x,\pi_k)-\zeta_N(x,\pi_j))| &\leq \sup_{x\in\mathcal{X}}\{|R_N(x,\pi_k)-\zeta_N(x,\pi_k)|+|R_N(x,\pi_j)-\zeta_N(x,\pi_j)|\}\\ &= \sup_{x\in\mathcal{X}}\{|R_N(x,\pi_k)-\zeta_N(x,\pi_k)|\}+\sup_{x\in\mathcal{X}}\{|R_N(x,\pi_j)-\zeta_N(x,\pi_j)|\}\\ &= o_p(N^{-1/2})+o_p(N^{-1/2})=o_p(N^{-1/2}).\end{aligned}$$

Lemma 1.7.2 *Under the regularity conditions, we have*

$$\hat{T}_1 - T^*_{1N} = o_p(1),$$

where

$$T^*_{1N} := \int_{\mathcal{X}}\sum_{k=1}^{K}\sum_{j=1}^{K}\left\{\sqrt{N}\left|\Gamma(x,\pi_k,\pi_j)+[\tau_N(x,\pi_k)-\tau_N(x,\pi_j)]-\mathbb{E}[\tau_N(x,\pi_k)-\tau_N(x,\pi_j)]\right|\right\}\frac{w(x,\pi_j,\pi_k)}{2}dx$$

and

$$\tau_N(x,\pi) = \tau_{N0}(x,\pi)+\zeta_N(x,\pi).$$

Hence, under the null hypothesis such that $\tau(x,\pi_k)=\tau(x,\pi_j)$ on $\mathcal{X}\times\Pi$, we have

$$\hat{T}_1 = T_{1N} + o_p(1),$$

where

$$T_{1N} := \int_{\mathcal{X}}\sum_{k=1}^{K}\sum_{j=1}^{K}\left\{\sqrt{N}\left|[\tau_N(x,\pi_k)-\tau_N(x,\pi_j)]-\mathbb{E}[\tau_N(x,\pi_k)-\tau_N(x,\pi_j)]\right|\right\}\frac{w(x,\pi_j,\pi_k)}{2}dx.^{8}$$

and

$$\begin{aligned}\tau_N(x,\pi) :=& \frac{1}{Nh^d}\sum_{i=1}^{N}(\{Y-\mathbb{E}[Y|X=x,\Pi=\pi,T=1]\}\frac{T_i\cdot\mathbb{1}(\Pi=\pi)}{P_1(x,\pi_k)}\\ &-\{Y-\mathbb{E}[Y|X=x,\Pi=\pi,T=0]\}\frac{(1-T_i)\cdot\mathbb{1}(\Pi=\pi)}{P_0(x,\pi_k)})\cdot K\left(\frac{x-X_i}{h}\right).\end{aligned}$$

Using the triangle inequality and the proof of Lemma B.2 in Chang, Lee, and Whang (2015), the proof of this Lemma is straightforward.

[8] Here, I use the true weight rather than the estimated weight but all results hold using the estimated weight since the estimator is uniformly consistent by assumption.

Consistency of the estimators of asymptotic variance

Lemma 1.7.3 *Under the regularity conditions, the following hold:*

1. $\sup_{x\in\mathcal{X}} |\hat{\tau}(x,\pi) - \tau(x,\pi)| = O_p((Nh^d)^{-1/2}\log N + h^s)\ \forall\pi \in \Pi$,
2. $\sup_{x\in\mathcal{X}} |\hat{\rho}_2(x,\pi) - \rho_2(x,\pi)| = O_p((Nh^d)^{-1/2}\log N + h^s)\ \forall\pi \in \Pi$.

This Lemma corresponds to Lemma B.3 in Chang, Lee, and Whang (2015). Hence I omit the proof.

Theorem 1.7.1 *Under the regularity conditions, we have*

$$\hat{\sigma}_1^2 = \sigma_1^2 + o_p(1).$$

Since $\mathrm{Cov}\left(\left|\sqrt{1-\rho(x,t,\pi_i,\pi_j,\pi_k,\pi_l)^2}\mathbb{Z}_1 + \rho(x,t,\pi_i,\pi_j,\pi_k,\pi_l)\mathbb{Z}_2\right|, |\mathbb{Z}_2|\right)$ is a continuous functional of ρ on $[-1,1]^d$ and $\hat{\rho}^2(x,t,\pi_i,\pi_j,\pi_k,\pi_l)$ is uniformly consistent for $\rho^2(x,t,\pi_i,\pi_j,\pi_k,\pi_l)$ using Lemma 1.7.3(2), Assumption 1.2.4 and Assumption 1.3.1(e). It is straightforward to show this result.

Theorem 1.7.2 *Under the regularity conditions, we have that*

$$\frac{T_{1N} - a_1}{\sigma_1} \xrightarrow{d} N(0,1).$$

The following fact which is Fact 6.1 in Giné, Mason, and Zaitsev (2003) and follows from Theorem 1 of Sweeting (1977) will be necessary.

Fact 1.7.1 *Let $\{(W_i, V_i)' : i = 1,\ldots,N\}$ be a sequence of iid random vectors in $\mathbb{R}^2$ such that each component has mean zero and variance one and finite absolute moments of the third order. Also, let $(Z_1, Z_2)'$ be a bivariate normal with $\mathbb{E}[Z_1] = \mathbb{E}[Z_2] = 0$, $Var(Z_1) = Var(Z_2) = 1$ and $\mathrm{Cov}(Z_1, Z_2) = \mathrm{Cov}(W_i, V_i) = \rho$. Then there exist universal positive constants A_1, A_2 and A_3 such that*

$$\left|\mathbb{E}\left|\frac{\sum_{i=1}^N W_i}{\sqrt{N}}\right| - \mathbb{E}|Z_1|\right| \leq \frac{A_1}{\sqrt{N}}\mathbb{E}|W_i|^3 \tag{1.25}$$

and, whenever $\rho^2 < 1$

$$\left|\mathbb{E}\left|\frac{\sum_{i=1}^N W_i}{\sqrt{N}}\cdot\frac{\sum_{i=1}^N V_i}{\sqrt{N}}\right| - \mathbb{E}|Z_1 Z_2|\right| \leq \frac{A_2}{(1-\rho^2)^{3/2}\sqrt{N}}(\mathbb{E}|W_i|^3 + \mathbb{E}|V_i|^3) \tag{1.26}$$

and

$$\left|\mathbb{E}\left[\frac{\sum_{i=1}^{N} W_i}{\sqrt{N}} \cdot \left|\frac{\sum_{i=1}^{N} V_i}{\sqrt{N}}\right|\right] - \mathbb{E}[Z_1|Z_2|]\right| \leq \frac{A_3}{(1-\rho^2)^{3/2}\sqrt{N}}(\mathbb{E}|W_i|^3 + \mathbb{E}|V_i|^3) \tag{1.27}$$

The following lemma is similar to lemma 6.1 in Giné, Mason, and Zaitsev (2003). To save space, I omit the proof and refer readers to the proof of lemma 6.1 in Giné, Mason, and Zaitsev (2003).

Lemma 1.7.4 *Suppose $\mathcal{H}$ is a finite class of uniformly bounded real-valued functions H, which are equal to zero outside a known compact set. Further, let $g(\pi, x)f(x)$ be continuously differentiable in x with $\sup_{(\pi,x)\in\Pi\times B}\left|\frac{d(g(\pi,x)f(x))}{dx}\right| < \infty$ where $B \subset \mathbb{R}^d$ is a compact set. Then uniformly in $H \in \mathcal{H}$*

$$\sup_{(\pi,x)\in\Pi\times B}\left|\frac{1}{h^d}\int_{\mathbb{R}^d} g(\pi,z)f(z)H\left(\frac{x-z}{h}\right)dz - g(\pi,x)f(x)\int_{\mathbb{R}^d} H\left(\frac{x-z}{h}\right)dz\right| \to 0 \text{ as } h \to 0. \tag{1.28}$$

Now I use the "Poissonization" technique used in Giné, Mason, and Zaitsev (2003), Lee et al. (2013) and Chang, Lee, and Whang (2015). Let $\mathcal{N}$ denote a Poisson random variable with mean N defined on the same probability space as the sequence $\{W_i : i \geq 1\} := \{(Y_i, X_i, T_i, \Pi_i) : i \geq 1\}$ and independent of this sequence. Define

$$\chi_{t,k} := \frac{\mathbb{E}[Y|X = x, \Pi = \pi_k, T = t]}{P_t(x, \pi_k)},$$

$$\chi(\pi_k, x, T) := [\chi_{1,k}(\pi_k, x) \cdot T - \chi_{0,k}(\pi_k, x) \cdot (1 - T)] \cdot \mathbb{1}(\Pi = \pi_k),$$

and

$$\psi(W_i, x, \pi_k) := [Y_i \cdot \mathbb{1}(\Pi_i = \pi_k)\phi(x, \pi_k, T_i) - \chi(\pi_k, x, T_i)]\frac{1}{h^d}K\left(\frac{x - X_i}{h}\right) + \tau(x; \pi_k).$$

Then

$$\tau_{\mathcal{N}}(x, \pi_k) = \tau_{\mathcal{N}0}(x, \pi_k) + \zeta_{\mathcal{N}}(x, \pi_k) = \frac{1}{N}\sum_{i=1}^{N}\psi(W_i, x, \pi_k).$$

Hence define,

$$\begin{aligned}\Gamma_N(x, \pi_k, \pi_j) :=& \tau_{\mathcal{N}}(x, \pi_k) - \tau_{\mathcal{N}}(x, \pi_j) \\ =& \frac{1}{N}\sum_{i=1}^{N}\psi(W_i, x, \pi_k) - \frac{1}{N}\sum_{i=1}^{N}\psi(W_i, x, \pi_j) \\ =& \frac{1}{N}\sum_{i=1}^{N}\Theta(W_i, x, \pi_k, \pi_j),\end{aligned}$$

where

$$\Theta(W_i, x, \pi_k, \pi_j) := \Big\{ [Y_i \cdot [\mathbb{1}(\Pi_i = \pi_k)\phi(x, T_i, \pi_k) - \mathbb{1}(\Pi_i = \pi_j)\phi(x, T_i, \pi_j)] - [\chi(\pi_k, x, T_i) - \chi(\pi_j, x, T_i)] \Big\} \cdot \frac{1}{h^d} K\left(\frac{x - X_i}{h}\right) + \Gamma(x, \pi_k, \pi_j).$$

Now we will poissonize $\Gamma_N(x, \pi_k, \pi_j)$. To do so, again define

$$\Gamma_{\mathcal{N}}(x, \pi_k, \pi_j) = \frac{1}{N} \sum_{i=1}^{\mathcal{N}} \Theta(W_i, x, \pi_k, \pi_j)$$

where the sum is zero if N=0. Note that by the law of iterated expectation and variance,

$$\mathbb{E}\Gamma_{\mathcal{N}}(x, \pi_k, \pi_j) = \mathbb{E}\Gamma_N(x, \pi_k, \pi_j) = \mathbb{E}[\psi(W, x, \pi_k)] - \mathbb{E}[\psi(W, x, \pi_j)], \tag{1.29}$$

$$\kappa_{\tau,N}(x, \pi_k, \pi_j) := NVar(\Gamma_{\mathcal{N}}(x, \pi_k, \pi_j)) = \mathbb{E}[\Theta^2(W, x, \pi_k, \pi_j)], \tag{1.30}$$

$$\kappa_{\tau,N}(x, \pi) := NVar(\tau_{\mathcal{N}}(x, \pi)) \tag{1.31}$$

and

$$NVar(\Gamma_N(x, \pi_k, \pi_j)) = \mathbb{E}[\Theta^2(W, x, \pi_k, \pi_j)] - \{\mathbb{E}[\Theta(W, x, \pi_k, \pi_j)]\}^2. \tag{1.32}$$

Let $\epsilon \in \left(0, \int_{\mathcal{X}} f(x)dx\right)$ be an arbitrary constant. For constant $\{M_l > 0 : l = 1, \ldots, d\}$, let $\mathcal{B}(M) = \Pi_{l=1}^d [-M_l, M_l] \subset \mathcal{X}$ denote a Borel set in $\mathbb{R}^d$ with nonempty interior with finite Lebesgue measure $\lambda(\mathcal{B}(M))$. For $v > 0$, define $\mathcal{B}(M, v)$ to be the v-contraction of $\mathcal{B}(M)$, i.e., $\mathcal{B}(M, v) = \{x \in \mathcal{B}(M) : \inf_{y \in \mathbb{R}^d \setminus B(M)} \{\|x - y\|\} \geq v\}$, Choose $M, v > 0$ and a Borel set B such that

$$B \subset \mathcal{B}(M, v), \tag{1.33}$$

$$\int_{\mathbb{R}^d \setminus \mathcal{B}(M)} f(x)dx := \alpha > 0, \tag{1.34}$$

$$\int_{\mathcal{X}} f(x)dx - \int_{B} f(x)dx > \epsilon. \tag{1.35}$$

According to Chang, Lee, and Whang (2015), such M, v and B exist by the absolute continuity of the density f. Lets define a poissonized version of T_{1N} (restricted to B)- the uniform asymptotic approximation of $\hat{T}_1$ under the null hypothesis- to be:

$$\begin{aligned} T_{1N}^{P}(B) := &\int_{B} \sum_{k=1}^{K} \sum_{j=1}^{K} \left\{ \sqrt{N} \left| [(\Gamma_{\mathcal{N}}(x, \pi_k, \pi_j)] - \mathbb{E}[(\Gamma_{\mathcal{N}}(x, \pi_k, \pi_j)] \right| \right\} \frac{w(x, \pi_j, \pi_k)}{2} dx \\ &- \int_{B} \sum_{k=1}^{K} \sum_{j=1}^{K} \left\{ \sqrt{N} \mathbb{E} \left| [(\Gamma_{\mathcal{N}}(x, \pi_k, \pi_j)] - \mathbb{E}[(\Gamma_{\mathcal{N}}(x, \pi_k, \pi_j)] \right| \right\} \frac{w(x, \pi_j, \pi_k)}{2} dx. \end{aligned}$$

Let

$$\sigma_{1N}^{2}(B) = \mathrm{Var}(T_{1N}^{P}(B)).$$

The following lemma shows that the variance of the approximated and poissonized version of $\hat{T}_1$ under the null in B converges to the variance of the "un-approximated" and "un-poissonized" version of $\hat{T}_1$ under the null in B

Lemma 1.7.5 *If the regularity conditions holds and B satisfies (1.33)-(1.35), then*

$$\lim_{N \to \infty} \sigma_{1N}^{2}(B) = \mathrm{P}(B) \cdot \mathrm{Var}(\hat{T}_1) = \sigma_{1,B}^{2}, \tag{1.36}$$

where

$$\begin{aligned} \sigma_{1,B}^{2} := &\int_{-1}^{1} \int_{B} \sum_{i=1}^{K} \sum_{j=1}^{K} \sum_{k=1}^{K} \sum_{l=1}^{K} \mathrm{Cov}(|\sqrt{1 - \rho^2(x, \pi_i, \pi_j, \pi_k, \pi_l, r)}\mathbb{Z}_1 + \rho(x, \pi_i, \pi_j, \pi_k, \pi_l, r)\mathbb{Z}_2|, |\mathbb{Z}_2|) \\ &\cdot \sqrt{\rho_2(x, \pi_i, \pi_j)\rho_2(x, \pi_k, \pi_l)} \cdot w(x, \pi_i, \pi_j) w(x', \pi_k, \pi_l) dx dr. \end{aligned} \tag{1.37}$$

proof 1.7.1 *Note that for each* $(x, \pi_i, \pi_j), (x', \pi_k, \pi_l) \in \mathbb{R}^d \times \Pi^2$ *such that* $||x - x'|| > h$, *the random variables* $\Gamma_{\mathcal{N}}(x, \pi_i, \pi_j) - \mathbb{E}[\Gamma_{\mathcal{N}}(x, \pi_i, \pi_j)]$ *and* $\Gamma_{\mathcal{N}}(x', \pi_k, \pi_l) - \mathbb{E}[\Gamma_{\mathcal{N}}(x', \pi_k, \pi_l)]$ *are independent because they are functions of independent increments of a Poisson process and the kernel K vanishes outside of the closed rectangle* $[-1, 1]^d$. *Therefore,*

$$\begin{aligned} \mathrm{Var}(T_{1N}^{P}(B)) = &\frac{1}{4} \int_{1} \mathrm{Cov}(\sqrt{N}|\Gamma_{\mathcal{N}}(x, \pi_i, \pi_j) - \mathbb{E}[\Gamma_{\mathcal{N}}(x, \pi_i, \pi_j)]|, \sqrt{N}|\Gamma_{\mathcal{N}}(x', \pi_k, \pi_l) - \mathbb{E}[\Gamma_{\mathcal{N}}(x', \pi_k, \pi_l)]|) \\ &\cdot w(x, \pi_i, \pi_j) w(x', \pi_k, \pi_l) dx dx' \end{aligned}$$

$$=\frac{1}{4}\int_1 \text{Cov}(\sqrt{N}|\Gamma_{\mathcal{N}}(x,\pi_i,\pi_j)-\mathbb{E}[\Gamma_{\mathcal{N}}(x,\pi_i,\pi_j)]|,\ \sqrt{N}|\Gamma_{\mathcal{N}}(x',\pi_k,\pi_l)-\mathbb{E}[\Gamma_{\mathcal{N}}(x',\pi_k,\pi_l)]|)$$
$$\cdot\ \mathbb{1}(h^{-1}(x-x')\in[-1,1]^d)\cdot w(x,\pi_i,\pi_j)w(x',\pi_k,\pi_l)dxdx',$$

where $\int_1 := \int_B\int_B\sum_{\pi_i\in\Pi}\sum_{\pi_j\in\Pi}\sum_{\pi_k\in\Pi}\sum_{\pi_l\in\Pi}\cdot$

Again, let

$$\begin{aligned}
S_{\tau,\mathcal{N}}(x,\pi,\pi') =& \frac{\sqrt{N}\{\Gamma_{\mathcal{N}}(x,\pi,\pi')-\mathbb{E}[\tau_{\mathcal{N}}(x,\pi,\pi')]\}}{\sqrt{\kappa_{\Gamma,N}(x,\pi,\pi')}}\\
=& \frac{\sqrt{N}\{\tau_{\mathcal{N}}(x,\pi)-\mathbb{E}[\tau_{\mathcal{N}}(x,\pi)]\}}{\sqrt{\kappa_{\Gamma,N}(x,\pi,\pi')}}+\frac{\sqrt{N}\{\tau_{\mathcal{N}}(x,\pi')-\mathbb{E}[\tau_{\mathcal{N}}(x,\pi')]\}}{\sqrt{\kappa_{\Gamma,N}(x,\pi,\pi')}}\\
\leq& \frac{\sqrt{N}\{\tau_{\mathcal{N}}(x,\pi)-\mathbb{E}[\tau_{\mathcal{N}}(x,\pi)]\}}{\sqrt{\kappa_{\tau,N}(x,\pi)}}+\frac{\sqrt{N}\{\tau_{\mathcal{N}}(x,\pi')-\mathbb{E}[\tau_{\mathcal{N}}(x,\pi')]\}}{\sqrt{\kappa_{\tau,N}(x,\pi')}}\\
:=& S_{\tau,\mathcal{N}}(x,\pi)+S_{\tau,\mathcal{N}}(x,\pi'). \qquad (1.38)
\end{aligned}$$

Now, with $K<\infty$ *and* $\int_B dx<\infty$, *by the Lemma 1.7.4, we can show that*

$$\sup_{(x,\pi,\pi')\in B\times\Pi^2}\left|\sqrt{\kappa_{\Gamma,N}(x,\pi,\pi')}-h^{-d/2}\sqrt{\rho_2(x,\pi,\pi')}\right|=O(h^{d/2}), \qquad (1.39)$$

where

$$\rho_2(x,\pi,\pi')=\left\{\sum_{t\in\{0,1\}}\sum_{\pi_0\in\{\pi,\pi'\}}\frac{\mathbb{E}[Y^2|X=x,T=t,\Pi=\pi_0]-(\mathbb{E}[Y|X=x,T=t,\Pi=\pi_0])^2}{P_t(x,\pi_0)}\right\}\cdot\int K^2(\xi)d\xi$$

We also have that

$$\int_B\int_B\sum_{\pi_i\in\Pi}\sum_{\pi_j\in\Pi}\sum_{\pi_k\in\Pi}\sum_{\pi_l\in\Pi}\mathbb{1}(h^{-1}(x-x')\in[-1,1]^d)\cdot w(x,\pi_i,\pi_j)w(x',\pi_k,\pi_l)dxdx'=O(h^d). \qquad (1.40)$$

And finally, by the Cauchy Schwartz Inequality, we have

$$\sup_{\{(x,\pi_i,\pi_j),(x',\pi_k,\pi_l)\}\in(B\times\Pi^2)^2}\left|\text{Cov}(|S_{\tau,\mathcal{N}}(x,\pi_i,\pi_j)|,|S_{\tau,\mathcal{N}}(x',\pi_k,\pi_l)|)\right|=O(1). \qquad (1.41)$$

Therefore, we have

$$\text{Var}(T_{1N}^P(B))=\bar{\sigma}_{1N}^2+o(1),$$

where

$$\bar{\sigma}^2_{1N} := \frac{1}{4}\int_1 \mathrm{Cov}(|S_{\tau,\mathcal{N}}(x,\pi_i,\pi_j)|, |S_{\tau,\mathcal{N}}(x',\pi_k,\pi_l)|)\cdot \sqrt{\rho_2(x,\pi_i,\pi_j)\rho_2(x',\pi_k,\pi_l)} \times h^{-d}\cdot \mathbb{1}(h^{-1}(x-x')\in[-1,1]^d)\cdot w(x,\pi_i,\pi_j)w(x',\pi_k,\pi_l)dxdx'$$

Now, for $b\in\{1,2\}$, let $Z_{bn}(x,\pi_i,\pi_j) = Z_{bn}(x,\pi_i)+Z_{bn}(x,\pi_j)$ where ($Z_{1n}(x,\pi_i)$, $Z_{2n}(x,\pi_i)$, $Z_{1n}(x,\pi_j)$, $Z_{2n}(x,\pi_j)$) are mean zero pairwise independent Gaussian processes.
Then $(Z_{1n}(x,\pi_i,\pi_j), Z_{2n}(x',\pi_k,\pi_l))$ for $(x,\pi_i,\pi_j),(x',\pi_k,\pi_l)\in\mathbb{R}^d\times\Pi^2$, is a mean zero bivariate Gaussian process. Now for each $(x,\pi_i,\pi_j)\in\mathbb{R}^d\times\Pi^2, (x',\pi_k,\pi_l)\in\mathbb{R}^d\times\Pi^2$, let $(Z_{1n}(x,\pi_i,\pi_j), Z_{2n}(x',\pi_k,\pi_l))$ and $(S_{\tau,\mathcal{N}}(x,\pi_i,\pi_j), S_{\tau,\mathcal{N}}(x',\pi_k,\pi_l))$ have the same covariance structure. That is,

$$(Z_{1n}(x,\pi_i,\pi_j), Z_{2n}(x,\pi_k,\pi_l)) \overset{d}{=} \left(\sqrt{1-\rho^*_N(x,\pi_i,\pi_j,\pi_k,\pi_l)^2}\mathbb{Z}_1 + \rho^*_N(x',\pi_i,\pi_j,\pi_k,\pi_l)\mathbb{Z}_2, \mathbb{Z}_2\right),$$

where $\mathbb{Z}_1$ and $\mathbb{Z}_2$ are independent standard normal random variables and

$$\rho^*_N(x,x',\pi_i,\pi_j,\pi_k,\pi_l) := \mathbb{E}[S_{\tau,\mathcal{N}}(x,\pi_i,\pi_j)S_{\tau,\mathcal{N}}(x',\pi_k,\pi_l)].$$

Let

$$\bar{\tau}^2_{N,0} = \int_{T_0}\int_B \sum_{\pi_i\in\Pi}\sum_{\pi_j\in\Pi}\sum_{\pi_k\in\Pi}\sum_{\pi_l\in\Pi}\mathrm{Cov}(|Z_{1n}(x,\pi_i,\pi_j)|, |Z_{2n}(x',\pi_k,\pi_l)|)\cdot\sqrt{\rho_2(x,\pi_i,\pi_j)\rho_2(x',\pi_k,\pi_l)} \times h^{-d}\cdot\mathbb{1}(h^{-1}(x-x')\in[-1,1]^d)\cdot w(x,\pi_i,\pi_j)w(x',\pi_k,\pi_l)dxdx'.$$

By a change of variables $x' = x+th$, we can write

$$\bar{\tau}^2_{N,0} = \int_B\int_B \sum_{\pi_i\in\Pi}\sum_{\pi_j\in\Pi}\sum_{\pi_k\in\Pi}\sum_{\pi_l\in\Pi}\mathrm{Cov}(|Z_{1n}(x,\pi_i,\pi_j)|, |Z_{2n}(x+th,\pi_k,\pi_l)|) \cdot\sqrt{\rho_2(x,\pi_i,\pi_j)\rho_2(x+th,\pi_k,\pi_l)}\times\mathbb{1}(x\in B)\mathbb{1}(x+th\in B) \cdot w(x,\pi_i,\pi_j)w(x+th,\pi_k,\pi_l)dxdt.$$

Note that

$$\rho^*_N(x,x',\pi_i,\pi_j,\pi_k,\pi_l) = N\mathbb{E}\left[\frac{\{\Gamma_{\mathcal{N}}(x,\pi_i,\pi_j)-\mathbb{E}[\Gamma_{\mathcal{N}}(x,\pi_i,\pi_j)]\}}{\sqrt{\kappa_{\Gamma,N}(x,\pi_i,\pi_j)}}\frac{\{\Gamma_{\mathcal{N}}(x',\pi_k,\pi_l)-\mathbb{E}[\Gamma_{\mathcal{N}}(x',\pi_k,\pi_l)]\}}{\sqrt{\kappa_{\Gamma,N}(x',\pi_k,\pi_l)}}\right]$$

$$= \frac{\mathbb{E}\left[Q(x,\pi_i,\pi_j)Q(x',\pi_i,\pi_j)Q(x,\pi_k,\pi_l)Q(x',\pi_k,\pi_l)\cdot \frac{1}{h^d}K\left(\frac{x-X}{h}\right)\frac{1}{h^d}K\left(\frac{x'-X}{h}\right)\right]}{\sqrt{\kappa_{\Gamma,N}(x,\pi_i,\pi_j)\kappa_{\Gamma,N}(x',\pi_k,\pi_l)}},$$

where

$$Q(x,\pi,\pi') := \left\{[Y\cdot[\mathbb{1}(\Pi=\pi)\hat{\phi}(x,T,\pi) - \mathbb{1}(\Pi=\pi')\hat{\phi}(x,T,\pi')] - [\chi(\pi,x,T)-\chi(\pi',x,T)]\right\}.$$

By Lemma 1.7.4 and a change of variable $x' = x + th$, we can show that

$$\rho_N^*(x,x',\pi_i,\pi_j,\pi_k,\pi_l) \to \frac{\int K(\xi)K(\xi+t)d\xi}{\int K(\xi)d\xi}.$$

Therefore, as in the proof of Lemma B.10 in Chang, Lee, and Whang (2015), by the bounded convergence theorem, we have that

$$\lim_{N\to\infty}\bar{\tau}_{N,0}^2 = \sigma_{1,B}^2.$$

Now, the desired result holds if

$$\bar{\tau}_{N,0}^2 \to \bar{\sigma}_{N,0}^2 \text{ as } N\to\infty. \tag{1.42}$$

To prove (1.42), let $\epsilon_0 \in (0,1]$ and let $c(\epsilon_0) = (1+\epsilon_0)^2 - 1$. Let Q_1 and Q_2 be two independent random variables that are independent of $(\{Y_i,X_i\}_{i=1}^\infty, \mathcal{N})$, each having a two-point distribution that gives two points, $\{\sqrt{c(\epsilon_0)}\}$ and $\{-\sqrt{c(\epsilon_0)}\}$, the equal mass of 1/2, so that $\mathbb{E}[Q_1] = \mathbb{E}[Q_2] = 0$ and $\mathrm{Var}(Q_1) = \mathrm{Var}(Q_2) = c(\epsilon_0)$. Let $S^Q_{\tau,\mathcal{N},1}(x,\pi,\pi') = \frac{S_{\tau,\mathcal{N}}(x,\pi,\pi')+2Q_1}{1+\epsilon_0} = \frac{S_{\tau,\mathcal{N}}(x,\pi)+Q_1+S_{\tau,\mathcal{N}}(x,\pi')+Q_1}{1+\epsilon_0} =: S^Q_{\tau,\mathcal{N},1}(x,\pi)+S^Q_{\tau,\mathcal{N},1}(x,\pi')$ and $S^Q_{\tau,\mathcal{N},2}(x,\pi,\pi') = \frac{S_{\tau,\mathcal{N}}(x,\pi\pi')+2Q_2}{1+\epsilon_0} = \frac{S_{\tau,\mathcal{N}}(x,\pi)+Q_2+S_{\tau,\mathcal{N}}(x,\pi')+Q_2}{1+\epsilon_0} =: S^Q_{\tau,\mathcal{N},2}(x,\pi)+S^Q_{\tau,\mathcal{N},2}(x,\pi')$. Define

$$\begin{aligned}\bar{\sigma}_{1N,Q}^2 := \int_1 &\mathrm{Cov}(|S^Q_{\tau,\mathcal{N},1}(x,\pi_i,\pi_j)|,|S^Q_{\tau,\mathcal{N},2}(x,\pi_k,\pi_l)|)\cdot\sqrt{\rho_2(x,\pi_i,\pi_j)\rho_2(x',\pi_k,\pi_l)}\\ &\times h^{-d}\cdot\mathbb{1}(h^{-1}(x-x')\in[-1,1]^d)\cdot w(x,\pi_i,\pi_j)w(x',\pi_k,\pi_l)dxdx',\end{aligned} \tag{1.43}$$

and let $Z^Q_{1,N}(x,\pi,\pi') = \frac{Z_{1,N}(x,\pi)+Z_{1,N}(x,\pi')+2Q_1}{1+\epsilon_0} =: Z^Q_{1,N}(x,\pi)+Z^Q_{1,N}(x,\pi')$ and $Z^Q_{2,N}(x,\pi,\pi') = \frac{Z_{2,N}(x,\pi)+Z_{2,N}(x,\pi')+2Q_2}{1+\epsilon_0} =: Z^Q_{2,N}(x,\pi)+Z^Q_{2,N}(x,\pi')$. Then $(Z^Q_{1,N}(x,\pi_i,\pi_j), Z^Q_{2,N}(x',\pi_k,\pi_l))$ is a mean zero multivariate Gaussian process such that, for each $(x,\pi_i,\pi_j)\in\mathbb{R}\times\Pi^2$ and $(x,\pi_k,\pi_l)\in\mathbb{R}\times\Pi^2$, $(Z^Q_{1,N}(x,\pi_i,\pi_j), Z^Q_{2,N}(x',\pi_k,\pi_l))$ and $(S^Q_{\tau,\mathcal{N},1}(x,\pi_i,\pi_j), S^Q_{\tau,\mathcal{N},2}(x',\pi_k,\pi_l))$ have the same covariance structure.

Also, define

$$\bar{\tau}_{N,Q}^2 = \int_B\int_B \sum_{\pi_i\in\Pi}\sum_{\pi_j\in\Pi}\sum_{\pi_k\in\Pi}\sum_{\pi_l\in\Pi}\mathrm{Cov}(|Z^Q_{1,N}(x,\pi_i,\pi_j)|,|Z^Q_{2,N}(x',\pi_k,\pi_l)|)$$

$$\cdot \sqrt{\rho_2(x,\pi_i,\pi_j)\rho_2(x',\pi_k,\pi_l)} \times h^{-d} \cdot \mathbb{1}(h^{-1}(x-x') \in [-1,1]^d)$$
$$\cdot w(x,\pi_i,\pi_j)w(x',\pi_k,\pi_l)dxdx'.$$

Using the triangle inequality, we have

$$\begin{aligned}
|\bar{\tau}^2_{N,Q} - \bar{\sigma}^2_{N,Q}| = &\Bigg| \int_B \int_B \sum_{\pi_i\in\Pi}\sum_{\pi_j\in\Pi}\sum_{\pi_k\in\Pi}\sum_{\pi_l\in\Pi} \Big(\mathrm{Cov}(|Z^Q_{1,N}(x,\pi_i,\pi_j)|, |Z^Q_{2,N}(x',\pi_k,\pi_l)|) \\
&- \mathrm{Cov}(|S^Q_{\tau,\mathcal{N},1}(x,\pi_i,\pi_j)|, |S^Q_{\tau,\mathcal{N},2}(x,\pi_k,\pi_l)|)\Big) \cdot \sqrt{\rho_2(x,\pi_i,\pi_j)\rho_2(x',\pi_k,\pi_l)} \\
&\times h^{-d} \cdot \mathbb{1}(h^{-1}(x-x') \in [-1,1]^d) \cdot w(x,\pi_i,\pi_j)w(x',\pi_k,\pi_l)dxdx' \Bigg| \\
\leq &\int_B \int_B \sum_{\pi_i\in\Pi}\sum_{\pi_j\in\Pi}\sum_{\pi_k\in\Pi}\sum_{\pi_l\in\Pi} \Big|\Big(\mathbb{E}|Z^Q_{1,N}(x,\pi_i,\pi_j)|\mathbb{E}|Z^Q_{2,N}(x',\pi_k,\pi_l)| \\
&- \mathbb{E}|S^Q_{\tau,\mathcal{N},1}(x,\pi_i,\pi_j)|\mathbb{E}|S^Q_{\tau,\mathcal{N},2}(x,\pi_k,\pi_l)|\Big) \cdot \sqrt{\rho_2(x,\pi_i,\pi_j)\rho_2(x',\pi_k,\pi_l)} \\
&\times h^{-d} \cdot \mathbb{1}(h^{-1}(x-x') \in [-1,1]^d) \cdot w(x,\pi_i,\pi_j)w(x',\pi_k,\pi_l)dxdx' \Big| \\
+ &\int_B \int_B \sum_{\pi_i\in\Pi}\sum_{\pi_j\in\Pi}\sum_{\pi_k\in\Pi}\sum_{\pi_l\in\Pi} \Big|\Big(\mathbb{E}|Z^Q_{1,N}(x,\pi_i,\pi_j)||Z^Q_{2,N}(x',\pi_k,\pi_l)| \\
&- \mathbb{E}|S^Q_{\tau,\mathcal{N},1}(x,\pi_i,\pi_j)||S^Q_{\tau,\mathcal{N},2}(x,\pi_k,\pi_l)|\Big) \cdot \sqrt{\rho_2(x,\pi_i,\pi_j)\rho_2(x',\pi_k,\pi_l)} \\
&\times h^{-d} \cdot \mathbb{1}(h^{-1}(x-x') \in [-1,1]^d) \cdot w(x,\pi_i,\pi_j)w(x',\pi_k,\pi_l)dxdx' \Big| \\
\leq &\int_B \int_B \sum_{\pi_i\in\Pi}\sum_{\pi_j\in\Pi}\sum_{\pi_k\in\Pi}\sum_{\pi_l\in\Pi} \Big|\Big(\mathbb{E}|Z^Q_{1,N}(x,\pi_i)|\mathbb{E}|Z^Q_{2,N}(x',\pi_k)| - \mathbb{E}|S_{\tau,\mathcal{N},1}(x,\pi_i)|\mathbb{E}|S_{\tau,\mathcal{N},2}(x',\pi_k)| \\
&+ \mathbb{E}|Z^Q_{1,N}(x,\pi_i)|\mathbb{E}|Z^Q_{2,N}(x',\pi_l)| - \mathbb{E}|S_{\tau,\mathcal{N},1}(x,\pi_i)|\mathbb{E}|S_{\tau,\mathcal{N},2}(x',\pi_l)| \\
&+ \mathbb{E}|Z^Q_{1,N}(x,\pi_j)|\mathbb{E}|Z^Q_{2,N}(x',\pi_k)| - \mathbb{E}|S_{\tau,\mathcal{N},1}(x,\pi_j)|\mathbb{E}|S_{\tau,\mathcal{N},2}(x',\pi_k)| \\
&+ \mathbb{E}|Z^Q_{1,N}(x,\pi_j)|\mathbb{E}|Z^Q_{2,N}(x',\pi_l)| - \mathbb{E}|S_{\tau,\mathcal{N},1}(x,\pi_j)|\mathbb{E}|S_{\tau,\mathcal{N},2}(x',\pi_l)|\Big) \cdot \sqrt{\rho_2(x,\pi_i,\pi_j)\rho_2(x',\pi_k,\pi_l)} \\
&\times h^{-d} \cdot \mathbb{1}(h^{-1}(x-x') \in [-1,1]^d) \cdot w(x,\pi_i,\pi_j)w(x',\pi_k,\pi_l)dxdx' \Big| \\
+ &\int_B \int_B \sum_{\pi_i\in\Pi}\sum_{\pi_j\in\Pi}\sum_{\pi_k\in\Pi}\sum_{\pi_l\in\Pi} \Big|\Big(\mathbb{E}|Z^Q_{1,N}(x,\pi_i)||Z^Q_{2,N}(x',\pi_k)| - \mathbb{E}|S_{\tau,\mathcal{N},1}(x,\pi_i)||S_{\tau,\mathcal{N},2}(x',\pi_k)| \\
&+ \mathbb{E}|Z^Q_{1,N}(x,\pi_i)||Z^Q_{2,N}(x',\pi_l)| - \mathbb{E}|S_{\tau,\mathcal{N},1}(x,\pi_i)||S_{\tau,\mathcal{N},2}(x',\pi_l)| \\
&+ \mathbb{E}|Z^Q_{1,N}(x,\pi_j)||Z^Q_{2,N}(x',\pi_k)| - \mathbb{E}|S_{\tau,\mathcal{N},1}(x,\pi_j)||S_{\tau,\mathcal{N},2}(x',\pi_k)| \\
&+ \mathbb{E}|Z^Q_{1,N}(x,\pi_j)||Z^Q_{2,N}(x',\pi_l)| - \mathbb{E}|S_{\tau,\mathcal{N},1}(x,\pi_j)||S_{\tau,\mathcal{N},2}(x',\pi_l)|\Big) \cdot \sqrt{\rho_2(x,\pi_i,\pi_j)\rho_2(x',\pi_k,\pi_l)} \\
&\times h^{-d} \cdot \mathbb{1}(h^{-1}(x-x') \in [-1,1]^d) \cdot w(x,\pi_i,\pi_j)w(x',\pi_k,\pi_l)dxdx' \Big|
\end{aligned}$$

$$:=\Delta_{1N,Q}(\pi_i,\pi_k)+\Delta_{1N,Q}(\pi_i,\pi_l)+\Delta_{1N,Q}(\pi_j,\pi_k)+\Delta_{1N,Q}(\pi_j,\pi_l)$$
$$+\Delta_{2N,Q}(\pi_i,\pi_k)+\Delta_{2N,Q}(\pi_i,\pi_l)+\Delta_{2N,Q}(\pi_j,\pi_k)+\Delta_{2N,Q}(\pi_j,\pi_l).$$

From the proof of Lemma B.10 in Chang, Lee, and Whang (2015), note that for all $\{\pi,\pi'\}\in\Pi$, $\Delta_{1NQ}(\pi,\pi')=o(1)$, *and* $\Delta_{2NQ}(\pi,\pi')=o(1)$. *Using the infinite divisibility property of Poisson processes, it is straightforward to verify that* $|\bar{\sigma}^2_{N,Q}-\bar{\sigma}^2_{N,0}|\to 0$ *and* $|\bar{\tau}^2_{N,Q}-\bar{\tau}^2_{N,0}|\to 0$ *as* $\epsilon_0\to 0$. *Hence the triangle inequality establishes (1.42), and thus the lemma has been proved.*

Let M be defined as in (1.33)-(1.35) and let

$$U_{_N}:=\frac{1}{\sqrt{N}}\left\{\sum_{i=1}^{N}\mathbb{1}[X_i\in\mathcal{B}(M)]-N\Pr(X\in\mathcal{B}(M))\right\}$$

and

$$V_{_N}:=\frac{1}{\sqrt{N}}\left\{\sum_{i=1}^{N}\mathbb{1}[X_i\in\mathbb{R}^d\setminus\mathcal{B}(M)]-N\Pr(X\in\mathbb{R}^d\setminus\mathcal{B}(M))\right\}$$

Also define

$$S_{_N}:=\frac{T^P_{_{1N}}(B)}{\sigma_{1N}(B)}$$

Lemma 1.7.6 *Under the regularity conditions, we have that*

$$(S_{_N},U_N)\xrightarrow{d}(\mathbb{Z}_1,\ \sqrt{1-\alpha}\mathbb{Z}_2)$$

where $\mathbb{Z}_1$ *and* $\mathbb{Z}_2$ *are independent* $N(0,1)$ *random variables and* α *is defined as in (1.34).*

proof 1.7.2 *Let*

$$\Delta_N(\pi_k,\pi_j,x)=\sqrt{N}\Big\{\left|\Gamma_{\mathcal{N}}(x,\pi_k,\pi_j)-\mathbb{E}[\Gamma_{\mathcal{N}}(x,\pi_k,\pi_j)]\right|$$
$$-\mathbb{E}\left|\Gamma_{\mathcal{N}}(x,\pi_k,\pi_j)-\mathbb{E}[\Gamma_{\mathcal{N}}(x,\pi_k,\pi_j)]\right|\Big\}\cdot w(x,\pi_k,\pi_j)$$

Construct a partition of $\mathcal{B}(M)$. *Consider a regular grid* $G_{\boldsymbol{i}}=(i_1h,(i_1+1)h]\times\cdots\times(i_dh,(i_d+1)h]$ *where* $\boldsymbol{i}=(i_1,\ldots,i_d)$, $i_1,\ldots,i_d$ *are integers. Define* $R_{\boldsymbol{i}}=G_{\boldsymbol{i}}\cap\mathcal{B}(M)$, $\mathcal{I}_i=\{\boldsymbol{i}\in\mathbb{Z}^d:(G_{\boldsymbol{i}}\cap\mathcal{B}(M))\neq\emptyset\}$. *Then, we see that* $\{R_{\boldsymbol{i}}:\boldsymbol{i}\in\mathcal{I}_i\subset\mathbb{Z}^d\}$ *is a partition of* $\mathcal{B}(M)$ *with* $\lambda(R_{\boldsymbol{i}})\le A_1h^d$, $m_N:=\#(\mathcal{I}_i)\le A_2h^{-d}$ *for some positive constants* A_1 *and* A_2.

In addition, set

$$\alpha_{\boldsymbol{i},N} = \frac{\int_{R_{\boldsymbol{i}}} \mathbb{1}(x \in B) \cdot \sum_{k=1}^{K} \sum_{j=1}^{K} \Delta_N(\pi_k, \pi_j, x) dx}{\sigma_N(B)}$$

and

$$u_{\boldsymbol{i},N} = \frac{1}{\sqrt{N}} \left\{ \sum_{j=1}^{N} \mathbb{1}(X_j \in R_{\boldsymbol{i}}) - N \Pr(X \in R_{\boldsymbol{i}}) \right\}.$$

Then, we have $S_N = \sum_{\boldsymbol{i} \in \mathcal{I}_i} \alpha_{\boldsymbol{i},N}$ *and* $U_N = \sum_{\boldsymbol{i} \in \mathcal{I}_i} u_{\boldsymbol{i},N}$. *We can verify that* $\mathrm{Var}(S_N) = 1$ *and* $\mathrm{Var}(U_N) = 1 - \alpha$. *For arbitrary* λ_1 *and* $\lambda_2 \in \mathbb{R}$, *let*

$$y_{\boldsymbol{i},N} = \lambda_1 \alpha_{\boldsymbol{i},N} + \lambda_2 u_{\boldsymbol{i},N}.$$

Notice that $\{y_{\boldsymbol{i},N} : \boldsymbol{i} \in \mathcal{I}_i\}$ *is an array of mean zero one-dependent random fields.*

We need to show that

$$\mathrm{Var}\left(\sum_{\boldsymbol{i} \in \mathcal{I}_i} y_{i,N} \right) = \mathrm{Var}(\lambda_1 S_N + \lambda_2 U_N) \to \lambda_1^2 + \lambda_2^2 (1 - \alpha) \tag{1.44}$$

and

$$\sum_{\boldsymbol{i} \in \mathcal{I}_i} \mathbb{E}|y_{i,N}|^r = o(1) \text{ for some } 2 < r < 3. \tag{1.45}$$

The results of the Lemma follows from the central limit theorem of Shergin (1993) and the Cramér-Wold device. To show, (1.44)*, which holds if we have*

$$\mathrm{Cov}(S_N, U_N) = O\left(\frac{1}{\sqrt{Nh^{2d}}} \right), \tag{1.46}$$

which implies that

$$\mathrm{Cov}\left(\int_B \sum_{k=1}^{K} \sum_{j=1}^{K} \left\{ \sqrt{N} \left| [(\Gamma_{\mathcal{N}}(x, \pi_k, \pi_j)] - \mathbb{E}[(\Gamma_{\mathcal{N}}(x, \pi_k, \pi_j)] \right| \right\} w(x, \pi_j, \pi_k) dx, U_N \right) = O\left(\frac{1}{\sqrt{Nh^{2d}}} \right). \tag{1.47}$$

For any $(x, \pi_k, \pi_j) \in B \times \Pi^2$ *we have*

$$\left(S_{\tau,N}(x, \pi_k, \pi_j), \frac{U_N}{\sqrt{\Pr(X \in B(M))}} \right) \overset{d}{=} \left(\frac{1}{\sqrt{N}} \sum_{i=1}^{N} Q_{\tau,N}^{(i)}(x, \pi_k, \pi_j), \frac{1}{\sqrt{N}} \sum_{i=1}^{N} U^{(i)} \right) \tag{1.48}$$

where $(Q^{(i)}_{\tau,N}(x,\pi_k,\pi_j), U^{(i)})$ *for* $i = 1,\ldots,N$ *are i.i.d. copies of* $(Q_{\tau,N}(x,\pi_k,\pi_j), U)$ *with* $Q_{\tau,N}(x,\pi_k,\pi_j)$ *is defined as*

$$Q_{\tau,N}(x,\pi_k,\pi_j) = \left[\sum_{i=1}^{\eta} \Theta(W_i,x,\pi_k,\pi_j) - \mathbb{E}[\Theta(W,x,\pi_k,\pi_j)]\right] / \sqrt{\mathbb{E}[\Theta^2(W_i,x,\pi_k,\pi_j)]}$$

and

$$U = \left[\sum_{i=1}^{\eta} \mathbb{1}[X_i \in \mathcal{B}(M)] - \Pr(X \in \mathcal{B}(M))\right] / \sqrt{\Pr(X \in \mathcal{B}(M))}.$$

where η *denote an independent Poisson random variable with mean 1 that is independent of* $\{W_i : n \geq 1\}$. *Let* $(Z_{1N}(x,\pi_k,\pi_j), Z_{2N})$ *for* $(x,\pi_k,\pi_j) \in \mathbb{R}\times\Pi^2$ *be a mean zero Gaussian process such that, for each* $(x,\pi_k,\pi_j) \in \mathbb{R}\times\Pi^2$, $(Z_{1N}(x,\pi_k,\pi_j), Z_{2N})$ *and the left-hand side of (1.48) has the same covariance structure. That is,*

$$(Z_{1N}(x,\pi_k,\pi_j), Z_{2N}) \overset{d}{=} (\sqrt{1-(\gamma_i^*(x,\pi_k,\pi_j))^2}\mathbb{Z}_1 + \gamma_i^*(x,\pi_k,\pi_j))\mathbb{Z}_2, \mathbb{Z}_2),$$

where $\mathbb{Z}_1$ *and* $\mathbb{Z}_2$ *are independent standard normal random variables and*

$$\gamma_i^*(x,\pi_k,\pi_j)) = \mathbb{E}\left[S_{\tau,\mathcal{N}}(x,\pi_k,\pi_j).\ \frac{U_N}{\Pr(X \in \mathcal{B}(M))}\right].$$

We can show that

$$\sup_{\Pi\times B}\left|\mathbb{E}\left[\frac{\tau_{\mathcal{N}}(x,\pi_k) - \mathbb{E}[\tau_{\mathcal{N}}(x,\pi_k)]}{\sqrt{\kappa_{\Gamma,N}(x,\pi_k,\pi_j)}} \cdot \frac{U_N}{\Pr(X \in \mathcal{B}(M))}\right]\right| = O(h^{d/2}),$$

Using the additive property of the Big-O notation, and the triangle inequality, notice that we have

$$\sup_{B\times\Pi^2} |\gamma_i^*(x,\pi_k,\pi_j))| = O(h^{d/2}), \tag{1.49}$$

which in turn is less than or equal to ϵ *for all sufficiently large N and any* $0 < \epsilon < 1/2$. *This result and (1.27) imply that*

$$\sup_{B\times\Pi^2}\left|\mathrm{Cov}\left(|S_{\tau,\mathcal{N}}(x,\pi_k,\pi_j)|, \frac{U_N}{\Pr(X \in \mathcal{B}(M))}\right) - \mathbb{E}[|Z_{1N}(x,\pi_k,\pi_j)|Z_{2N}]\right| \leq O\left(\frac{1}{\sqrt{N\cdot h^{2d}}}\right) \tag{1.50}$$

Using the additive property of the Big-O notation, and the triangle inequality, this implies that

$$\sup_B \left| \text{Cov}\left(\left| \sum_{k=1}^{K} \sum_{j=1}^{K} S_{\tau,\mathcal{N}}(x, \pi_k, \pi_j) \right|, \frac{U_N}{\Pr(X \in \mathcal{B}(M))} \right) - \mathbb{E}\left[\left| \sum_{k=1}^{K} \sum_{j=1}^{K} Z_{1N}(x, \pi_k, \pi_j) \right| Z_{2N} \right] \right| \leq O\left(\frac{1}{\sqrt{N \cdot h^{2d}}} \right). \tag{1.51}$$

On the other hand,

$$\begin{aligned} \sup_{B \times \Pi^2} |\mathbb{E}[Z_{1N}(x, \pi_k, \pi_j)|Z_{2N}| &= \sup_{B \times \Pi^2} |\gamma_i^*(x, \pi_k, \pi_j) \mathbb{E}[|Z_{1N}(x, \pi_k, \pi_j)| Z_{1N}(x, \pi_k, \pi_j)] \\ &\leq \sup_{B \times \Pi^2} |\gamma_i^*(x, \pi_k, \pi_j)| \mathbb{E}[Z_{1N}^2(x, \pi_k, \pi_j)] \\ &= \sup_{B \times \Pi^2} |\gamma_i^*(x, \pi_k, \pi_j)| = O(h^{d/2}). \end{aligned}$$

Using the law of iterated expectations and (1.49). This also implies that

$$\sup_B \left| \mathbb{E}\left[\left| \sum_{k=1}^{K} \sum_{j=1}^{K} Z_{1N}(x, \pi_k, \pi_j) \right| Z_{2N} \right] \right| = O(h^{d/2}), \tag{1.52}$$

using the additive property of the Big-O notation, and the triangle inequality.

Therefore, (1.51) and (1.52) imply that

$$\sup_B |\text{Cov}(\sqrt{N}| \sum_{k=1}^{K} \sum_{j=1}^{K} [(\Gamma_{\mathcal{N}}(x, \pi_k, \pi_j)] - \mathbb{E}[(\Gamma_{\mathcal{N}}(x, \pi_k, \pi_j)]|, U_N)| \leq O\left(\frac{1}{Nh^{2d}} + h^{d/2} \right)$$

which when combined with $\lambda(B) < \infty$ yields (1.47) *and hence* (1.44) *as desired.*

Next we establish (1.45). Chang, Lee, and Whang (2015) shows that for any $\pi_k \in \Pi$,

$$\sum_{\boldsymbol{i} \in \mathcal{I}_i} \mathbb{E} \left| \frac{\int_{R_i} \mathbb{1}(x \in B) \sqrt{N_k} \{ |\tau_{\mathcal{N}}(x, \pi_k) - \mathbb{E}\tau_{\mathcal{N}}(x, \pi_k)| - \mathbb{E}[|\tau_{\mathcal{N}}(x, \pi_k) - \mathbb{E}\tau_{\mathcal{N}}(x, \pi_k)|] \} w(x, \pi_k)}{\sigma_{N_k}(B)} \right|^r \leq O(N_k \cdot h^{rd/2}) = o(1) \tag{1.53}$$

Notice that,

$$\begin{aligned}\sum_{\boldsymbol{i}\in\mathcal{I}_i}\mathbb{E}|\alpha_{\boldsymbol{i},N}| &= \sum_{\boldsymbol{i}\in\mathcal{I}_i}\mathbb{E}\left|\int_{R_{\boldsymbol{i}}}\mathbb{1}(x\in B)\cdot\sum_{k=1}^{K}\sum_{j=1}^{K}\Delta_i(\pi_k,\pi_j,x)dx\right|^r \\ &= \sum_{\boldsymbol{i}\in\mathcal{I}_i}\mathbb{E}\left|\sum_{k=1}^{K}\sum_{j=1}^{K}\int_{R_{\boldsymbol{i}}}\mathbb{1}(x\in B)\cdot\Delta_i(\pi_k,\pi_j,x)dx\right|^r \\ &\leq \sum_{\boldsymbol{i}\in\mathcal{I}_i}\sum_{k=1}^{K}\sum_{j=1}^{K}\mathbb{E}\left|\int_{R_{\boldsymbol{i}}}\mathbb{1}(x\in B)\cdot\Delta_i(\pi_k,\pi_j,x)dx\right|^r \\ &= o(1)\end{aligned} \tag{1.54}$$

using (1.53), the triangle inequality and the additive property of the Big-O notation.

Also, from existing results we can verify that

$$\sum_{\boldsymbol{i}\in\mathcal{I}_i}\mathbb{E}|u_{\boldsymbol{i},N}|^r\to 0 \tag{1.55}$$

Therefore, combining (1.44) *and* (1.45)*, we have (1.45). This now completes the proof of the Lemma*

We are now ready to prove asymptotic normality

Lemma 1.7.7 *Under the regularity conditions, the following holds*

$$\begin{aligned}&\lim_{N\to\infty}\int_B\sum_{k=1}^{K}\sum_{j=1,j\neq k}^{K}\Bigg\{\sqrt{N}\mathbb{E}\left[\left|[(\Gamma_{\mathcal{N}}(x,\pi_k,\pi_j)]-\mathbb{E}[(\Gamma_{\mathcal{N}}(x,\pi_k,\pi_j)]\right|\right] \\ &-\mathbb{E}|\mathbb{Z}|\frac{K(K-1)}{2}\kappa^{1/2}_{\tau,N}(x,\pi_k,\pi_j)\Bigg\}w(x,\pi_j,\pi_k)dx=0\end{aligned}$$

and

$$\begin{aligned}&\lim_{N\to\infty}\int_B\sum_{k=1}^{K}\sum_{j=1,j\neq k}^{K}\Bigg\{\sqrt{N}\mathbb{E}\left[\left|[(\Gamma_{\mathcal{N}}(x,\pi_k,\pi_j)]-\mathbb{E}[(\Gamma_{\mathcal{N}}(x,\pi_k,\pi_j)]\right|\right] \\ &-\mathbb{E}|\mathbb{Z}_1|\cdot\frac{K(K-1)}{2}\kappa^{1/2}_{\tau,N}(x,\pi_k,\pi_j)\Bigg\}w(x,\pi_j,\pi_k)dx=0,\end{aligned}$$

where $\mathbb{Z}$ is a standard normal random variable.

Using Lemma 1.7.4, and similar arguments in the proof of Lemma 6.3 of Giné, Mason, and Zaitsev (2003), the results are established.

Define

$$L_N(B) = \frac{\sqrt{N}}{\sigma_N(B)} \int_B \sum_{k=1}^{K} \sum_{j=1, j\neq k}^{K} \Big\{ \Big| [(\Gamma_N(x, \pi_k, \pi_j)] - \mathbb{E}[(\Gamma_N(x, \pi_k, \pi_j)] \Big| - \mathbb{E} \Big| [(\Gamma_N(x, \pi_k, \pi_j)] - \mathbb{E}[(\Gamma_N(x, \pi_k, \pi_j)] \Big| \Big\} w(x, \pi_j, \pi_k) dx.$$

Lemma 1.7.8 *Under the regularity conditions, we have*

$$L_N(B) \xrightarrow{d} \mathbb{Z}$$

as $N \to \infty$, where $\mathbb{Z}$ is a standard normal random variable.

proof 1.7.3 *Notice that*

$$S_N = \frac{\sqrt{N}}{\sigma_N(B)} \int_B \sum_{k=1}^{K} \sum_{j=1, j\neq k}^{K} \Big\{ \sqrt{N} \Big| [(\Gamma_{\mathcal{N}}(x, \pi_k, \pi_j)] - \mathbb{E}[(\Gamma_{\mathcal{N}}(x, \pi_k, \pi_j)] \Big| \Big\} w(x, \pi_j, \pi_k) - \frac{\sqrt{N}}{\sigma_N(B)} \int_B \sum_{k=1}^{K} \sum_{j=1, j\neq k}^{K} \Big\{ \sqrt{N} \mathbb{E} \Big| [(\Gamma_{\mathcal{N}}(x, \pi_k, \pi_j)] - \mathbb{E}[(\Gamma_{\mathcal{N}}(x, \pi_k, \pi_j)] \Big| \Big\} w(x, \pi_j, \pi_k) dx$$

By the de-poissonization arguments of Beirlant and Mason (1995), we have

$$\begin{aligned}(S_N | \mathcal{N} = N) \stackrel{d}{=} & \frac{\sqrt{N}}{\sigma_N(B)} \int_B \sum_{k=1}^{K} \sum_{j=1, j\neq k}^{K} \Big\{ \sqrt{N} \Big| [(\Gamma_N(x, \pi_k, \pi_j)] - \mathbb{E}[(\Gamma_N(x, \pi_k, \pi_j)] \Big| \Big\} w(x, \pi_j, \pi_k) \\ & - \frac{\sqrt{N}}{\sigma_N(B)} \int_B \sum_{k=1}^{K} \sum_{j=1, j\neq k}^{K} \Big\{ \sqrt{N} \mathbb{E} \Big| [(\Gamma_N(x, \pi_k, \pi_j)] - \mathbb{E}[(\Gamma_N(x, \pi_k, \pi_j)] \Big| \Big\} w(x, \pi_j, \pi_k) dx \\ \to & \mathbb{Z}\end{aligned}$$

Now from Lemma 1.7.7, we know that

$$\lim_{N\to\infty} \int_B \sum_{k=1}^{K} \sum_{j=1, j\neq k}^{K} \Big\{ \sqrt{N} \mathbb{E} \Big| [(\Gamma_{\mathcal{N}}(x, \pi_k, \pi_j)] - \mathbb{E}[(\Gamma_{\mathcal{N}}(x, \pi_k, \pi_j)] \Big| \Big\} w(x, \pi_j, \pi_k) dx - \int_B \sum_{k=1}^{K} \sum_{j=1, j\neq k}^{K} \Big\{ \sqrt{N} \mathbb{E} \Big| [(\Gamma_{\mathcal{N}}(x, \pi_k, \pi_j)] - \mathbb{E}[(\Gamma_{\mathcal{N}}(x, \pi_k, \pi_j)] \Big| \Big\} w(x, \pi_j, \pi_k) dx = 0.$$

Hence, we have

$$L_N(B) \xrightarrow{d} \mathbb{Z}$$

as $N \to \infty$, *as required*

We are now ready to prove Theorem 1.7.2. We can show that for any $\pi_k \in \Pi$,

$$\begin{aligned}
&\lim_{N_k\to\infty} \sup \mathbb{E}\left(\sqrt{N_k} \int_{B_l^c} \{|\tau_{N_k}(x,\pi_k) - \mathbb{E}\tau_{N_k}(x,\pi_k)| - \mathbb{E}[|\tau_{N_k}(x,\pi_k) - \mathbb{E}\tau_{N_k}(x,\pi_k)|]\} w(x,\pi_k)dx \right)^2 \\
&\leq C_2 \lambda(\mathcal{X}) \left(\sup_{x\in\mathcal{X}} \mathbb{E}[|Y^2 \hat{\phi}(\pi_k, x, T)^2 || X = x] + E[|\chi(\pi_k, x, T)^2|] \right) \int_{B_l^c} f(x)dx,
\end{aligned} \tag{1.56}$$

where C_2 is a positive constant and $\{B_l : l > 1\}$ is a sequence of Borel sets in $\mathbb{R}^d$ that has a finite Lebesgue measure $\lambda(B_l)$ and satisfies (1.33)-(1.35) with $B_l = B$ for each l and let

$$\lim_{l\to\infty} \int_{B_l^c} f(x)dx = 0. \tag{1.57}$$

Also, for each $l \geq 1$, by Lemma 1.7.5, we have

$$\lim_{l\to\infty} \sigma^2_{1,B_l} = \sigma^2_1. \tag{1.58}$$

We can show that

$$\begin{aligned}
&\lim_{N\to\infty} \sup \mathbb{E}\Bigg(\sqrt{N} \int_{B_l^c} \sum_{k=1}^{K} \sum_{j=1}^{K} \Bigg\{ \Big| [(\Gamma_{\mathcal{N}}(x,\pi_k,\pi_j)] - \mathbb{E}[(\Gamma_{\mathcal{N}}(x,\pi_k,\pi_j)] \Big| - \\
&\mathbb{E}\Big| [(\Gamma_{\mathcal{N}}(x,\pi_k,\pi_j)] - \mathbb{E}[(\Gamma_{\mathcal{N}}(x,\pi_k,\pi_j)] \Big| \Bigg\} w(x,\pi_j,\pi_k)dx \Bigg)^2 \\
&\leq \lim_{N\to\infty} \sup \mathbb{E}\Bigg(\sqrt{N} \int_{B_l^c} \sum_{k=1}^{K} \sum_{j=1}^{K} \{|\tau_{\mathcal{N}_k}(x,\pi_k) - \mathbb{E}[\tau_{i_k}(x,\pi_k)]| - \mathbb{E}|\tau_{\mathcal{N}_k}(x,\pi_k) - \mathbb{E}[\tau_{i_k}(x,\pi_k)]| + \\
&|\tau_{\mathcal{N}_j}(x,\pi_j) - \mathbb{E}[\tau_{i_j}(x,\pi_j)]| - \mathbb{E}|\tau_{\mathcal{N}_j}(x,\pi_j) - \mathbb{E}[\tau_{i_j}(x,\pi_j)]| \} w(x,\pi_j,\pi_k)dx \Bigg)^2 \\
&\leq \lim_{N\to\infty} \sup \mathbb{E}\Bigg(N \int_{B_l^c} 2 \sum_{k=1}^{K} \sum_{j=1}^{K} \Bigg(\{|\tau_{\mathcal{N}_k}(x,\pi_k) - \mathbb{E}[\tau_{i_k}(x,\pi_k)]| - \mathbb{E}|\tau_{\mathcal{N}_k}(x,\pi_k) - \mathbb{E}[\tau_{i_k}(x,\pi_k)]| + \\
&|\tau_{\mathcal{N}_j}(x,\pi_j) - \mathbb{E}[\tau_{i_j}(x,\pi_j)]| - \mathbb{E}|\tau_{\mathcal{N}_j}(x,\pi_j) - \mathbb{E}[\tau_{i_j}(x,\pi_j)]| \} w(x,\pi_j,\pi_k)dx \Bigg)^2 \Bigg) \\
&\leq \lim_{N\to\infty} \sup \mathbb{E}\Bigg(N \int_{B_l^c} 4 \sum_{k=1}^{K} \sum_{j=1}^{K} \Bigg(\{|\tau_{\mathcal{N}_k}(x,\pi_k) - \mathbb{E}[\tau_{i_k}(x,\pi_k)]| - \mathbb{E}|\tau_{\mathcal{N}_k}(x,\pi_k) - \mathbb{E}[\tau_{i_k}(x,\pi_k)]| \Bigg)^2 +
\end{aligned}$$

$$
\begin{aligned}
&\left(|\tau_{\mathcal{N}_j}(x,\pi_j) - \mathbb{E}[\tau_{i_j}(x,\pi_j)]| - \mathbb{E}|\tau_{\mathcal{N}_j}(x,\pi_j) - \mathbb{E}[\tau_{i_j}(x,\pi_j)]|\}w(x,\pi_j,\pi_k)dx\right)^2\Bigg) \\
&= \lim_{N\to\infty}\sup\Bigg\{4N\mathbb{E}\Bigg(\int_{B_l^c} K\sum_{k=1}^{K}\left(|\tau_{\mathcal{N}_k}(x,\pi_k) - \mathbb{E}[\tau_{i_k}(x,\pi_k)]| - \mathbb{E}|\tau_{\mathcal{N}_k}(x,\pi_k) - \mathbb{E}[\tau_{i_k}(x,\pi_k)]|w(x,\pi_j,\pi_k)dx\right)^2\Bigg)+ \\
&\mathbb{E}\Bigg(\int_{B_l^c} K\sum_{j=1}^{K}\left(|\tau_{\mathcal{N}_j}(x,\pi_j) - \mathbb{E}[\tau_{i_j}(x,\pi_j)]| - \mathbb{E}|\tau_{\mathcal{N}_j}(x,\pi_j) - \mathbb{E}[\tau_{i_j}(x,\pi_j)]|w(x,\pi_j,\pi_k)dx\right)^2\Bigg)\Bigg\} \\
&= \lim_{N\to\infty}\sup\Bigg\{4NK\int_{B_l^c}\sum_{k=1}^{K}\mathbb{E}\Bigg(|\tau_{\mathcal{N}_k}(x,\pi_k) - \mathbb{E}[\tau_{i_k}(x,\pi_k)]| - \mathbb{E}|\tau_{\mathcal{N}_k}(x,\pi_k) - \mathbb{E}[\tau_{i_k}(x,\pi_k)]|w(x,\pi_j,\pi_k)dx\Bigg)^2 + \\
&4NK\int_{B_l^c}\sum_{j=1}^{K}\mathbb{E}\Bigg(|\tau_{\mathcal{N}_j}(x,\pi_j) - \mathbb{E}[\tau_{i_j}(x,\pi_j)]| - \mathbb{E}|\tau_{\mathcal{N}_j}(x,\pi_j) - \mathbb{E}[\tau_{i_j}(x,\pi_j)]|w(x,\pi_j,\pi_k)dx\Bigg)^2\Bigg\} \\
&\leq \lim_{N\to\infty}\sup\Bigg\{4K\sum_{k=1}^{K}\mathbb{E}\Bigg(\sqrt{N}\int_{B_l^c}|\tau_{\mathcal{N}_k}(x,\pi_k) - \mathbb{E}[\tau_{i_k}(x,\pi_k)]| - \mathbb{E}|\tau_{\mathcal{N}_k}(x,\pi_k) - \mathbb{E}[\tau_{i_k}(x,\pi_k)]|w(x,\pi_j,\pi_k)dx\Bigg)^2\Bigg\}+ \\
&\lim_{N\to\infty}\sup\Bigg\{4K\sum_{j=1}^{K}\mathbb{E}\Bigg(\sqrt{N}\int_{B_l^c}|\tau_{\mathcal{N}_j}(x,\pi_j) - \mathbb{E}[\tau_{i_j}(x,\pi_j)]| - \mathbb{E}|\tau_{\mathcal{N}_j}(x,\pi_j) - \mathbb{E}[\tau_{i_j}(x,\pi_j)]|w(x,\pi_j,\pi_k)dx\Bigg)^2\Bigg\} \\
&\leq 4K\sum_{k=1}^{K}\lim_{N\to\infty}\sup\mathbb{E}\Bigg(\sqrt{N}\int_{B_l^c}|\tau_{\mathcal{N}_k}(x,\pi_k) - \mathbb{E}[\tau_{i_k}(x,\pi_k)]| - \mathbb{E}|\tau_{\mathcal{N}_k}(x,\pi_k) - \mathbb{E}[\tau_{i_k}(x,\pi_k)]|w(x,\pi_j,\pi_k)dx\Bigg)^2 + \\
&4K\sum_{j=1}^{K}\lim_{N\to\infty}\sup\mathbb{E}\Bigg(\sqrt{N}\int_{B_l^c}|\tau_{\mathcal{N}_j}(x,\pi_j) - \mathbb{E}[\tau_{i_j}(x,\pi_j)]| - \mathbb{E}|\tau_{\mathcal{N}_j}(x,\pi_j) - \mathbb{E}[\tau_{i_j}(x,\pi_j)]|w(x,\pi_j,\pi_k)dx\Bigg)^2 \\
&\leq 4K\sum_{k=1}^{K}C_2\lambda(\mathcal{X})\left(\sup_{x\in\mathcal{X}}\mathbb{E}[|Y^2\hat{\phi}(\pi_k,x,T)^2||X=x] + E[|\chi(\pi_k,x,T)^2|]\right)\int_{B_l^c}f(x)dx+ \\
&4K\sum_{j=1}^{K}C_2\lambda(\mathcal{X})\left(\sup_{x\in\mathcal{X}}\mathbb{E}[|Y^2\hat{\phi}(\pi_k,x,T)^2||X=x] + E[|\chi(\pi_k,x,T)^2|]\right)\int_{B_l^c}f(x)dx \\
&= 8KC_2\lambda(\mathcal{X})\cdot\left\{\sum_{k=1}^{K}\left(\sup_{x\in\mathcal{X}}\mathbb{E}[|Y^2\hat{\phi}(\pi_k,x,T)^2||X=x] + E[|\chi(\pi_k,x,T)^2|]\right)\right\}\cdot\int_{B_l^c}f(x)dx.
\end{aligned}
$$

The first inequality is as a result of the triangle inequality. The second and third inequalities are as a result of the of the fact that $(\sum_i a_i)^2 \leq \sum_i a_i^2$. The fourth and fifth inequalities are as a result of the linear properties of limsup. The sixth inequality is as a result of the the inequality in (1.56).

Using this result, with the results in (1.57)–(1.58) and Theorem 4.2 in Billingsley (1968), we conclude that

$$
\int_{\mathcal{X}}\sum_{k=1}^{K}\sum_{j=1,j\neq k}^{K}\left\{\sqrt{N}\left|[\tau_N(x,\pi_k) - \tau_N(x,\pi_j)] - \mathbb{E}[\tau_N(x,\pi_k) - \tau_N(x,\pi_j)]\right|\right\}w(x,\pi_j,\pi_k)dx \xrightarrow{d} \sigma_0\mathbb{Z}.
$$

The proof is complete because we can use Lemma (1.7.7) to show that

$$\lim_{N\to\infty}\int_B \sum_{k=1}^{K}\sum_{j=1,j\neq k}^{K}\left\{\sqrt{N}\mathbb{E}\left[\left|[(\Gamma_N(x,\pi_k,\pi_j)]w(x,\pi_j,\pi_k)-\mathbb{E}[(\Gamma_N(x,\pi_k,\pi_j)]\right|\right]-a_i\right\}dx=0.$$

1.7.4.2 Sketch Proof of Asymptotic normality of $\hat{T}_2$

Uniform Asymptotic approximation of $\hat{T}_2$

Write

$$\hat{\tau}(x,\pi)=\tau(x,\pi)+(\tau_{N0}(x,\pi)-\mathbb{E}(\tau_{N0}(x,\pi)))+(\mathbb{E}(\tau_{N0}(x,\pi))-\tau(x,\pi))+R_N(x,\pi)$$

where

$$\tau_{N0}(x,\pi):=\frac{1}{Nh^d}\sum_{i=1}^{N}Y_i\mathbb{1}(\Pi_i=\pi)\left[\frac{T_i}{P_1(x;\pi)}-\frac{(1-T_i)}{P_0(x;\pi)}\right]\cdot K\left(\frac{x-X_i}{h}\right)$$

and

$$\begin{aligned}R_N(x,\pi):=&\frac{1}{Nh^d}\sum_{i=1}^{N}Y_i\mathbb{1}(\Pi_i=\pi)\left[\frac{T_i}{P_1(x;\pi)}-\frac{(1-T_i)}{P_0(x;\pi)}\right]\\&\times\left(T_i\frac{P_1(x,\pi)-\hat{P}_1(x,\pi)}{\hat{P}_1(x,\pi)}+(1-T_i)\frac{P_0(x,\pi)-\hat{P}_0(x,\pi)}{\hat{P}_0(x,\pi)}\right)\cdot K\left(\frac{x-X_i}{h}\right).\end{aligned}$$

Therefore,

$$\begin{aligned}\hat{\Gamma}(x,x',\pi_k)=&\tau(x,\pi_k)-\tau(x',\pi_k)+(\tau_{N0}(x,\pi_k)-\tau_{N0}(x',\pi_k)-\mathbb{E}(\tau_{N0}(x,\pi_k)-\tau_{N0}(x',\pi_k)))\\&+(\mathbb{E}(\tau_{N0}(x,\pi_k)-\tau_{N0}(x',\pi_k))-\tau(x,\pi_k)-\tau(x',\pi_k)))+R_N(x,\pi_k)-R_N(x',\pi_k)\end{aligned}$$

Now, define

$$\begin{aligned}\zeta_N(x,\pi)=&\mathbb{E}[Y|X=x,\Pi=\pi,T=1]-\mathbb{E}[Y|X=x,\Pi=\pi,T=0]\\&-\mathbb{E}[Y|X=x,\Pi=\pi,T=1]\frac{1}{Nh^dP_1(x,\pi)}\sum_{i=1}^{N}T_i\mathbb{1}(\Pi=\pi)K\left(\frac{x-X_i}{h}\right)\\&+\mathbb{E}[Y|X=x,\Pi=\pi,T=0]\frac{1}{Nh^dP_0(x,\pi)}\sum_{i=1}^{N}(1-T_i)\mathbb{1}(\Pi=\pi)K\left(\frac{x-X_i}{h}\right).\end{aligned}$$

The following lemma shows that $R_N(x,\pi)$ can be approximated by $\zeta_N(x,\pi)$ uniformly over x at a

rate faster than $N^{-1/2}$.

Under the regularity conditions, we find that for $\pi_k \in \mathbf{\Pi}$,

$$\sup_{x\in\mathcal{X}} |(R_N(x,\pi_k) - R_N(x',\pi_k)) - (\zeta_N(x,\pi_k) - \zeta_N(x',\pi_k))| = o_p(N^{-1/2}).$$

From, Lemma B.1 in Chang, Lee, and Whang (2015) for $\pi_k \in \Pi$, we have

$$\sup_{x\in\mathcal{X}} |R_N(x,\pi_k) - \zeta_N(x,\pi_k))| = o_p(N^{-1/2}).$$

Now,

$$\begin{aligned}
&\sup_{x,x'\in\mathcal{X}\times\mathcal{X}} |(R_N(x,\pi_k) - R_N(x',\pi_k)) - (\zeta_N(x,\pi_k) - \zeta_N(x',\pi_k))| \\
&\quad\leq \sup_{x,x'\in\mathcal{X}\times\mathcal{X}} \{|R_N(x,\pi_k) - \zeta_N(x,\pi_k)| + |R_N(x',\pi_k) - \zeta_N(x',\pi_k)|\} \\
&\quad\leq \sup_{x\in\mathcal{X}}\{|R_N(x,\pi_k) - \zeta_N(x,\pi_k)|\} + \sup_{x'\in\mathcal{X}}\{|R_N(x',\pi_k) - \zeta_N(x',\pi_k)|\} \\
&\quad= o_p(N^{-1/2}) + o_p(N^{-1/2}) = o_p(N^{-1/2}).
\end{aligned}$$

Under the regularity conditions, we have

$$\hat{T}_2 - T^*_{2N} = o_p(1),$$

where

$$\begin{aligned}
T^*_{2N} := \int_{\mathcal{X}}\int_{\mathcal{X}} \sum_{k=1}^{K} \Bigg\{ \sqrt{N} \Bigg| &\Gamma(x,x',\pi_k) + [\tau_N(x,\pi_k) - \tau_N(x',\pi_k)] \\
&- \mathbb{E}[\tau_N(x,\pi_k) - \tau_N(x',\pi_k)] \Bigg| \Bigg\} \frac{w(x,x',\pi_k)}{2} dx dx'
\end{aligned}$$

and

$$\tau_N(x,\pi) = \tau_{N0}(x,\pi) + \zeta_N(x,\pi).$$

Hence, under the null hypothesis such that $\tau(x,\pi_k) = \tau(x',\pi_k)$ on $\mathcal{X}\times\Pi$, we have

$$\hat{T}_2 = T_{2N} + o_p(1),$$

where

$$T_{2N} := \int_{\mathcal{X}}\int_{\mathcal{X}}\sum_{k=1}^{K}\left\{\sqrt{N}\,|[\tau_N(x,\pi_k)-\tau_N(x',\pi_k)]-\mathbb{E}[\tau_N(x,\pi_k)-\tau_N(x',\pi_k)]|\right\}\frac{w(x,x',\pi_k)}{2}dxdx'$$

and

$$\begin{aligned}\tau_N(x,\pi) :=& \frac{1}{Nh^d}\sum_{i=1}^{N}\Bigg(\{Y-\mathbb{E}[Y|X=x,\Pi=\pi,T=1]\}\frac{T_i\cdot\mathbb{1}(\Pi=\pi)}{P_1(x,\pi_k)}\\ &-\{Y-\mathbb{E}[Y|X=x,\Pi=\pi,T=0]\}\frac{(1-T_i)\cdot\mathbb{1}(\Pi=\pi)}{P_0(x,\pi_k)}\Bigg)\cdot K\left(\frac{x-X_i}{h}\right).\end{aligned}$$

Now, rewrite T_{2N} as an average and poissonize. To begin, define

$$\chi_{t,k} := \frac{\mathbb{E}[Y|X=x,\Pi=\pi_k,T=t]}{P_t(x,\pi_k)}$$

$$\chi(\pi_k,x,T) := [\chi_{1,k}(\pi_k,x)\cdot T-\chi_{0,k}(\pi_k,x)\cdot(1-T)]\cdot\mathbb{1}(\Pi=\pi_k)$$

$$\psi(W_i,x,\pi_k) := [Y_i\cdot\mathbb{1}(\Pi_i=\pi_k)\phi(x,\pi_k,T_i)-\chi(\pi_k,x,T_i)]\frac{1}{h^d}K\left(\frac{x-X_i}{h}\right)+\tau(x;\pi_k).$$

Then

$$\tau_{N}(x,\pi_k)=\tau_{N0}(x,\pi_k)+\zeta_{N}(x,\pi_k)=\frac{1}{N}\sum_{i=1}^{N}\psi(W_i,x,\pi_k)$$

Hence define,

$$\begin{aligned}\Gamma_N(x,x',\pi_k) :=& \tau_{N}(x,\pi_k)-\tau_{N}(x',\pi_k)\\ =&\frac{1}{N}\sum_{i=1}^{N}\psi(W_i,x,\pi_k)-\frac{1}{N}\sum_{i=1}^{N}\psi(W_i,x',\pi_k)\\ =&\frac{1}{N}\sum_{i=1}^{N}\Theta(W_i,x,x',\pi_k)\end{aligned}$$

where

$$\begin{aligned}\Theta(W_i,x,x',\pi_k) :=& Y_i\mathbb{1}(\Pi=\pi_k)\cdot\{\phi(x,\pi_k,T_i)K_h(x-X_i)-\phi(x',\pi_k,T_i)K_h(x'-X_i)\}\\ &-[\chi(\pi_k,x,T_i)K_h(x-X_i)-\chi(\pi_k,x',T_i)K_h(x'-X_i)]+\Gamma(x,x',\pi_k).\end{aligned}$$

Next, we will poissonize $\Gamma_N(x, x', \pi_k)$. To do so, define

$$\Gamma_{\mathcal{N}}(x, \pi_k, \pi_j) = \frac{1}{N} \sum_{i=1}^{\mathcal{N}} \Theta(W_i, x, x', \pi_k)$$

where the empty sum is defined to be zero. Note that by the laws of iterated expectation and variance,

$$\mathbb{E}\Gamma_{\mathcal{N}}(x, x', \pi_k) = \mathbb{E}\Gamma_N(x, x', \pi_k) = \mathbb{E}[\psi(W, x, \pi_k)] - \mathbb{E}[\psi(W, x', \pi_k)], \tag{1.59}$$

$$\kappa_{\tau,N}(x, x', \pi_k) := NVar(\Gamma_{\mathcal{N}}(x, x', \pi_k)) = \mathbb{E}[\Theta^2(W, x, x', \pi_k)] \tag{1.60}$$

and

$$NVar(\Gamma_N(x, x', \pi_k)) = \mathbb{E}[\Theta^2(W, x, x', \pi_k)] - \{\mathbb{E}[\Theta(W, x, x', \pi_k)]\}^2. \tag{1.61}$$

Lets define a poissonized version of T_{2N} (restricted to B)- the uniform asymptotic approximation of $\hat{T}_2$ under the null hypothesis- to be:

$$\begin{aligned} T_{2N}^P(B) := &\int_B \int_B \sum_{k=1}^{K} \left\{ \sqrt{N} \left| [(\Gamma_{\mathcal{N}}(x, x', \pi_k)] - \mathbb{E}[(\Gamma_{\mathcal{N}}(x, x', \pi_k)] \right| \right\} \frac{w(x, x', \pi_k)}{2} dx \\ &- \int_B \int_B \sum_{k=1}^{K} \left\{ \sqrt{N} \mathbb{E} \left| [(\Gamma_{\mathcal{N}}(x, x', \pi_k)] - \mathbb{E}[(\Gamma_{\mathcal{N}}(x, x', \pi_k)] \right| \right\} \frac{w(x, x', \pi_k)}{2} dx. \end{aligned}$$

Lets now prove the asymptotic variance of the poissonized test statistic. If the regularity conditions holds and B satisfies (1.33)-(1.35), then

$$\lim_{N \to \infty} \sigma_{2N}^2(B) = \sigma_{2,B}^2, \tag{1.62}$$

where

$$\sigma_{2,B}^2 := \frac{h^{2d}}{4} \int_B \int_{T_0} \int_{T_0} \int_{T_0} \sum_{k=1}^{K} \mathrm{Cov}\left(\left| \sqrt{1 - \rho_1^*(x, q, r, s, \pi_k)^2} \mathbb{Z}_1 + \rho_1^*(x, q, r, s, \pi_k) \mathbb{Z}_2 \right|, |\mathbb{Z}_2| \right) dr ds dq dx, \tag{1.63}$$

with

$$\rho_1^*(x, q, r, s, \pi) := \frac{\rho_1(x, r, \pi_k) - \rho_1(x, s, \pi_k) - \rho_1(x, q, r, \pi_k) + \rho_1(x, q, s, \pi)}{2\sqrt{(\rho_2(x, \pi_k) - \rho_1(x, q, \pi_k))(\rho_2(x, \pi_k) - \rho_1(x, r, s, \pi_k))}}$$

which reduces to

$$\rho_1^*(x,q,r,s,\pi) := \frac{\int K(\xi+r)K(\xi) - K(\xi+s)K(\xi) - K(\xi+q)K(\xi+r) + K(\xi+q)K(\xi+s)d\xi}{2\cdot\sqrt{(\int K^2(\xi) - K(\xi+q)K(\xi)d\xi)(\int K^2(\xi) - K(\xi+q)K(\xi+s)d\xi)}}$$

since

$$\rho_1(x,r,s,\pi) := \left\{\sum_{t\in\{0,1\}} \frac{\mathbb{E}[Y^2|X=x,T=t,\Pi=\pi] - (\mathbb{E}[Y|X=x,T=t,\Pi=\pi])^2}{P_t(x,\pi)}\right\}\cdot\int K(\xi+s)K(\xi+r)d\xi.$$

Part 1

Note that for each $(x,x',\pi_k),(x'',x''',\pi_k)\in\mathbb{R}^d\times\Pi^2$ such that $\|x-x''\|>h, \|x-x'''\|>h, \|x'-x''\|>h$ and, $\|x'-x'''\|>h$, the random variables $\Gamma_{\mathcal{N}}(x,x',\pi_k) - \mathbb{E}[\Gamma_{\mathcal{N}}(x,x',\pi_k)]$ and $\Gamma_{\mathcal{N}}(x'',x''',\pi_k) - \mathbb{E}[\Gamma_{\mathcal{N}}(x'',x''',\pi_k)]$ are independent because they are functions of independent increments of a Poisson process and the kernel K vanishes outside of the closed rectangle $[-1,1]^d$. Therefore,

$$\begin{aligned}
\mathrm{Var}(T_{2N}^P(B)) =& \frac{1}{4}\int_2 \mathrm{Cov}(\sqrt{N}|\Gamma_{\mathcal{N}}(x,x',\pi_k) - \mathbb{E}[\Gamma_{\mathcal{N}}(x',x,\pi_k)]|,\ \sqrt{N}|\Gamma_{\mathcal{N}}(x'',x''',\pi_k) - \mathbb{E}[\Gamma_{\mathcal{N}}(x'',x''',\pi_k)]|)\\
&\cdot w(x',x,\pi_k)w(x'',x''',\pi_k)dxdx'dx''dx'''\\
=& \frac{1}{4}\int_2 \mathrm{Cov}(\sqrt{N}|\Gamma_{\mathcal{N}}(x,x',\pi_k) - \mathbb{E}[\Gamma_{\mathcal{N}}(x',x,\pi_k)]|,\ \sqrt{N}|\Gamma_{\mathcal{N}}(x'',x''',\pi_k) - \mathbb{E}[\Gamma_{\mathcal{N}}(x'',x''',\pi_k)]|)\\
&\cdot \mathbb{1}(h^{-1}(z-z')\in[-1,1]^d, z\in\{x,x'\}\text{ and }z'\in\{x'',x''''\})w(x',x,\pi_k)w(x'',x''',\pi_k)dxdx'dx''dx''',
\end{aligned}$$

where $\int_2 := \int_B\int_B\int_B\int_B\sum_{k=1}^K$. Furthermore, let

$$S_{\tau,\mathcal{N}}(x,x',\pi) = \frac{\sqrt{N}\{\Gamma_{\mathcal{N}}(x,x',\pi) - \mathbb{E}[\Gamma_{\mathcal{N}}(x,x',\pi)]\}}{\sqrt{\kappa_{\tau,N}(x,x',\pi)}}. \tag{1.64}$$

Now, with $\lambda(\Pi\times B^2)<\infty$, by the Lemma (1.7.4), we can show that for any $\pi\in\Pi$

$$\sup_{(x,x')\in B^2}\left|\sqrt{\kappa_{\tau,N}(x,x',\pi)} - h^{-d/2}\sqrt{2\rho_2(x,\pi) - 2\rho_1(x,q,\pi)}\right| = O(h^{d/2}), \tag{1.65}$$

where

$$\rho_2(x,\pi) = \left\{\sum_{t\in\{0,1\}} \frac{\mathbb{E}[Y^2|X=x,T=t,\Pi=\pi] - (\mathbb{E}[Y|X=x,T=t,\Pi=\pi])^2}{P_t(x,\pi)}\right\}\cdot\int K^2(\xi)d\xi \tag{1.66}$$

and

$$\rho_1(x,q,\pi) = \left\{\sum_{t\in\{0,1\}} \frac{\mathbb{E}[Y^2|X=x,T=t,\Pi=\pi] - (\mathbb{E}[Y|X=x,T=t,\Pi=\pi])^2}{P_t(x,\pi)}\right\}\cdot\int K(\xi)K(\xi+q)d\xi. \tag{1.67}$$

This is because $\mathbb{E}[\Theta^2(W,x,x',\pi)] = \mathbb{E}[\psi^2(W,x,\pi)] + \mathbb{E}[\psi^2(W,x',\pi)] - 2\mathbb{E}[\psi(W,x,\pi)\psi(W,x',\pi)]$ and by change of variable $x' = x + qh$. We also have that

$$\begin{aligned}&\int_1 \mathbb{1}(h^{-1}(z-z') \in [-1,1]^d, z \in \{x,x'\} \text{ and } z' \in \{x'',x''''\}) w(x',x,\pi_k) w(x'',x''',\pi_k) dx dx' dx'' dx''' \\ &= O(h^d). \end{aligned} \tag{1.68}$$

And finally, by the Cauchy Schwartz Inequality, we have

$$\sup_{\{(x,x',\pi_k),(x'',x''',\pi_k)\} \in (B^2 \times \Pi)^2} \left| \text{Cov}(|S_{\tau,\mathcal{N}}(x,x',\pi_k)|, |S_{\tau,\mathcal{N}}(x'',x''',\pi_k)|) \right| = O(1). \tag{1.69}$$

Therefore,

$$\text{Var}(T_{2N}^P(B)) = \bar{\sigma}_{2N}^2 + o(1),$$

where by change of variable $x' = x + qh$, $x'' = x + rh$ and $x''' = x + sh$, we have

$$\begin{aligned}\bar{\sigma}_{2N}^2 :=& \frac{1}{4} \int_2 \text{Cov}(|S_{\tau,\mathcal{N}}(x,x+qh,\pi_k)|, |S_{\tau,\mathcal{N}}(x+rh,x+sh,\pi_k)|) \cdot \mathbb{1}(h^{-1}(z-z') \in [-1,1]^d, z \in \{x,x'\} \text{ and } z' \in \{x'',x'''\}) \\ & h^{-d} \sqrt{[2\rho_2(x,\pi_k) - 2\rho_1(x,q,\pi_k)][2\rho_2(x,\pi_k) - 2\rho_1(x,r,s,\pi_k)]} w(x',x,\pi_k) w(x'',x''',\pi_k) dx dx' dx'' dx'''. \end{aligned} \tag{1.70}$$

Part 2

Now, let $(Z_{1N}(x,x',\pi_k), Z_{2N}(x'',x''',\pi_k))$ for $(x,x',\pi_k), (x'',x''',\pi_k) \in \mathbb{R}^d \times \mathbb{R}^d \times \Pi$, be a mean zero multivariate Gaussian process such that, for each $(x,x',\pi) \in \mathbb{R}^d \times \mathbb{R}^d \times \Pi$, $(Z_{1N}(x,x',\pi_k), Z_{2N}(x'',x''',\pi_k))$ and $(S_{\tau,\mathcal{N}}(x,x',\pi_k), S_{\tau,\mathcal{N}}(x'',x''',\pi_k))$ have the same covariance structure. That is,

$$Z_{1N}(x,x',\pi_k), Z_{2N}(x'',x''',\pi_k)) \overset{d}{=} \left(\sqrt{1 - \rho_N^*(x,x',x'',x''',\pi_k)^2} \mathbb{Z}_1 + \rho_N^*(x,x',x'',x''',\pi_k) \mathbb{Z}_2, \mathbb{Z}_2 \right),$$

where $\mathbb{Z}_1$ and $\mathbb{Z}_2$ are independent standard normal random variables and

$$\rho_N^*(x,x',x'',x''',\pi_k) := \mathbb{E}[S_{\tau,\mathcal{N}}(x,x',\pi_k) S_{\tau,\mathcal{N}}(x'',x''',\pi_k)].$$

Let

$$\begin{aligned}\bar{\tau}_{2N}^2 =& \frac{1}{4} \int_1 \text{Cov}(|Z_{1n}(x,x',\pi_k)|, |Z_{2n}(x'',x''',\pi_k)|) \cdot \mathbb{1}(h^{-1}(z-z') \in [-1,1]^d, z \in \{x,x'\} \text{ and } z' \in \{x'',x'''\}) \\ & h^{-d} \sqrt{[2\rho_2(x,\pi_k) - 2\rho_1(x,q,\pi_k)][2\rho_2(x,\pi_k) - 2\rho_1(x,r,s,\pi_k)]} w(x',x,\pi_k) w(x'',x''',\pi_k) dx dx' dx'' dx'''. \end{aligned}$$

Define

$$\rho_1(x,q_1,q_2,\pi)=\left\{\sum_{t\in\{0,1\}}\frac{\mathbb{E}[Y^2|X=x,T=t,\Pi=\pi]-(\mathbb{E}[Y|X=x,T=t,\Pi=\pi])^2}{P_t(x,\pi)}\right\}\cdot\int K(\xi+q_1)K(\xi+q_2)d\xi \quad (1.71)$$

By a change of variable $x'=x+qh$, $x''=x+rh$ and $x'''=x+sh$, we can show that

$$\rho_N^*(x,x',x'',x''',\pi_k)\to\rho_1^*(x,q,r,s,\pi_k):=\frac{\rho_1(x,r,\pi_k)-\rho_1(x,s,\pi_k)-\rho_1(x,q,r,\pi_k)+\rho_1(x,q,s,\pi)}{\sqrt{4\cdot(\rho_2(x,\pi_k)-\rho_1(x,q,\pi_k))(\rho_2(x,\pi_k)-\rho_1(x,r,s,\pi_k))}}.$$

Using the bounded convergence theorem, we can show that

$$\lim_{N\to\infty}\bar{\tau}_{2N}^2=\sigma_{2,B}^2$$

and, the desired result holds if

$$\bar{\tau}_{2N}^2\to\bar{\sigma}_{2N}^2 \text{ as } N\to\infty. \quad (1.72)$$

To prove (1.72), let $\epsilon_0\in(0,1]$ and let $c(\epsilon_0)=(1+\epsilon_0)^2-1$. Let Q_1 and Q_2 be two independent random variables that are independent of $(\{Y_i,X_i\}_{i=1}^\infty,\mathcal{N})$, each having a two-point distribution that gives two points, $\{\sqrt{c(\epsilon_0)}\}$ and $\{-\sqrt{c(\epsilon_0)}\}$, the equal mass of 1/2, so that $\mathbb{E}[Q_1]=\mathbb{E}[Q_2]=0$ and $\text{Var}(Q_1)=\text{Var}(Q_2)=c(\epsilon_0)$.

Let $S^Q_{\tau,\mathcal{N},1}(x,x',\pi):=\frac{S_{\tau,\mathcal{N}}(x,x',\pi)+Q_1}{1+\epsilon_0}$ and $S^Q_{\tau,\mathcal{N},2}(x,x',\pi):=\frac{S_{\tau,\mathcal{N}}(x,x',\pi)+Q_2}{1+\epsilon_0}$, and define

$$\begin{aligned}\bar{\sigma}^2_{2N,Q}:=\frac{1}{4}\int_2&\text{Cov}(|S^Q_{\tau,\mathcal{N},1}(x,x+qh,\pi_k)|,|S^Q_{\tau,\mathcal{N},2}(x+rh,x+sh,\pi_k)|)\cdot\\&\mathbb{1}(h^{-1}(z-z')\in[-1,1]^d,z\in\{x,x'\}\text{ and }z'\in\{x'',x'''\})\\&h^{-d}\sqrt{[2\rho_2(x,\pi_k)-2\rho_1(x,q,\pi_k)][2\rho_2(x,\pi_k)-2\rho_1(x,r,s,\pi_k)]}\cdot\\&w(x',x,\pi_k)w(x'',x''',\pi_k)dxdx'dx''dx'''.\end{aligned} \quad (1.73)$$

Now, let $Z^Q_{1,N}(x,x',\pi):=\frac{Z_{1,N}(x,x',\pi)+Q_1}{1+\epsilon_0}$ and $Z^Q_{2,N}(x,x',\pi):=\frac{Z_{2,N}(x,x',\pi)+Q_2}{1+\epsilon_0}$.
Then $(Z^Q_{1,N}(x,x',\pi),Z^Q_{2,N}(x'',x''',\pi))$ is a mean zero multivariate Gaussian process such that, for each $(x,x',\pi)\in\mathbb{R}\times\Pi^2$ and $(x'',x''',\pi)\in\mathbb{R}^2\times\Pi$, $(Z^Q_{1,N}(x,x',\pi),Z^Q_{2,N}(x'',x''',\pi))$ and $(S^Q_{\tau,\mathcal{N},1}(x,x',\pi),S^Q_{\tau,\mathcal{N},2}(x'',x''',\pi))$ have the same covariance structure.

Also, define

$$\begin{aligned}\bar{\tau}^2_{2N,Q}=\frac{1}{4}\int_2&\text{Cov}(|Z^Q_{1n}(x,x',\pi_k)|,|Z^Q_{2n}(x'',x''',\pi_k)|)\cdot\mathbb{1}(h^{-1}(z-z')\in[-1,1]^d,z\in\{x,x'\}\text{ and }z'\in\{x'',x'''\})\\&h^{-d}\sqrt{[2\rho_2(x,\pi_k)-2\rho_1(x,q,\pi_k)][2\rho_2(x,\pi_k)-2\rho_1(x,r,s,\pi_k)]}w(x',x,\pi_k)w(x'',x''',\pi_k)dxdx'dx''dx'''.\end{aligned}$$

Note that

$$\begin{aligned}
\left|\bar{\sigma}^2_{2N,Q} - \bar{\tau}^2_{2N,Q}\right| =& \Bigg|\frac{1}{4}\int_B\int_B\int_B\int_B \sum_{k=1}^{K} \mathrm{Cov}(|Z^Q_{1n}(x,x',\pi_k)|, |Z^Q_{2n}(x'',x''',\pi_k)|) - \\
& \mathrm{Cov}(|S^Q_{\tau,\mathcal{N},1}(x,x',\pi_k)|, |S^Q_{\tau,\mathcal{N},2}(x'',x''',\pi_k)|) \\
& \cdot \mathbb{1}(h^{-1}(z-z') \in [-1,1]^d, z \in \{x,x'\} \text{ and } z' \in \{x'',x'''\}) dx dx' dx'' dx''' \Bigg| \\
\leq& \frac{1}{4}\int_B\int_B\int_B\int_B \sum_{k=1}^{K} \Bigg| \mathbb{E}[|Z^Q_{1n}(x,x',\pi_k)|]\mathbb{E}[|Z^Q_{2n}(x'',x''',\pi_k)|] - \\
& \mathbb{E}[|S^Q_{\tau,\mathcal{N},1}(x,x',\pi_k)||S^Q_{\tau,\mathcal{N},2}(x'',x''',\pi_k)|] \Bigg| \\
& \cdot \mathbb{1}(h^{-1}(z-z') \in [-1,1]^d, z \in \{x,x'\} \text{ and } z' \in \{x'',x'''\}) dx dx' dx'' dx''' \\
& + \frac{1}{4}\int_B\int_B\int_B\int_B \sum_{k=1}^{K} \Bigg| \mathbb{E}[|Z^Q_{1n}(x,x',\pi_k)||Z^Q_{2n}(x'',x''',\pi_k)|] - \\
& \mathbb{E}[|S^Q_{\tau,\mathcal{N},1}(x,x',\pi_k)||S^Q_{\tau,\mathcal{N},2}(x'',x''',\pi_k)|] \Bigg| \\
& \cdot \mathbb{1}(h^{-1}(z-z') \in [-1,1]^d, z \in \{x,x'\} \text{ and } z' \in \{x'',x'''\}) dx dx' dx'' dx''' \\
:=& \Delta_{1N,Q} + \Delta_{2N,Q} \qquad (1.74)
\end{aligned}$$

Adapting the proofs in Giné, Mason, and Zaitsev (2003), we have $\Delta_{1N,Q} + \Delta_{2N,Q} = o(1)$ as required.

Lemma 1.7.9 *Under the regularity conditions, for each* $\pi_k \in \mathbf{\Pi}$, $\mathrm{Cov}(\hat{\tau}(x,\pi_k), \hat{\tau}(x',\pi_k)) \to 0$, *as* $N \to \infty$.

proof 1.7.4

$$\begin{aligned}
\mathrm{Cov}(\hat{\tau}(x,\pi_k), \hat{\tau}(x',\pi_k)) :=& \frac{1}{N^2h^{2d}}\mathrm{Cov}\Bigg(\sum_{i=1}^{N} Y_i \mathbb{1}(\Pi_i = \pi_k)\hat{\phi}(T_i,x,\pi_k)K\left(\frac{x-X_i}{h}\right), \\
& \sum_{i=1}^{N} Y_i \mathbb{1}(\Pi_i = \pi)\hat{\phi}(T_i,x',\pi_k)K\left(\frac{x'-X_i}{h}\right)\Bigg) \\
=& \frac{1}{N^2h^{2d}} \sum_{i=1}^{N} \mathrm{Cov}\Bigg(Y_i \mathbb{1}(\Pi_i = \pi_k)\hat{\phi}(T_i,x,\pi_k)K\left(\frac{x-X_i}{h}\right), \\
& Y_i \mathbb{1}(\Pi_i = \pi_k)\hat{\phi}(T_i,x',\pi_k)K\left(\frac{x'-X_i}{h}\right)\Bigg)
\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{Nh^{2d}} \cdot \operatorname{Cov}\Bigg(Y \mathbb{1}(\Pi = \pi_k) \hat{\phi}(T, x, \pi_k) K\left(\frac{x - X}{h}\right), \\
&\quad Y \mathbb{1}(\Pi = \pi_k) \hat{\phi}(T, x', \pi_k) K\left(\frac{x' - X}{h}\right)\Bigg)
\end{aligned} \tag{1.75}
$$

Therefore as $N \to \infty$, $\operatorname{Cov}(\hat{\tau}(x, \pi_k), \hat{\tau}(x', \pi_k)) \to 0$ as required.

Also note that for each $\pi_k \in \mathbf{\Pi}$ and for all $x \in \mathcal{X}$, $\hat{\tau}(x, \pi_k)$ is asymptotically normal. Therefore using Lemma 1.7.9, it is obvious that $\{\hat{\tau}(x, \pi_k)\}_{x \in \mathcal{X}}$ is mutually independent asymptotically. With this fact, we can prove the asymptotic normality of the test statistic $\hat{S}_2$ using similar arguments used to prove the asymptotic normality of the test statistic $\hat{S}_1$.

1.7.5 Proof of Theorem 1.3.2

I prove the first part of Theorem 1.3.2. To save space, I omit the proof of the second part of Theorem 1.3.2 because it is similar to the first part.

$$
\begin{aligned}
\Pr(\hat{S}_1 > z_{1-\alpha}) &= \Pr(\hat{T}_{1N} > a_{1N} + \hat{\sigma} z_{1-\alpha}) \\
&= \Pr(T^*_{1N} > a_{1N} + \hat{\sigma} z_{1-\alpha}) + o(1) \\
&= \Pr(T_{1N} > a_{1N} + \hat{\sigma} z_{1-\alpha}) + o(1) \\
&\to \alpha.
\end{aligned} \tag{1.76}
$$

The second equality holds by Lemma 1.7.2, the third equality holds because $T_{1N} = T^*_{1N}$ under the null hypothesis. The convergence to α follows from Theorems 1.7.1 and 1.7.2.

1.7.6 Proof of Theorem 1.3.3

I prove the first part of Theorem 1.3.3. We can use similar arguments to prove the second part (ii).

$$
\begin{aligned}
\Pr(\hat{S}_1 > z_{1-\alpha}) &= \Pr(\hat{T}_{1N} > a_{1N} + \hat{\sigma} z_{1-\alpha}) \\
&= \Pr\left(\frac{\hat{T}_{1N}}{\sqrt{N}} > \frac{a_{1N} + \hat{\sigma} z_{1-\alpha}}{\sqrt{N}}\right) \\
&= \Pr\left(\frac{\hat{T}_{1N}}{\sqrt{N}} > 0\right) - \Pr\left(0 < \frac{\hat{T}_{1N}}{\sqrt{N}} < \frac{a_{1N} + \hat{\sigma} z_{1-\alpha}}{\sqrt{N}}\right) \\
&= \Pr\left(\frac{\hat{T}_{1N}}{\sqrt{N}} > 0\right) - o(1) \\
&\to 1,
\end{aligned}
$$

where the third equality holds because $(a_{1N} + \hat{\sigma} z_{1-\alpha})/\sqrt{N} = o(1)$ and the last convergence to one follows from the definition of the alternative hypothesis and the fact that $|\hat{\tau}(x,\pi) - \tau(x,\pi)| = o_p(1) \ \ \forall x \in \mathcal{X}, \pi \in \Pi$.

1.7.7 Proof of Theorem 1.3.4

I prove the first part (i) of Theorem 1.3.4. We can use similar arguments to prove the second part (ii). Under H_{1a}, using similar arguments as in Theorem 1.7.2 we show that

$$\frac{T^*_{1N} - \tilde{a}_{N1}}{\sigma_1} \xrightarrow{d} N(0,1), \tag{1.77}$$

where $\tilde{a}_{N1} = 2^{-1} \int_{\mathcal{X}} \sum_{k=1}^{K} \sum_{j=1}^{K} \mathbb{E}[|\mathbb{Z}_1 \cdot h^{-d/2} \sqrt{\rho_2(x,\pi_k,\pi_j)} + h^{-d/4} \delta(x,\pi_k,\pi_j)|] \cdot w(x,\pi_k,\pi_j) dx$. Also, using same arguments as in the proof of Theorem 4.3 in Chang, Lee, and Whang (2015), we can show that

$$\lim_{N\to\infty} \{\tilde{a}_{N1} - a_{N1}\} = \frac{1}{2\sqrt{2\pi}} \int \sum_{k=1}^{K} \sum_{j=1}^{K} \delta^2(x,\pi_k,\pi_j) dx$$

Therefore,

$$\begin{aligned}
\Pr(\hat{S}_1 > z_{1-\alpha}) &= \Pr(\hat{T}_{N1} > \hat{a}_{1N} + \hat{\sigma}_1 z_{1-\alpha}) \\
&= \Pr(T^*_{1N} > \hat{a}_{1N} + \hat{\sigma}_1 z_{1-\alpha}) + o(1) \\
&= \Pr\left(\frac{T^*_{1N} - \tilde{a}_{N1}}{\sigma_1} > \frac{\hat{a}_{1N} - a_{N1}}{\sigma_1} + \frac{\hat{\sigma}_1}{\sigma_1} z_{1-\alpha} - \frac{a_{N1} - \tilde{a}_{N1}}{\sigma_1} \right) \\
&\to 1 - \Phi\left(z_{1-\alpha} - \frac{1}{\sqrt{2\pi}\sigma_1} \int \sum_{k=1}^{K} \sum_{j=1}^{K} \delta^2(x,\pi_k,\pi_j) dx \right)
\end{aligned}$$

Chapter 2

Randomization Inference of Heterogeneous Treatment Effect under Network Interference

2.1 Introduction

The no interference assumption is a common assumption in causality studies, particularly in experiments where individuals are randomly assigned to different treatments or interventions. It assumes that an individual's treatment assignment does not affect the outcomes of other individuals Cox (1958). For example, in a clinical trial, the no interference assumption assumes that an individual's response to the treatment does not depend on whether or not other individuals received the treatment.

However, in reality, this assumption may not always hold. Modern society is inextricably interconnected through social networks or other forms of interactions; interference can occur when the treatment or intervention of one individual affects the outcomes of other individuals via social connections. For example, if a treatment involves a group intervention, such as a community health program, individuals will likely interact with each other, and the effects of the treatment may spread beyond the immediate target population. In such cases, relaxing the no interference assumption can provide a more accurate representation of the actual outcomes and help researchers better estimate and infer the effects of the intervention.

Motivated by this fact, over the past decade, there has been increasing attention paid to the incorporation of interference into standard models of causal inference, especially when analyzing network data sets. For instance, Liu and Hudgens (2014) develop a framework for causal inference that accounts for interference within groups. However, this can be an arduous task in complex networks, and researchers often need to make simplifying assumptions about the interference structure in causality studies.

One such simplifying assumption is the *exposure mapping* construct of Aronow et al. (2017), which summarizes the impacts of other individuals' treatments into lower dimensional sufficient statistics. This may reduce the number of missing potential outcomes and makes it possible to estimate causal effects in the presence of network interference. For example, Leung (2020) studies the estimation of treatment effects in network populations by assuming that the fraction of treated neighbors is the appropriate sufficient statistic of how the treatments of neighbors affect one's outcome. Therefore, an individual's treatment and neighborhood treatment ratio are the causal variables of interest. Different specifications of the exposure mapping may require different assumptions about the nature of the network interference (see Manski (2013)).

In this chapter, we add to the growing stock of research papers that allow network interference into the potential outcomes model. Our distinct goal is to provide statistical methods to credibly infer heterogeneous treatment effects (HTEs) using experimental network data sets. The knowledge of HTEs is useful in the design of welfare-maximizing policies as it allows for the targeting of specific subgroups that would benefit the most from a particular intervention. For instance, a study

by Viviano (2019) demonstrates how to use HTE estimates to improve a weather insurance policy take-up among rice farmers in rural China. Similarly, Han et al. (2022) use HTE estimates to design multinomial success rules in populations where interactions occur within non-overlapping groups. To achieve our goal, we develop randomization testing methods that are valid asymptotically for three useful notions of HTEs in the presence of network interference: (i) the null hypothesis of constant treatment effects across the population; (ii) heterogeneous treatment effects across network exposure values only; and (iii) heterogeneous treatment effects across network exposure values and covariate-defined discrete groups only.

Our reliance on a randomization-based testing method for HTEs in the current chapter is motivated by two main reasons. First, since units are linked through social networks, we cannot assume that the variables of units are independent. Therefore, the traditional asymptotic-based inferential methods are not directly applicable. The second advantage of the randomization-based testing method is that it provides exact p-values without imposing restrictive conditions on the data generating process (DGP) for *sharp* null hypotheses.[1] Moreover, recent studies show that even for non-sharp null hypotheses, there are conditional randomization-based methods that can generate exact conditional p-values without any assumptions on the DGP.

The null hypotheses we consider in this chapter are *not* sharp due to two reasons. First, our null hypotheses contain nuisance parameters which are unknown values in the science tables one will construct under the null. Hence, one can only partially impute the potential outcomes that depend on these nuisance parameters under the null hypotheses. The problem of nuisance parameters in null hypotheses is not exclusive to network interference situations, but there may be more of these parameters present in this setting. (See Ding et al. (2016) for a hypothesis of constant treatment effects under no interference). Second, under network interference, the number of potential outcomes depends on the exposure mapping one imposes. Therefore, without additional restrictions on the underlying DGP, we cannot impute all missing potential outcomes under the null hypotheses. As a result of these two reasons, the conventional randomization method of inference by Fisher (1925) is not directly applicable to the null hypotheses we test in the current chapter without appropriate modifications.

In the following sections, we thoroughly discuss our proposed approaches to "sharpen" the null hypotheses. Here, we provide a summary. First, we propose two methods to deal with the presence of the nuisance parameters. Additionally, we offer a conditional randomization method to handle the issue of multiple potential outcomes. The idea behind the conditioning method is that by focusing on a subset of treatment assignment vectors and a subset of units, a non-sharp null hypothesis becomes

[1] Under a sharp null hypothesis, all potential outcomes for each unit can be imputed.

sharp. In Section 2.3.2, we show how the conditioning method in this chapter differs from those that exist in the literature, particularly those of Athey et al. (2018) and Basse et al. (2019).

This chapter makes three main contributions. First, we introduce three new hypotheses of constant treatment effects under network interference. By testing these null hypotheses individually, researchers can infer HTEs in populations with network interferences. However, jointly testing all three null hypotheses allows researchers to determine if treatment effect variations are driven by pretreatment variables, network exposure variables, or are idiosyncratic. It can help policy-makers better understand the underlying mechanisms driving heterogeneity in treatment effects.

Second, we propose a novel conditional randomization method that provides a reliable testing procedure for our null hypotheses. It is a generalization of existing methods in the literature. We show that the proposed method produces valid p-values in the limit.

Finally, we propose techniques for handling nuisance parameters in the null hypotheses. Specifically, we offer practitioners two effective techniques and show that they produce valid p-values for large samples when combined with our conditional randomization method. In other words, our proposed conditioning method and the techniques to deal with the nuisance parameters make the null hypotheses sharp and ensure that the randomization hypothesis in Lehmann and Romano (2022) holds.

The organization of the rest of the chapter is as follows. We review existing related work in the second part of this section. Section 2.2 describes the setup and the hypothesis testing problem. In Section 2.3, we discuss the proposed testing procedure and our main results. Monte Carlo simulation design and results are in Section 2.4. Our concluding remarks are in Section 2.5. All proofs, useful theorems, and lemmas are in the Appendix.

2.1.1 Related Literature

The study of HTEs spans multiple fields and is often under the assumption of no interference. Many existing papers focus on systematic HTEs explained by pretreatment variables (Crump et al. (2008), Wager and Athey (2018), and Sant'Anna (2021)). In an influential paper, Bitler et al. (2006) provide an in-depth critique of this approach to testing for HTEs. They argue that heterogeneity of conditional average treatments across covariate-defined subgroups often does not imply individual treatment effect variation unless one assumes constant subgroup treatment effect. Inspired by the results in Bitler et al. (2006), Ding et al. (2016) study randomization inference for HTEs beyond that which can be accounted for by pretreatment variables. They use a method in Berger and Boos (1994) to deal with the nuisance parameters in their null hypothesis. They prove the validity of their maximum p-value approach and acknowledge that it leads to the under-rejection of the null

hypothesis. Similarly, more recently, Chung and Olivares (2021) propose a permutation test for HTEs using the Doob-Meyer theorem (martingale transformation) to handle the nuisance parameters in the null hypothesis. They show that their testing procedure is asymptotically valid as the Khmaladzation of the empirical process that the null hypothesis characterizes will render the estimation errors of the nuisance parameters asymptotically negligible.

Fisher's method of randomization inference (Fisher (1925)) proposes and tests for the sharp null hypothesis of zero treatment effect. While this method is innovative because it abstracts from distributional assumptions and shows that physical randomization of treatment is the logical basis of inference, many researchers criticize it for its limited scope of application. Motivated by this criticism, Athey et al. (2018) propose an abstract concept of an artificial experiment that differs from the original experiment to make a non-sharp null hypothesis sharp. In econometrics and statistics, this method is now widely known as the conditional randomization method of inference. Several subsequent studies, such as Basse et al. (2019) and Zhang and Zhao (2022), aim to generalize the framework to construct conditioning events to "sharpen" non-sharp null hypotheses. These studies investigate the calculations of exact p-values for a large class of null hypotheses about treatment effects in settings where we have experimental network data. However, the null hypotheses we study in the present chapter are different and contain nuisance parameters.

Among all the aforementioned references, the current work is most closely related to Ding et al. (2016) and Chung and Olivares (2021), but there are some differences. First, we allow for network interference which introduces dependencies among observations. In other words, the treated sample is not independent of the control sample. This means that the unconditional asymptotic null distribution of test statistics like the Kolmogorov-Smirnov, Cramer-Von-Mises, and ratio of variances (F-test) is not trivial. Second, we provide a new method to handle the nuisance parameters. Finally, due to multiple potential outcomes in the current framework, our null hypotheses are non-sharp even if the nuisance parameters are known by the econometrician. Thus, in contrast to the unconditional randomization procedure the authors employ in these papers, we propose a new conditional randomization approach that "sharpens" our non-sharp hypotheses. Our method proves to be effective in generating valid p-values in the limit.

2.2 Framework

2.2.1 Setup

Following Athey et al. (2018), we consider the following framework. Suppose we have a population of N units (with i indexing the units) connected through a single network[2] that we denote by a symmetric $N \times N$ adjacency matrix $\mathbf{A}$. The ijth element of the adjacency matrix, A_{ij}, equals one if units i and j interact, and zero otherwise. Henceforth, we refer to units i and j as neighbors if $A_{ij} = 1$. We assume there are no self-loops, i.e., $A_{ii} = 0$ for all i.

Also, we assume that an experimenter randomly assigns each unit i to a binary treatment $T_i \in \{0, 1\}$. Therefore, we have a vector $\mathbf{T}$ which denotes the N-component vector of treatments. The experimenter assigns the treatments using the treatment assignment mechanism $p : \{0, 1\}^N \mapsto [0, 1]$, where $p(\mathbf{t})$ is the probability of $\mathbf{T} = \mathbf{t}$, with $p(\mathbf{t}) \geq 0$ and $\sum_{\mathbf{t}\in\{0,1\}^N} p(\mathbf{t}) = 1$. In addition, we let $\mathbf{Y} : \{0, 1\}^N \mapsto \mathbb{Y}^N$ represent the mapping of potential outcomes, where the i^{th} element of $\mathbf{Y}$ is $Y_i : \{0, 1\}^N \mapsto \mathbb{Y} \subset \mathbb{R}$. Thus, $Y_i(\mathbf{t})$ denotes the potential outcome for unit i if $\mathbf{T} = \mathbf{t}$, and the i^{th} element of $\mathbf{Y}(\mathbf{t})$. If the observed value of the treatment assignment vector is $\mathbf{t}^{obs}$, then, we encode the N-component vector of observed outcomes as $\mathbf{Y}^{obs} = \mathbf{Y}(\mathbf{t}^{obs})$ with the i^{th} element $Y_i^{obs} = Y_i(\mathbf{t}^{obs})$, which represents the realized outcome of unit i. Alternatively, we can characterize the observed outcomes as a function of the vector of treatments $\mathbf{T}$ and all the vector of potential outcomes $\mathbf{Y}(\mathbf{t})$, with $p(\mathbf{t}) > 0$.

Furthermore, for each unit, there is an L dimensional vector of pretreatment variables $X_i \in \mathbb{X} \subset \mathbb{R}^L$, with the $N \times L$ matrix of pretreatment variables denoted by $\mathbf{X}$. Note that X_i includes the local network characteristics of unit i. Therefore, $(\mathbf{Y}^{obs}, \mathbf{T}, \mathbf{A}, \mathbf{X})$ is the data available to the researcher. The current chapter assumes a design-based approach where $\mathbf{A}$ and $\mathbf{X}$, as well as the potential outcome mapping $\mathbf{Y}(\cdot)$ are fixed, but $\mathbf{Y}^{obs}$ is random due to the randomness of the treatment assignment.

2.2.2 Network Exposure Mapping

In this subsection, we introduce the concept of network exposure mapping and its usefulness in incorporating network interference into the potential outcomes framework. Failure to account for network interference in causal analyses may lead to misleading statistical results and economic conclusions. However, allowing general network interference aggravates the missing data problem of causal inference and could make the potential outcomes model intractable. As a result, a salient element of causal inference in the presence of network interference is a *network exposure mapping*

[2]The network is exogenous, i.e., a fixed characteristic of the population and units are not strategically interacting.

(Aronow et al. (2017)), which imposes testable restrictions on the nature of interactions. It is related to the concept of level sets in Athey et al. (2018). Formally, we define our network exposure mapping as

$$\pi : \{1, \ldots, N\} \times \{0, 1\}^N \mapsto \mathbf{\Pi}, \tag{2.1}$$

that maps $(i, \mathbf{T})$ into $\Pi \in \mathbf{\Pi} \subset \mathbb{R}$, where $\mathbf{\Pi}$ is an arbitrary set of possible treatment exposure values. We assume that the functional form of π is arbitrary but known to the econometrician. In addition, we suppose that the exposure mapping is the same for all units. However, for notational simplicity, we let $\pi(i, \mathbf{T}) = \pi_i(\mathbf{T})$. This should not be misconstrued for variations in exposure mapping across units.

Given a network exposure mapping, a reasonable assumption that generalizes the standard stable unit treatment value assumption (SUTVA) and imposes restrictions on the nature of interactions is that, for any two treatment vectors $\mathbf{t} \neq \mathbf{t}'$ where $\mathbf{t} = (t_i, \mathbf{t}_{-i}) \in \{0, 1\}^N$ and $\mathbf{t}' = (t_i, \mathbf{t}'_{-i}) \in \{0, 1\}^N$, $Y_i(\mathbf{t}) = Y_i(\mathbf{t}')$ if $\pi_i(\mathbf{t}) = \pi_i(\mathbf{t}')$. This assumption states that potential outcomes depend on treatment and network exposure value. Therefore, borrowing terminology from Manski (2013), the tuples $(T, \Pi) \in \{0, 1\} \times \mathbf{\Pi}$ represent the *effective treatments*. Thus, $Y_i(\mathbf{t}) = Y_i(t, \pi)$ is the potential outcome for unit i if the treatment vector $\mathbf{T} = \mathbf{t}$ is such that $T_i = t$ and $\pi_i(\mathbf{t}) = \pi$. For example, if we define network exposure mapping as the fraction of treated neighbors, then the potential outcomes depend on the treatment assigned to a unit and the fraction of treated neighbors.

Let us formalize the assumptions that describe the network and the nature of interactions.

Assumption 2.2.1 (No Second and Higher-Order Spillovers) *Let $\mathbb{M}(i, j)$ be length of the shortest path between units i and j where, $\mathbb{M}(i, j) = \infty$ if there is no path between i and j. Given the definition of $\mathbb{M}(\cdot, \cdot)$, $Y_i(\mathbf{t}') = Y_i(\mathbf{t})$ for all i, and for all pairs of assignment vectors $\mathbf{t}, \mathbf{t}' \in \{0, 1\}^N$ if $t_j = t'_j$ for all units j where $\mathbb{M}(i, j) < 2$.*

Assumption 2.2.2 (Uniformly Bounded Degrees) *For each unit i, $\exists\ M < \infty$ such that*

$$\lim_{N \to \infty} \sum_{j=1}^{N} A_{ij} \leq M \tag{2.2}$$

Assumption 2.2.1 is a testable restriction that permits spillover effects of the first order but no higher-order spillovers. In other words, altering the treatment of direct neighbors may change one's outcome, but altering the treatment of neighbors-of-neighbors does not affect one's outcome. It is a convenient and testable restriction (Athey et al. (2018)) that ensures sparsity of the network. Note that the testing procedures we propose are still valid if we impose other less restrictive network sparsity conditions like no third order or more spillovers in Leung (2020).

Assumption 2.2.2 is a crucial assumption that imposes a uniform upper bound on the asymptotic degree of each node (the number of connections of each unit). It ensures that the number of treatment exposure values converges to a finite number as the population size increases. For instance, if the treatment exposure variable is the fraction of treated neighbors and $M = 3$, then this assumption ensures that the set of exposure values converges to $\mathbf{\Pi} = \{0, 1/3, 2/3, 1\}$. For our asymptotic results, Assumption 2.2.2 ensures that the number of potential outcomes under consideration does not vary with the population size. We acknowledge that for some network exposure mappings, an asymptotic bound on the average degree of nodes - which is less restrictive than the condition in (2.2) - is plausible. However, note that this assumption is the rule in most randomized experiments with networks rather than the exception. For instance, in a randomized experiment on rice-producing households, Cai et al. (2015) restricts the number of neighboring households to five. We highlight in Section 2.3 that for improved statistical power, a small M is desirable because it ensures that we have "sufficient" units with the different values of exposure, i.e., $\Pr(\Pi = \pi) > 0$ for all $\pi \in \mathbf{\Pi}$.

2.2.3 The Hypothesis Testing Problem

Let us formally describe the testing problem and clearly outline the disparities between our hypotheses and the conventional hypothesis testing problem that assumes no interference. Using an arbitrary network exposure mapping, we consider three non-sharp null hypotheses that characterize the notions of homogeneous treatment effects under network interference. First, we have the null hypothesis

$$H_0 : Y_i(1, \pi) - Y_i(0, \pi) = \tau \text{ for some } \tau, \ \forall \ \pi \in \mathbf{\Pi}, \text{ for } i = 1, \ldots, N. \tag{2.3}$$

It is the null hypothesis of *constant treatment effect across the population*. In words, this hypothesis asserts that there are *no forms of variations* in treatment effects, i.e., systematic and idiosyncratic variations in treatment effects are absent. It is the strongest form of constant treatment effects under network interference. Note that testing H_0 is salient in the design of treatment assignment rules that aim to maximize welfare. For instance, if we fail to reject this null hypothesis, then the welfare maximizing rule does not depend on the exposure variable under consideration. Under the *no interference assumption*, Ding et al. (2016) and Chung and Olivares (2021) respectively design a randomization and permutation test for an analogous hypothesis which is not sharp due to the presence of an unknown nuisance parameter. In contrast, our null hypothesis (2.3) is non-sharp not only because τ is a nuisance parameter, but also due to the multiplicity of potential outcomes induced by network interference.

Second, we consider the null hypothesis

$$H_0^{\Pi} : Y_i(1,\pi) - Y_i(0,\pi) = \tau(\pi) \text{ for some } \tau(\cdot), \ \ \forall\ \pi \in \mathbf{\Pi}, \ \text{for } i = 1,\ldots,N. \tag{2.4}$$

In other words, this null hypothesis asserts that *treatment effects may only vary systematically across treatment exposure values*. In simpler terms, the treatment effects for units with the same exposure value are constant, but variations exist across exposure values. If one rejects H_0^{Π}, we may target treatment using the exposure variable. It is important to note that H_0^{Π} is different from the null hypothesis of no interference in Aronow (2012) and Athey et al. (2018), which is a restriction on the treatment response or potential outcomes functions, whereas, H_0^{Π} is a restriction on treatment effects. Moreover, there may be interference, yet treatment effects are constant across exposure values.

Finally, we consider the null hypothesis

$$H_0^{X,\Pi} : Y_i(1,\pi; x) - Y_i(0,\pi; x) = \tau(\pi; x) \ \ \forall\ x \in \mathbb{X}, \ \forall\ \pi \in \mathbf{\Pi}, \ \text{for } i = 1,\ldots,N. \tag{2.5}$$

Hypothesis $H_0^{X,\Pi}$ implies that treatment effects *may only vary systematically across pretreatment and exposure values.* In simpler terms, the treatment effects for units with the same pretreatment and exposure value are constant. Here, we assume that the L pretreatment variables are either discrete or can define an L-number of non-overlapping subgroups. Specifically, we suppose that the pretreatment variables $\mathbf{X}$ can define L discrete subgroups. Hence, with a slight abuse of notations, we have $\mathbb{X} = \{x_1 \ldots, x_L\}$. Thus, the pretreatment and exposure variables create an $L \times K$-number of non-overlapping subgroups. This hypothesis is also crucial in designing covariate-dependent eligibility rules to maximize social welfare in practical settings where resources are scarce.

Since one cannot observe the individual level treatment effects $Y_i(1,\pi) - Y_i(0,\pi)$ for $\pi \in \mathbf{\Pi}$, we rewrite the null hypotheses above as statements about the distributions of the treatment and control groups for each exposure value. Let $F_{1\pi}$ and $F_{0\pi}$ denote the distributions of $Y(1,\pi)$ and $Y(0,\pi)$ respectively for $\pi \in \mathbf{\Pi}$. Then, for the null hypotheses (2.3)-(2.5), there is a corresponding testable statement about the functional of distributions. For instance, $H_0 : Y_i(1,\pi) = Y_i(0,\pi) + \tau, \ \forall \pi \in \mathbf{\Pi}$ implies that $H_0^{CDF} : F_{1\pi}(y) = F_{0\pi}(y - \tau) \ \ \forall \pi \in \mathbf{\Pi}, \ \forall y \in \mathbb{Y}$. In addition, equal variances of $Y(1,\pi)$ and $Y(0,\pi)$, for $\pi \in \mathbf{\Pi}$ is a testable implication of H_0. Also, note that the unequal variances is evidence against H_0, although the converse may not be true (see Ding et al. (2016) for other testable implications).

To complete the description of the testing problem, it is worthwhile to compare our null hypothesis of constant treatment effects across the population in (2.3) to the hypothesis of constant treatment effect under the SUTVA assumption in Chung and Olivares (2021) and Ding et al. (2016).

In other words, does the rejection (non-rejection) of the null hypothesis of constant treatment effect under no interference implies a rejection (non-rejection) of the null hypothesis of constant treatment effect under network interference? A testable implication of the hypothesis of constant treatment effects under no interference is $H_0^0 : F_1(y) = F_0(y - \tau), \forall y \in \mathbb{Y}$ for some τ, where F_1 and F_0 denote the distributions of the traditional potential outcomes $Y(1)$ and $Y(0)$ respectively. Despite the close resemblance between H_0^0 and H_0^{CDF}, they are not equivalent because $(Y(1), Y(0))$ and $(Y(1,\pi), Y(0,\pi), \pi \in \mathbf{\Pi})$ are different vectors.

2.3 The Testing Procedure: Randomization Inference

We study reliable randomization testing procedures for the null hypotheses. In particular, we investigate the three elements of randomization inference (RI): (i) a randomized treatment assignment mechanism, (ii) a test statistic, and (iii) a sharp null hypothesis. We assume that the treatment assignment mechanism is known, and there exist data from a completely randomized experiment.[3] We acknowledge that a completely randomized experiment design introduces dependencies between treatment assigned to units. However, under Assumption 2.2.2, such dependencies occur with a sufficiently low probability (Sävje et al. (2021)). In Subsections 2.3.1-2.3.2, we discuss the other two ingredients of RI in relation to the null hypotheses (2.3)-(2.5).

2.3.1 Test Statistics

The choice of the test statistic has implications on the statistical power of the RI for the null hypotheses (2.3)-(2.5). We propose test statistics that are functionals of estimated conditional variances of outcomes of the treatment and control groups for each exposure value. For all three null hypotheses, we look at two testing approaches: a multiple testing approach with many test statistics and a single testing approach with one combined test statistic.

Let $|\mathbf{\Pi}| = K$, where $|\cdot|$ denotes set cardinality. Focusing on the null hypotheses (2.3) and (2.4), we propose K test statistics that corresponds to each $\pi \in \mathbf{\Pi}$, i.e.,

$$\text{TS}_k(\mathbf{Y}^{obs}, \mathbf{T}, \mathbf{A}) := \max\left\{\frac{\hat{\sigma}_1^2(\pi_k)}{\hat{\sigma}_0^2(\pi_k)}, \frac{\hat{\sigma}_0^2(\pi_k)}{\hat{\sigma}_1^2(\pi_k)}\right\}, \text{ for } k = \{1, \ldots, K\}, \tag{2.6}$$

where $\hat{\sigma}_t^2(\pi_k)$ is an estimate of the conditional variance of observed outcome for units with treatment

[3]This is a convenient assumption. We can extend the results to other treatment assignments regimes like Bernoulli assignments, stratified assignments, or the two-stage random saturation design of Baird et al. (2018).

t and exposure value π_k. Note that if a researcher knows that $\sigma_t^2(\pi_k) > \sigma_{t'}^2(\pi_k)$ for $t \neq t' \in \{0, 1\}$, for all $\pi_k \in \mathbf{\Pi}$ then a simpler test statistic $\hat{\sigma}_1^2(\pi_k)/\hat{\sigma}_0^2(\pi_k)$ for all $\pi_k \in \mathbf{\Pi}$ is attractive. To test (2.3) and (2.4) using these test statistics, note that one can apply any multiple testing procedures that give a decision for all $\pi_k \in \mathbf{\Pi}$. Alternatively, we can combine all the K test statistics into one using a scalar-valued function, i.e.,

$$\text{TS}(\mathbf{Y}^{obs}, \mathbf{T}, \mathbf{A}) := f(\text{TS}_1, \ldots, \text{TS}_k, \ldots, \text{TS}_K), \tag{2.7}$$

where $f(\cdot, \ldots, \cdot)$ is some known scalar-valued function.

For null hypothesis (2.5), we propose test statistics that correspond to each subgroup defined by pretreatment variables and the exposure variable, i.e.,

$$\text{TS}_{k,l}(\mathbf{Y}^{obs}, \mathbf{T}, \mathbf{A}, \mathbf{X}) := \max\left\{\frac{\hat{\sigma}_1^2(\pi_k, x_l)}{\hat{\sigma}_0^2(\pi_k, x_l)}, \frac{\hat{\sigma}_0^2(\pi_k, x_l)}{\hat{\sigma}_1^2(\pi_k, x_l)}\right\}, k = \{1, \ldots, K\}, l = \{1, \ldots, L\}, \tag{2.8}$$

where $\hat{\sigma}_t^2(\pi_k, x_l)$ is an estimate of the conditional variance of observed outcome for units with treatment t, exposure value π_k and pretreatment value (or covariate defined subgroup) x_l. Alternatively, we can combine all the test statistics into one using a scalar-valued function, i.e.,

$$\text{TS}^{X,\Pi}(\mathbf{Y}^{obs}, \mathbf{T}, \mathbf{A}, \mathbf{X}) := f(\text{TS}_{k,l} : x_l \in \mathbb{X}, k = 1, \ldots, K). \tag{2.9}$$

For the two combined test statistics, we propose an equally-weighted average of the individual test statistics. In other words, $f(\cdot)$ is the summation across individual test statistics.

2.3.2 Sharp Null Hypothesis and Conditional Randomization Inference

Let us examine the third component of randomization inference, sharp null hypotheses, and how it connects to our proposed null hypotheses. As we mentioned in the introduction, for a given treatment assignment mechanism p, our null hypotheses are *not* sharp because of the presence of nuisance parameters and the multiplicity of potential outcomes. Therefore, we cannot impute all potential outcomes for each unit under the null hypotheses (2.3)-(2.5). For the rest of this subsection, we assume that the nuisance parameters are *known*[4], and propose a solution to deal with the "non-sharpness" stemming from the multiplicity of potential outcomes. We defer the proposed remedies to the nuisance parameter(s) problem to Subsection 2.3.3.

[4]For instance, when the nuisance parameters are set to zero, the null hypotheses impose no treatment effects or constant treatment effects of zero.

Suppose the nuisance parameters τ, $\tau(\pi)$ and $\tau(\pi; x)$ in the null hypotheses (2.3)-(2.5) are known. It is evident that unlike the null hypothesis H_0^0 under no interference in Ding et al. (2016) and Chung and Olivares (2021), (2.3)-(2.5) remain *non-sharp* due to multiple potential outcomes induced by network interference.

We propose a novel conditional randomization inference (CRI) method, where we use subsets of units in the population (often referred to as *focal units* (Athey et al. (2018))) and subsets of treatment assignment vectors to estimate the conditional null distributions. See Zhang and Zhao (2022) for an overview of the literature on CRI. We proceed to describe our proposed CRI method for the three null hypotheses. For each of the hypotheses, we describe the CRI method given (a) the multiple test statistics and (b) the combined test statistic.

2.3.2.1 H_0: Constant Treatment effects across the population

We begin the description of the "non-sharpness" of H_0 and its CRI method using the test statistics $\text{TS}_k, k = 1, \dots K$ in (2.6). To distinguish between the treatments assigned by the experimenter and permuted treatment vectors, we introduce additional notation. Let $\Lambda(\mathbf{T})$ denote a permutation of $\mathbf{T}$ that satisfies the treatment assignment mechanism. Thus, $\Lambda(T_i)$ is the treatment of unit i under a permutation of assigned treatment.

Under H_0 with τ known, H_0 is not sharp, and each $\text{TS}_k, k = 1, \dots, K$ is not imputable.[5] The direct implication of the non-imputability is that, under H_0, there exists a treatment vector $\Lambda(\mathbf{T}) = \mathbf{t}' \in \{0, 1\}^N$ where $p(\mathbf{t}') > 0$, but $\sum_{i=1}^N \mathbb{I}(\Lambda(T_i) = t, \pi_i(\mathbf{t}') = \pi) = 0$ for either $t = 0$ or $t = 1$, and $\pi \in \mathbf{\Pi}$. Here, $\mathbb{I}(\cdot)$ is the indicator function. The following example illustrates the "non-sharpness" of H_0.

Example 2.3.1 *Assume there are* $N = 10$ *units in an undirected social network, shown in Figure 2.1. Given the data* $(Y_i^{obs}, T_i, \Pi_i)_{i=1}^N$, *where* $\Pi_i = \pi_i(\mathbf{T}) := \mathbb{I}(\sum_{j=1}^N T_j A_{ij} / \sum_{j=1}^N A_{ij} \geq 0.5) \in \{0, 1\}$, *Table 2.1 shows that under* H_0 *in* (2.3)*, each unit has two missing potential outcomes which we represent by question marks. For example, the potential outcomes* $Y_1(1, 0)$ *and* $Y_1(0, 0)$ *are missing for unit 1. Thus,* H_0 *is not sharp. Now, let us consider* $\Lambda(\mathbf{T}) = \tilde{\mathbf{t}} = (1, 1, 1, 1, 1, 0, 0, 0, 0, 0)$, *a permutation of the observed treatment values. Note that,* $\tilde{\mathbf{t}}$ *produces a new vector of exposure values* $(1, 1, 1, 1, 1, 1, 0, 0, 0, 0)$. *Also, note that the test statistic* TS_0 *(i.e.,* TS_k *with* $\pi = 0$*) is not imputable under* H_0 *since there are no units with treatment equal to* 1 *and exposure value of* 0 *i.e.,* $\sum_{i=1}^N \mathbb{I}(\Lambda(T_i) = 1, \pi_i(\tilde{\mathbf{t}}) = 0) = 0$.

[5] A test statistic is imputable if, under the null hypothesis, we can compute a value of the statistic for all treatment vectors that is possible under a given assignment mechanism. See Basse et al. (2019) for a formal technical definition.

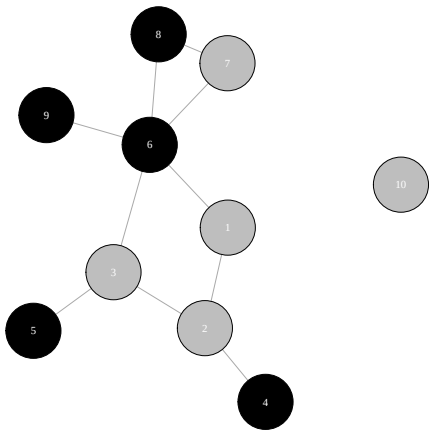


Figure 2.1: Undirected Social Network. (Note: Grey nodes are the control units and black nodes are treatment units)

Units	Observed Variables			Counterfactual Outcomes				Permuted Variables		
i	T_i	Π_i	Y_i^{obs}	$Y_i(1,0)$	$Y_i(0,0)$	$Y_i(1,1)$	$Y_i(0,1)$	$\Lambda(T_i)$	$\pi_i(\tilde{\mathbf{t}})$	Y_i^P
1	0	1	y_1	?	?	$y_1+\tau$	y_1	1	1	$y_1+\tau$
2	0	0	y_2	$y_2+\tau$	y_2	?	?	1	1	?
3	0	1	y_3	?	?	$y_3+\tau$	y_3	1	1	$y_3+\tau$
4	1	0	y_4	y_4	$y_4-\tau$	?	?	1	1	?
5	1	0	y_5	y_5	$y_5-\tau$	?	?	1	1	?
6	1	0	y_6	y_6	$y_6-\tau$	?	?	0	1	?
7	0	1	y_7	?	?	$y_7+\tau$	y_7	0	0	?
8	1	1	y_8	?	?	y_8	$y_8-\tau$	0	0	?
9	1	1	y_9	?	?	y_9	$y_9-\tau$	0	0	?
10	0	0	y_{10}	$y_{10}+\tau$	y_{10}	?	?	0	0	y_{10}

Table 2.1: A Science Table under H_0, using Example 2.3.1. NB: Y_i^P represents the new outcome of unit i under H_0 for the new treatment vector $\tilde{\mathbf{t}}$.

To "sharpen" H_0 and obtain the null distributions of a test statistic TS_k, for any k, we propose a RI method that requires conditioning on a subset of treatment vectors and a subset of experimental

units. First, let $\mathcal{T}_k$ represent the subset of treatment assignment vectors that imputes TS_k. We define it as

$$\mathcal{T}_k := \{\mathbf{t}' \in \{0,1\}^N : p(\mathbf{t}') > 0 \text{ and } R(t, \mathbf{t}', \pi_k) > \epsilon, t = 0, 1, \text{ for some } \epsilon \in [0, 1)\}, \tag{2.10}$$

with,

$$R(t, \mathbf{t}', \pi_k) := \frac{\sum_{i=1}^N \mathbb{I}\{t_i' = t, \pi_i(\mathbf{t}') = \pi_k, \pi_i(\mathbf{t}^{obs}) = \pi_k\}}{\sum_{i=1}^N \mathbb{I}\{\pi_i(\mathbf{t}^{obs}) = \pi_k\}}, \tag{2.11}$$

where, $R(\cdot, \cdot, \cdot)$ denotes the empirical probability or relative frequency. In other words, $\mathcal{T}_k$ is the subset of treatment assignment vectors satisfying the treatment assignment mechanism p, and also ensures that a "sufficient" number of units with the different treatments ($t = 0, 1$) have exposure values *fixed* at π_k. Note that ϵ controls the minimum number of units with an exposure value of π_k under the permuted treatment vector. Therefore, ϵ is a tuning parameter that affects the computation time and statistical power of our CRI method. Higher values of ϵ make our test conservative and increases computation time. If we let $\{0, 1\} = \{\pi_0, \pi_1\}$ in Example 2.3.1, then $\tilde{\mathbf{t}} \notin \mathcal{T}_0$.

By construction, the sets $\mathcal{T}_k$, for $k = 1, \ldots K$, depend on the sample size and the exposure mapping. For each k, the cardinality of $\mathcal{T}_k$ is more likely to be larger when the sample size is large. Also, exposure mappings with a smaller range of values (e.g., threshold functions of the treatment vectors of neighbors that produce binary exposure values as in Example 2.3.1) are more likely to produce larger $\mathcal{T}_k$ sets. The power and size distortions of the test depend on $|\mathcal{T}_k|$.

Second, let $F_k(\mathbf{t})$ denote the indicator variable for the focal units we use to compute TS_k when the treatment assignment vector is $\mathbf{t}$. Therefore, for unit i, we define

$$F_{ik}(\mathbf{t}') := \mathbb{I}(\pi_i(\mathbf{t}') = \pi_k, \pi_i(\mathbf{t}^{obs}) = \pi_k), \quad \forall \mathbf{t}' \in \mathcal{T}_k. \tag{2.12}$$

In other words, the focal units (for a given treatment vector that belongs to the subset of treatment assignment vectors) are the units whose exposure values *remain the same* as the exposure values under the observed treatment assignment vector. By construction, for $\mathbf{t}', \mathbf{t}'' \in \mathcal{T}_k$, where $\mathbf{t}' \neq \mathbf{t}''$, it is possible that $F_{ik}(\mathbf{t}') \neq F_{ik}(\mathbf{t}'')$, meaning unit i may be a focal unit to compute $\text{TS}_k(\mathbf{Y}^{obs}, \mathbf{t}', \mathbf{A})$ but a non-focal unit to impute $\text{TS}_k(\mathbf{Y}^{obs}, \mathbf{t}'', \mathbf{A})$ and vice-versa. Therefore, the number of focal units $\sum_{i=1}^N F_{ik}(\mathbf{t}')$ is random across $\mathbf{t}' \in \mathcal{T}_k$. In particular, if N_k denotes the units in the population with observed exposure value π_k, then for $\mathbf{t}' \in \mathcal{T}_k$,

$$\sum_{i=1}^N F_{ik}(\mathbf{t}') = N_k \cdot \sum_{t \in \{0,1\}} R(t, \mathbf{t}', \pi_k). \tag{2.13}$$

From (2.13), notice that we control the number of focal units when choosing the subset of treatment assignment vectors $\mathcal{T}_k$ via the tuning parameter ϵ. Larger values of ϵ lead to a higher number of focal units for each treatment assignment vector.

Using definitions (2.10) and (2.12), and a slight abuse of notation, the test statistics in (2.6) conditional on the subsets of focal units and treatment assignment vectors can be technically written as

$$\text{TS}_k(\mathbf{Y}^{obs}, \mathbf{T}, \mathbf{A}; \mathcal{T}_k, F_k) := \text{TS}(\mathbf{T}|\mathbf{A}, \mathbf{Y}^{obs}, \mathbf{T} \in \mathcal{T}_k, F_k(\mathbf{T}) = 1).$$

Consequently, for each $\text{TS}_k, k = 1, \ldots, K$, we can compute the conditional p-value (2.3) as

$$pval_k(\mathbf{T}, \mathbf{Y}^{obs}; \mathcal{T}_k, F_k) = \mathbb{E}_{\mathcal{T}_k}[\mathbb{I}\{\text{TS}_k(\mathbf{Y}_{F_k}(\mathbf{t}^*), \mathbf{t}^*, \mathbf{A}; \mathcal{T}_k, F_k) \geq \text{TS}_k(\mathbf{Y}^{obs}, \mathbf{T}, \mathbf{A}; \mathcal{T}_k, F_k)\}|\mathbf{t}^* \in \mathcal{T}_k, H_0], \tag{2.14}$$

where $\mathbf{Y}_{F_k}(\mathbf{t}^*)$ denotes the component vector of outcomes for focal units (i.e., $F_k(\mathbf{t}^*) = 1$) under H_0. Also, note that the expectation notation $\mathbb{E}_{\mathcal{T}_k}$ is to emphasize that the probability is with respect to $\mathbf{t}^* \in \mathcal{T}_k$.

Now, we focus on the conditioning method for the combined test statistic of H_0. Formally, we write the combined test statistic as

$$\text{TS}(\mathbf{Y}^{obs}, \mathbf{T}, \mathbf{A}) := f(\text{TS}_1, \ldots, \text{TS}_k, \ldots, \text{TS}_K) = \sum_{k=1}^{K} \text{TS}_k. \tag{2.15}$$

This test statistic is also not imputable for all treatment assignment vectors. In this instance, the treatment assignment vectors that guarantee the "imputability" of the test statistic are those that ensure that there are "sufficient" units to compute all the individual TS_k, for $k = 1, \ldots, K$. It is a more stringent requirement compared to the conditions for the individual test statistics. Hence, the conditioning method depends on the choice of the test statistic. It is a novel insight that merits further exploration in future research. Let $\mathcal{T}$ represent the subset of treatment assignment vectors that impute TS, i.e.,

$$\begin{aligned}\mathcal{T} :=& \{\mathbf{t}' \in \{0,1\}^N : p(\mathbf{t}') > 0 \text{ and } R(t, \mathbf{t}', \pi_k) > \epsilon, \forall t = 0, 1, \forall k = 1, \ldots, K, \epsilon \in [0, 1)\} \\ =& \cap_{k=1}^{K} \mathcal{T}_k.\end{aligned} \tag{2.16}$$

In general, by construction, for all $k = 1, \ldots, K$, $|\mathcal{T}| \leq |\mathcal{T}_k|$. Therefore, the sample size and the exposure mapping requirements for this test statistic may be more demanding.

For the combined test statistic, the identity of focal units also depends on the treatment assignment vector. If we let $F(\mathbf{t})$ denote the indicator variable for the focal units that we use to compute

the test statistic TS when the treatment assignment vector is $\mathbf{t}$, then for each unit i, we have

$$F_i(\mathbf{t}') := \mathbb{I}(\pi_i(\mathbf{t}') = \pi, \pi_i(\mathbf{t}^{\mathbf{obs}}) = \pi) \;\; \forall \pi = \pi_1, \ldots, \pi_K \;\; \forall \mathbf{t}' \in \mathcal{T}. \tag{2.17}$$

From the construction of F, note that the focal units are not fixed but vary with the treatment assignment vector. In particular, the number of focal units for $\mathbf{t}' \in \mathcal{T}_k$ is

$$\sum_{i=1}^{N} F_i(\mathbf{t}') = \sum_{k=1}^{K} \left(N_k \sum_{t \in \{0,1\}} R(t, \mathbf{t}', \pi_k) \right). \tag{2.18}$$

Equation (2.18) also shows that we control the number of focal units when choosing $\mathcal{T}$ via the tuning parameter.

Now, using (2.16), (2.17), and a slight abuse of notation, we rewrite the conditional combined test statistics conditional on the focal units and $\mathcal{T}$ as

$$\text{TS}(\mathbf{Y}^{obs}, \mathbf{T}, \mathbf{A}; \mathcal{T}, F) := \text{TS}(\mathbf{T} | \mathbf{A}, \mathbf{Y}^{obs}, \mathbf{T} \in \mathcal{T}, F(\mathbf{T}) = 1),$$

and the conditional p-value as

$$pval(\mathbf{T}, \mathbf{Y}^{obs}; \mathcal{T}, F) = \mathbb{E}_{\mathcal{T}}[\mathbb{I}\{\text{TS}(\mathbf{Y}_F(\mathbf{t}^*), \mathbf{t}^*, \mathbf{A}; \mathcal{T}, F) \geq \text{TS}(\mathbf{Y}^{obs}, \mathbf{T}^{obs}, \mathbf{A}; \mathcal{T}, F)\} | \mathbf{t}^* \in \mathcal{T}, H_0], \tag{2.19}$$

where $\mathbf{Y}_F(\mathbf{t}^*)$ is the vector of outcomes for the focal units ($F(\mathbf{t}^*) = 1$) under the H_0, and the expectation notation $\mathbb{E}_{\mathcal{T}}$ is to emphasize that the probability is with respect $\mathbf{t}^* \in \mathcal{T}$.

The drawback of our proposed CRI method is that, with a small sample size, $N_k, k = 1, \ldots, K$ may be small. Therefore, selecting focal units less than N_k may result in imprecise estimates that lead to invalid p-values. Nevertheless, using the test statistics TS_k and TS, we show the asymptotic validity of the proposed CRI method when the nuisance parameter τ is known. We introduce new notations to study the asymptotic properties of the testing procedure. Let N_0 denote the number of units in the control group, with $N - N_0 = N_1$, and N_{0k} represents the number of units in the control group that have exposure value π_k, with $N_k - N_{0k} = N_{1k}$, for $k = 1, \ldots, K$.

Assumption 2.3.1 *(i) There exist* $\rho \in (0, 1)$ *such that* $\lim_{N\to\infty} N_0/N = \rho$. *(ii) There exist* $\rho_k \in (0, 1)$ *such that* $\lim_{N\to\infty} N_{0k}/N_k = \rho_k$, *where* $k = 1, \ldots, K$.

Assumption 2.3.1 (i) is standard for asymptotic inference in causal studies under no interference. Assumption 2.3.1 (ii) describes the behavior of the fraction of control units among units with a specific exposure value in the limit. It is worthwhile to note that Assumption 2.2.2 is necessary for

Assumption 2.3.1 (ii).

The following theorem states the asymptotic validity results for the proposed CRI method given τ and the test statistics for H_0.

Theorem 2.3.1 (Asymptotic validity CRI method with known nuisance parameters) *Suppose Assumptions 2.2.1-2.3.1 holds, and the variation in potential outcomes is finite across the population. i) Given that* $0 \leq \epsilon < 1$, *under the null hypothesis* H_0 *in* (2.3)*, and for the true value of* τ, *the conditional randomization test using the test statistic* TS_k *is asymptotically valid at any significant level* α, *i.e.,*

$$\lim_{N_k \to \infty} \Pr(pval_k(\mathbf{T}, \mathbf{Y}^{obs}; \mathcal{T}_k, F_k) \leq \alpha) \leq \alpha \text{ for any } \alpha \in [0, 1]. \tag{2.20}$$

ii) Given that $0 \leq \epsilon < 1$, *under the null hypothesis* H_0 *in* (2.3)*, and for the true value of* τ, *the conditional randomization test using the test statistic* TS *is valid at any level* α, *i.e.,*

$$\lim_{N \to \infty} \Pr(pval(\mathbf{T}, \mathbf{Y}^{obs}; \mathcal{T}, F) \leq \alpha) \leq \alpha \text{ for any } \alpha \in [0, 1], \tag{2.21}$$

where the probability is with respect to $\mathbf{T}$.

The proof of Theorem 2.3.1(i) is in Appendix 2.6. Note that as the tuning parameter ϵ approaches 1, our CRI method attains finite sample validity. However, higher values of ϵ require a higher computation time and may lead to low statistical power.

Finally, we can easily show that any multiple testing procedure that uses the marginal p-values $pval_k, k = 1, \ldots, K$, is asymptotically valid. Hence we omit the formal proof.

Remark 2.3.1 *For both test statistics, we acknowledge that it is impossible to credibly make an inference about* H_0 *if we do not observe all the exposure values given the treatment assigned by the experimenter, i.e., if* $\mathbf{t}^{\mathbf{obs}} \notin \mathcal{T}$ *and* $\mathbf{t}^{\mathbf{obs}} \notin \mathcal{T}_k$, *for* $k = 1, \ldots K$. *In such scenarios, a natural refinement is to test the null hypothesis at the observed exposure values.*

2.3.2.2 H_0^{Π}: Constant treatment effect within exposure values

Let us revisit Example 2.3.1. Under H_0^{Π}, we have the corresponding science table in Table 2.2.

The only difference between Tables 2.1 and 2.2 is that we have two nuisance parameters $\tau(0)$ and $\tau(1)$[6] in Table 2.2. Thus, in general, we require K nuisance parameters to test H_0^{Π}. If one knows these nuisance parameters, then the same principles for choosing the subset of treatment assignments and units, as well as the proposed test statistics for H_0 in (2.3), are applicable in testing for H_0^{Π} as well.

[6]The definition of the network exposure mapping in this example produces two exposure values

Units	Observed Variables			Counterfactual Outcomes				Permuted Variables		
i	T_i	Π_i	Y_i^{obs}	$Y_i(1,0)$	$Y_i(0,0)$	$Y_i(1,1)$	$Y_i(0,1)$	$\Lambda(T_i)$	$\pi_i(\tilde{\mathbf{t}})$	Y_i^P
1	0	1	y_1	?	?	$y_1 + \tau(1)$	y_1	1	1	$y_1 + \tau(1)$
2	0	0	y_2	$y_2 + \tau(0)$	y_2	?	?	1	1	?
3	0	1	y_3	?	?	$y_3 + \tau(1)$	y_3	1	1	$y_3 + \tau(1)$
4	1	0	y_4	y_4	$y_4 - \tau(0)$	?	?	1	1	?
5	1	0	y_5	y_5	$y_5 - \tau(0)$	?	?	1	1	?
6	1	0	y_6	y_6	$y_6 - \tau(0)$	?	?	0	1	?
7	0	1	y_7	?	?	$y_7 + \tau(1)$	y_7	0	0	?
8	1	1	y_8	?	?	y_8	$y_8 - \tau(1)$	0	0	?
9	1	1	y_9	?	?	y_9	$y_9 - \tau(1)$	0	0	?
10	0	0	y_{10}	$y_{10} + \tau(0)$	y_{10}	?	?	0	0	y_{10}

Table 2.2: A Science Table under H_0^{Π}, using Example 2.3.1. NB: Y_i^P represents the new outcome of unit i under H_0^{Π} for the new treatment vector $\tilde{\mathbf{t}}$.

2.3.2.3 $H_0^{X,\Pi}$: Constant treatment effect across pretreatment variables and exposure values

We first describe the "non-sharpness" of $H_0^{X,\Pi}$ and its CRI method using the test statistics $\text{TS}_{k,l}$, $k = 1, \ldots, K$, and $l = 1, \ldots, L$ in (2.8). Based on arguments similar to those we employ for H_0, note that the test statistic $\text{TS}_{k,l}$ is not imputable. The following example builds on Example 2.3.1 and shows the "non-sharpness" of $H_0^{X,\Pi}$.

Example 2.3.2 *Assume there are $N = 10$ units in an undirected social network, as in Figure 2.2. Our realized data is $(Y_i, T_i, \Pi_i, X_i)_{i=1}^N$, where Π_i is defined in Example 2.3.1. Table 2.3 shows that each unit has two missing potential outcomes. Hence, $H_0^{X,\Pi}$ is also not sharp. Now, using the permuted treatment vector $\tilde{\mathbf{t}} = (1,1,1,1,1,0,0,0,0,0)$, we can deduce that none of the test statistics $\text{TS}_{k,l}, k = 1, \ldots, K$ and $l = 1, \ldots, L$ is imputable under $H_0^{X,\Pi}$ in (2.5). In other words, for each $\pi \in \{0,1\}$, and $x \in \{f, m\}$ there exist $\Lambda(\mathbf{T}) = \mathbf{t}' \in \{0,1\}^N$ where $p(\mathbf{t}') > 0$, but $\sum_{i=1}^N \mathbb{I}(\Lambda(T_i) = t, \pi_i(\mathbf{t}') = \pi, X_i = x) = 0$ for either $t = 0$ or $t = 1$, and $\pi \in \mathbf{\Pi}$.*

Units	Observed Variables				Counterfactual Outcomes				Permuted Variables		
i	T_i	Π_i	X_i	Y_i^{obs}	$Y_i(1,0)$	$Y_i(0,0)$	$Y_i(1,1)$	$Y_i(0,1)$	$\Lambda(T_i)$	$\pi_i(\tilde{\mathbf{t}})$	Y_i^P
1	0	1	m	y_1	?	?	$y_1 + \tau(1;m)$	y_1	1	1	$y_1 + \tau(1;m)$
2	0	0	m	y_2	$y_2 + \tau(0;m)$	y_2	?	?	1	1	?
3	0	1	m	y_3	?	?	$y_3 + \tau(1;m)$	y_3	1	1	$y_3 + \tau(1;m)$
4	1	0	f	y_4	y_4	$y_4 - \tau(0;f)$	?	?	1	1	?
5	1	0	f	y_5	y_5	$y_5 - \tau(0;f)$	?	?	1	1	?
6	1	0	m	y_6	y_6	$y_6 - \tau(0;m)$	?	?	0	1	?
7	0	1	f	y_7	?	?	$y_7 + \tau(1;f)$	y_7	0	0	?
8	1	1	m	y_8	?	?	y_8	$y_8 - \tau(1;m)$	0	0	?
9	1	1	f	y_9	?	?	y_9	$y_9 - \tau(1;f)$	0	0	?
10	0	0	f	y_{10}	$y_{10} + \tau(0)$	y_{10}	?	?	0	0	y_{10}

Table 2.3: A Science Table under $H_0^{X,\Pi}$, using Example 2.3.1. NB: Y_i^P represents the new outcome of unit i under $H_0^{X,\Pi}$ for the new treatment vector $\tilde{\mathbf{t}}$.

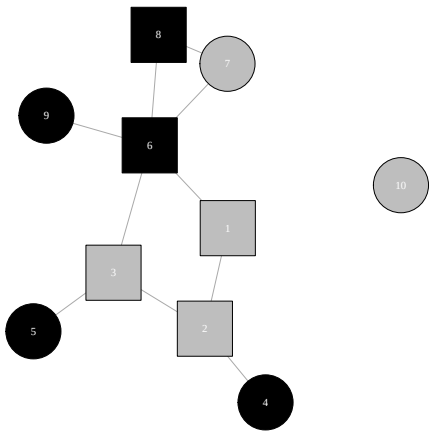


Figure 2.2: Undirected Social Network. (Note: Grey nodes are the control units and black nodes are treatment units. Circle nodes are females and square nodes are males.)

To estimate the null distributions of any of the test statistics $\text{TS}_{k,l}$, where the $L \times K$ nuisance parameters in $H_0^{X,\Pi}$ are known, we employ a CRI method characterized by the subset of treatment assignment vectors and focal units. First, let $\mathcal{T}_{k,l}$ denote the subset of treatment assignment vectors

that impute $\text{TS}_{k,l}$. Formally, we define $\mathcal{T}_{k,l}$ as

$$\mathcal{T}_{k,l} := \{\mathbf{t}' \in \{0,1\}^N : p(\mathbf{t}') > 0 \text{ and } R(t, \mathbf{t}', \pi_k, x_l) > \epsilon,\ t = 0, 1, \epsilon \in [0, 1)\}, \tag{2.22}$$

with

$$R(t, \mathbf{t}', \pi_k, x_l) := \frac{\sum_{i=1}^{N} \mathbb{I}\{t_i' = t, \pi_i(\mathbf{t}') = \pi_k, \pi_i(\mathbf{t}^{obs}) = \pi_k, \mathbf{X}_i = x_l\}}{\sum_{i=1}^{N} \mathbb{I}\{\pi_i(\mathbf{t}^{obs}) = \pi_k, \mathbf{X}_i = x_l\}}, \tag{2.23}$$

where the $R(\cdot, \cdot, \cdot, \cdot)$ is an empirical probability. In simpler terms, $\mathcal{T}_{k,l}$ is the subset of treatment assignment vectors (that respect the assignment mechanism), where there is a "sufficient" number of units in each of the arms of treatment that for a given pretreatment variable value x_l, and the exposure value remains fixed at π_k.

By construction, the sets $\mathcal{T}_{k,l},\ k = 1, \ldots, K$ and $l = 1, \ldots, L$, depend on the sample size, the number of pretreatment variables defined subgroups, and the number of exposure values. For each k and l, the cardinality of $\mathcal{T}_{k,l}$ is bigger when we have a large sample and the number of covariate-defined subgroups is small. Also, exposure mappings with a smaller range of values are more likely to lead to a larger $|\mathcal{T}_{k,l}|$.

Second, let $F_{k,l}(\mathbf{t})$ represent the indicator variable for the focal units when the treatment assignment vector is $\mathbf{t}$. Therefore, for unit i, we have

$$F_{ik,l}(\mathbf{t}') := \mathbb{I}(\pi_i(\mathbf{t}') = \pi_k, \pi_i(\mathbf{t}^{obs}) = \pi_k, \mathbf{X} = x_l)\ \ \forall \mathbf{t}' \in \mathcal{T}_{k,l}, \tag{2.24}$$

Therefore, for $\mathbf{t}' \in \mathcal{T}_{k,l}$, the number of focal units is

$$\sum_{i=1}^{N} F_{ik,l}(\mathbf{t}') = N_{k,l} \cdot \sum_{t \in \{0,1\}} R(t, \mathbf{t}', \pi_k, x_l) \tag{2.25}$$

where $N_{k,l}$ represents the number of units with $X = x_l$ and has observed exposure value of π_k. Equation (2.25) also indicates that we control the expected number of focal units using the tuning parameter ϵ. In addition, when we compare the equations (2.13) and (2.25), it is evident that the number of focal units is smaller under the null $H_0^{X,\Pi}$ for the single test statistics. It suggests that one may require more data to test $H_0^{X,\Pi}$ compared to H_0 and H_0^{Π}.

Next, using definitions (2.22), (2.24), and slightly abusing notation, the conditional test statistics can be written technically as

$$\text{TS}_{k,l}(\mathbf{Y}^{obs}, \mathbf{T}, \mathbf{A}, \mathbf{X}; \mathcal{T}_{k,l}, F_{k,l}) := \text{TS}(\mathbf{T}|\mathbf{A}, \mathbf{Y}^{obs}, \mathbf{T} \in \mathcal{T}_{k,l}, F_{k,l}(\mathbf{T}) = 1, \mathbf{X} = x_l).$$

Therefore, for each $\text{TS}_{k,l}$, we can also compute the conditional p-value as

$$pval_{kl}(\mathbf{T}, \mathbf{Y}^{obs}, \mathbf{X}; \mathcal{T}_{k,l}, F_{k,l}) = \mathbb{E}_{\mathcal{T}_{k,l}}[\mathbb{I}\{\text{TS}_{k,l}(\mathbf{Y}_{F_{k,l}}(\mathbf{t}^*), \mathbf{t}^*, \mathbf{A}, \mathbf{X}; \mathcal{T}_{k,l}, F_{k,l}) \geq \text{TS}_{k,l}(\mathbf{Y}^{obs}, \mathbf{T}, \mathbf{A}, \mathbf{X}; \mathcal{T}_{k,l}, F_{k,l})\} | \mathbf{t}^* \in \mathcal{T}_{k,l}, H_0^{X,\Pi}], \tag{2.26}$$

where $\mathbf{Y}_{F_{k,l}}(\mathbf{t}^*)$ is the vector of outcomes for focal units (i.e., $F_{k,l}(\mathbf{t}^*) = 1$) and the expectation notation $\mathbb{E}_{\mathcal{T}_{k,l}}$ is to emphasize that the probability is over the set $\mathcal{T}_{k,l}$.

Let us now consider the combined test statistic for $H_0^{X,\Pi}$. Formally, we write the combined test statistic as the sum of the single $L \times K$ test statistics, i.e.,

$$\text{TS}^{X,\Pi}(\mathbf{Y}^{obs}, \mathbf{T}, \mathbf{A}, \mathbf{X}) := f(\text{TS}_{k,l} : x_l \in \mathbb{X}, k = 1, \ldots, K) = \sum_{k=1}^{K} \sum_{l=1}^{L} \text{TS}_{k,l}. \tag{2.27}$$

This test statistic is also not imputable under $H_0^{X,\Pi}$. The treatment assignment vectors that guarantee the imputability of $\text{TS}^{X,\Pi}$ are those that ensure that there are a "sufficient" number of units to compute each $\text{TS}_{k,l}$. Let $\mathcal{T}^{X,\Pi}$ represent the subset of treatment assignment vectors that impute $\text{TS}^{X,\Pi}$. Formally,

$$\begin{aligned} \mathcal{T}^{X,\Pi} :=& \{\mathbf{t}' \in \{0,1\}^N : p(\mathbf{t}') > 0 \text{ and } R(t, \mathbf{t}', \pi_k, x_l) \geq \epsilon, \ \forall t = 0, 1, \\ & \forall k = 1, \ldots, K \quad \forall l = 1, \ldots, L \text{ and } \epsilon \in (0, 1) \ \} \\ =& \cap_{k=1}^{K} \mathcal{T}_{k,l} \end{aligned} \tag{2.28}$$

For instance, in Example 2.3.2, we can deduce that $\tilde{\mathbf{t}} \notin \mathcal{T}^{X,\Pi}$. By construction, for all $k = 1, \ldots, K$, and $l = 1, \ldots, L$, $|\mathcal{T}^{X,\Pi}| \leq |\mathcal{T}_{k,l}|$. Therefore, the sample size and exposure mapping requirements for the combined test statistic may be more demanding than those for each $\text{TS}_{k,l}$.

The definition of the focal units, however, is the same as the one we define for the combined test statistics of H_0 in (2.17). It is because pretreatment variables are not causal but control variables. Again, with a slight abuse of notation, the combined test statistics $\text{TS}^{X,\Pi}$ conditional on the subset of treatment assignment vectors and the focal units can also be written as

$$\text{TS}^{X,\Pi}(\mathbf{Y}^{obs}, \mathbf{T}, \mathbf{X}, \mathbf{A}; \mathcal{T}^{X,\Pi}, F) := \text{TS}(\mathbf{T} | \mathbf{A}, \mathbf{Y}^{obs}, \mathbf{T} \in \mathcal{T}^{X,\Pi}, F(\mathbf{T}) = 1)$$

and the resulting conditional p-value is

$$pval(\mathbf{T}, \mathbf{Y}^{obs}, \mathbf{X}; \mathcal{T}^{X,\Pi}, F) = \mathbb{E}_{\mathcal{T}^{X,\Pi}}[\mathbb{I}\{\text{TS}(\mathbf{Y}_F(\mathbf{t}^*), \mathbf{t}^*, \mathbf{X}, \mathbf{A}; \mathcal{T}^{X,\Pi}, F) \geq \text{TS}(\mathbf{Y}^{obs}, \mathbf{T}, \mathbf{X}, \mathbf{A}; \mathcal{T}^{X,\Pi}, F)\} | \mathbf{t}^* \in \mathcal{T}_{k,l}, H_0^{X,\Pi}] \tag{2.29}$$

where the expectation notation $\mathbb{E}_{\mathcal{T}^{X,\Pi}}$ is to emphasize that the probability is with respect to $\mathbf{t}^* \in \mathcal{T}^{X,\Pi}$.

Finally, note that in practice, the sets $\mathcal{T}_k$, $\mathcal{T}$, $\mathcal{T}_{k,l}$ and $\mathcal{T}^{X,\Pi}$ may be large and one may have to

approximate the p-values in (2.14), (2.19), (2.26), and (2.29). We recommend drawing a random sample of elements of size B from these sets to compute the p-values. The following algorithm summarizes our proposed CRI testing procedure with known nuisance parameters.

Algorithm 1: Conditional Randomization Algorithm for HTE with known nuisance parameters

Data: $(\mathbf{Y}^{obs}, \mathbf{T}, \mathbf{A}, \mathbf{X}, \{\Pi_i\}_i^N)$

Result: the estimated p-values: e.g., $\widehat{pval}_k(\mathbf{T}, \mathbf{Y}^{obs}; \mathcal{T}_k, F_k)$

Compute the appropriate test statistic using the observed data: e.g.,$\text{TS}_k(\mathbf{Y}^{obs}, \mathbf{T}; \mathcal{T}_k, F_k)$.

for *{b = 1 to B}* **do**

  Choose ϵ and draw $\mathbf{t}^{(b)}$ independently from the appropriate subset of treatment assignment vectors , e.g.,$\mathcal{T}_k$.

  Compute the test statistic using $\mathbf{t}^{(b)}$ and the focal units, e.g., $\text{TS}_k(\mathbf{Y}_{F_k}(\mathbf{t}^{(b)}), \mathbf{t}^{(b)}; \mathcal{T}_k, F_k)$

Compute the empirical p-value, e.g.,
$\widehat{pval}_k(\mathbf{T}, \mathbf{Y}^{obs}; \mathcal{T}_k, F_k) = B^{-1} \sum_{b=1}^{B} \mathbb{I}\{\text{TS}_k(\mathbf{Y}_{F_k}(\mathbf{t}^{(b)}), \mathbf{t}^{(b)}; \mathcal{T}_k, F_k) \geq \text{TS}_k(\mathbf{Y}^{obs}, \mathbf{T}; \mathcal{T}_k, F_k)\}$

Under any of the null hypotheses, we can show that the empirical p-values are also asymptotically valid. For instance, we can show that $\lim_{N_k \to \infty} \Pr(\widehat{pval}_k(\mathbf{T}, \mathbf{Y}^{obs}; \mathcal{T}_k, F_k) \leq \alpha) \leq \alpha$ for any $\alpha \in [0, 1]$, where the probability reflects variations in both treatment assignment and the sampling of treatment vectors from $\mathcal{T}_k$. In fact, this validity results hold for any B. However, the larger the value of B, the less the approximation error, $\widehat{pval}_k - pval$. Finally, note that we can easily extend Theorem 2.3.1 to H_0^{Π} and $H_0^{X,\Pi}$.

It is worthwhile comparing the proposed CRI method to existing conditioning methods in the literature. In general, our approach of choosing the subset of treatment assignments and focal unit resembles the *artificial experiment* concept by Athey et al. (2018) and the *conditioning mechanism* contrast by Basse et al. (2019). Nevertheless, there are three important distinctions. First, the choice of the subset of treatment assignments vectors and the focal units are intertwined. Specifically, we propose choosing the subset of treatment assignment vectors to maximize the number of focal units, and we choose the set of focal units to maximize the cardinality of the subset of treatment assignment vectors. In contrast, the approaches of the papers we cite above suggest a prior random or non-random selection of focal units that do not depend on the subset of treatment assignment vectors. Our procedure guarantees a larger subset of treatment assignment vectors leading to higher statistical power for the hypotheses we consider. A further examination into which of these approaches works better, in general, is open for future research.

Second and closely related, we allow focal units to vary across the elements in the subset of

treatment assignment vectors. Simply put, the number of focal units can differ based on the treatment assignment vectors, which sets it apart from previous methods. Specifically, the number of focal units is a binomial random variable (across the subset of treatment assignment vectors), which depends on the tuning parameter ϵ. Although allowing the focal units to vary across treatment assignments may lead to invalid p-values in small samples, it increases the cardinality of the subset of treatment assignment vectors leading to an improvement in asymptotic statistical power and size. For example, suppose we have a population of four units with two of them treated, and paired into two dyads. Then, testing the null hypothesis of *no spillovers*, the approaches of Athey et al. (2018) and Basse et al. (2019) produce a subset of treatment assignments of size two. However, allowing focal units to vary across treatment assignments leads to a subset of treatment assignments of size four. Depending on the test statistic, all four treatment assignments may be useful in approximating the null distribution of the test statistic, leading to improvement in statistical power and size. Again, a generalization of our proposed conditioning method that allows for varying focal units is open for future research.

Finally, the focal units in our proposed CRI approach indirectly depend on realized treatment through network exposure. In contrast, the *artificial experiment* approach explicitly advocates against using realized treatments or outcomes to select the focal units. Our results show that the indirect use of realized treatments to pick the focal units does not affect the validity of the conditional randomization method of inference, rather, it leads to improvements in statistical power and computation.

In summary, our proposed CRI method is a methodological contribution of the current chapter. It offers fresh perspectives on choosing focal units and the subset of treatment assignment vectors. It is a generalization of the existing approaches. For instance, if we set $R(t, \mathbf{t}', \pi_k)$ in equation (2.10) such that $R(0, \mathbf{t}', \pi_k) + R(1, \mathbf{t}', \pi_k) = 1$ for all $\mathbf{t}' \in \mathcal{T}_k$, then $\mathcal{T}_k$ will coincide with the subset of treatment assignment vectors one would obtain using the existing approaches. The focal units become fixed and uniformly larger (across the subset of treatment assignment vectors). However, it is computationally expensive and may reduce the cardinality of the subset of treatment assignment vectors leading to low statistical power and high size distortions.

2.3.3 Dealing with Unknown Nuisance Parameters

Knowledge of the parameters $\tau, \tau(\pi)$, and $\tau(\pi; x)$ for all $\pi \in \mathbf{\Pi}$ and for all $x \in \mathbb{X}$ implies that we can directly apply our proposed CRI method, and Theorem 2.3.1 hold. In practice, however, these parameters are unknown. A natural but naive approach is to replace the unknown nuisance parameters with their sample counterparts and proceed to apply the CRI method. Unfortunately,

this leads to invalid p-values in finite samples and asymptotically. To develop intuition into why the naive approach does not work, let us look at the definition of the *randomization hypothesis* (Lehmann and Romano, 2022, p. 832).

Definition 2.3.1 (Randomization Hypothesis) *Under the null hypotheses* (2.3)-(2.5)*, the distribution of the original sample data* $(\mathbf{Y}^{obs}, \mathbf{T}, \Pi, \mathbf{X})$ *is the same as the distribution of the samples we generate by replacing* $\mathbf{T}$ *with a new treatment vector* $\Lambda(\mathbf{T})$ *that respects the treatment assignment mechanism.*

Definition 2.3.1 is a crucial assumption that guarantees the validity of randomization-based inferential methods. The practical consequence of this definition is that, under the null hypotheses, the distribution of the test statistic that we compute using the original data is the same as the distributions of the test statistic we compute using data we generate with a new treatment vector that respects the treatment assignment mechanism.

Unfortunately, estimated nuisance parameters lead to a breakdown of the randomization hypothesis, and we illustrate this using the following example.

Example 2.3.3 *Suppose our sample comprises four* $(N = 4)$ *unconnected individuals with the original sample data given as* $(\mathbf{Y}^{obs} = \{y_1, y_2, y_3, y_4\}, \mathbf{T} = \{1, 1, 0, 0\}, \Pi = \{\pi_1, \pi_1, \pi_1, \pi_1\})$*. Consider the null hypothesis* H_0 *in* (2.3) *and estimate the nuisance parameter* τ *using the difference in means estimator, i.e.,* $\hat{\tau} = (y_1 + y_2 - y_3 - y_4)/2$*. Now, substituting* $\hat{\tau}$ *for* τ *in* (2.3)*, we can impute all missing potential outcomes. If we choose a new treatment vector* $\Lambda(\mathbf{T}) = \mathbf{t}'' = \{0, 0, 1, 1\}$ *and assume that the exposure values remain the same under* $\mathbf{t}''$*, then our new sample data is* $(\mathbf{Y}'' = \{(y_1 - y_2 + y_3 + y_4)/2, (y_2 - y_1 + y_3 + y_4)/2, (y_1 + y_2 + y_3 - y_4)/2, (y_1 + y_2 - y_3 + y_4)/2\}, \Lambda(\mathbf{T}) = \{0, 0, 1, 1\}, \Pi'' = \{\pi_1, \pi_1, \pi_1, \pi_1\})$*. When we compare* $\mathbf{Y}^{obs}$ *to* $\mathbf{Y}''$*, it is obvious that under* H_0*, the distributions of the original sample and the new sample are not the same. Therefore, the distribution of test statistics that inherit the underlying distributional properties of the two samples will also be different even if the* H_0 *is true.*

The Example 2.3.3 highlights the ensuing problem of replacing nuisance parameters with their sample counterparts in the hypotheses we consider. It shows that the randomization tests with estimated nuisance parameters are more likely to reject the null hypotheses, even if the null hypotheses are true due to a breakdown of the randomization hypothesis. It is *not* only a finite sample problem. Even asymptotically, the randomization hypothesis would not hold. Next, we propose two solutions - sample splitting technique and confidence interval technique - to handle the unknown nuisance parameters.

2.3.3.1 Sample Splitting Technique

The sample splitting (SS) technique we propose overcomes the problem of estimating nuisance parameters by randomly splitting the original data into two *balanced* sub-samples. We show that the p-values from jointly using the SS technique and our proposed CRI method are asymptotically valid. The SS technique comprises the following steps:

1. Randomly split the full sample into two equal and balanced sub-samples $\mathcal{U}^{est}$ and $\mathcal{U}^{inf}$. We use $\mathcal{U}^{est}$ for the estimation of the nuisance parameters and $\mathcal{U}^{inf}$ to test the hypothesis. Therefore, $\mathcal{U}^{est}$ and $\mathcal{U}^{inf}$ denote the estimation and inference sub-samples respectively.

2. Estimate the nuisance parameter(s) using $\mathcal{U}^{est}$.

3. Randomize the *full treatment vector* while respecting the treatment assignment mechanism, obtain the exposure values for the full sample, and compute the test statistics using $\mathcal{U}^{inf}$.

Due to network interference, one must randomize the full treatment vector rather than the treatment vector of the inference sub-sample. It is the notable difference between our sample splitting method compared to other splitting schemes for different purposes in the econometrics, statistics, and machine learning literature.

To understand why the proposed technique is effective, let us revisit Example 2.3.3. We established that the randomization hypothesis fails in this example when we replace the nuisance parameter with its sample counterpart.

Now, let us apply the SS technique. A possible balanced split of the original sample data into two is $\mathcal{U}^{est} = (\mathbf{Y}^{obs}_{est} = \{y_1, y_3\}, \mathbf{T}_{est} = \{1, 0\}, \Pi_{est} = \{\pi_1, \pi_1\})$ and $\mathcal{U}^{inf} = (\mathbf{Y}^{obs}_{inf} = \{y_2, y_4\}, \mathbf{T}_{inf} = \{0, 1\}, \Pi_{inf} = \{\pi_1, \pi_1\})$. We compute the difference in means estimator of τ using $\mathcal{U}^{est}$, i.e., $\hat{\tau}^{est} = y_1 - y_3$. Under H_0, the new treatment vector $\Lambda(\mathbf{T}) = \mathbf{t}'' = \{0, 0, 1, 1\}$ produces a new inference sub-sample $\mathcal{U}''^{inf} = (\mathbf{Y}''_{inf} = \{y_2 - \hat{\tau}^{est}, y_4 + \hat{\tau}^{est}\}, \mathbf{t}''_{inf} = \{0, 1\}, \Pi''_{inf} = \{\pi_1, \pi_1\}) = (\mathbf{Y}''_{inf} = \{(y_2 - y_1 + y_3), (y_4 + y_1 - y_3)\}, \mathbf{t}''_{inf} = \{0, 1\}, \Pi''_{inf} = \{\pi_1, \pi_1\})$.

If we compare $\mathcal{U}^{inf}$ to $\mathcal{U}''^{inf}$, the distributions of the test statistic conditional on $\mathcal{U}^{inf}$ (the original inference sub-sample) and conditional on $\mathcal{U}''^{inf}$ are the same since $\hat{\tau}^{est}$ is non-stochastic in the new inference sub-sample $\mathcal{U}''^{inf}$. Thus, the randomization hypothesis (which breaks down when we use the full sample data for estimation and inference) holds when we apply the sample splitting technique.

Under the null hypothesis H_0 in (2.3), and for the test statistic $\text{TS}_k(\mathbf{Y}^{obs}, \mathbf{T}, \mathbf{A}; \mathcal{T}_k, F_k)$, we denote the p-value using the SS technique on the CRI method as $pval_k(\mathbf{T}, \mathbf{Y}^{obs}; \mathcal{T}_k, F_k, \hat{\tau}^{est})$. Similarly, for

the combined test statistic $\text{TS}(\mathbf{Y}^{obs}, \mathbf{T}, \mathbf{A}; \mathcal{T}, F)$, we denote the p-values we obtain from applying the SS technique on our CRI method as $pval(\mathbf{T}, \mathbf{Y}^{obs}; \mathcal{T}, F, \hat{\tau}^{est})$.

The disadvantage of the SS technique is the loss of information due to sample splitting. Therefore, applying the SS technique to our CRI method in small samples may result in invalid p-values. However, we establish the *asymptotic* validity of the ensuing p-values when we apply the SS technique to our proposed CRI method in the following theorem.

Theorem 2.3.2 (Asymptotic Validity of the SS technique on CRI method) *Suppose Assumptions 2.2.1-2.3.1 holds, and the variation in potential outcomes is finite across the population. i) Given that* $0 \leq \epsilon < 1$, *under the null hypothesis* H_0 *in* (2.3) *, using the test statistic* $\text{TS}_k(\mathbf{Y}^{obs}, \mathbf{T}, \mathbf{A}; \mathcal{T}_k, F_k)$, *and given that* $\hat{\tau}^{est}$ *is a consistent estimate of* τ *conditional on* $\mathcal{U}^{est}$, *a balanced half sub-sample of the full data. Then,*

$$\lim_{N\to\infty} \Pr(pval_k(\mathbf{T}, \mathbf{Y}^{obs}; \mathcal{T}_k, F_k, \hat{\tau}^{est}) \leq \alpha) \leq \alpha \text{ for any } \alpha \in [0,1] \tag{2.30}$$

ii) Given that $0 \leq \epsilon < 1$, *under the null hypothesis* H_0 *in* (2.3), *using the test statistic* $\text{TS}(\mathbf{Y}^{obs}, \mathbf{T}, \mathbf{A}; \mathcal{T}, F)$, *and given that* $\hat{\tau}^{est}$ *is a consistent estimate of* τ *conditional on* $\mathcal{U}^{est}$, *a balanced half sub-sample of the full data. Then,*

$$\lim_{N\to\infty} \Pr(pval(\mathbf{T}, \mathbf{Y}^{obs}; \mathcal{T}, F, \hat{\tau}^{est}) \leq \alpha) \leq \alpha \text{ for any } \alpha \in [0,1] \tag{2.31}$$

where the probabilities are taken over $\mathbf{T}$.

Note that we can easily extend Theorem 2.3.2 to H_0^{Π} and $H_0^{X,\Pi}$. We summarize the testing procedure of applying the SS technique to the CRI method in Algorithm 2.

2.3.3.2 Confidence Interval Technique

This technique is an adaptation of the "confidence interval" (CI) technique of Berger and Boos (1994), which is put into practice in the context of *unconditional* randomization inference by Ding et al. (2016). We extend the technique to conditional randomization inference. The idea is to compute the maximum p-value across a finite number of estimates of the nuisance parameter in a given confidence interval. The pointwise computation of the p-values at each nuisance parameter(s) in the confidence interval suggests that one can view the parameter(s) as constant(s); therefore, the randomization hypothesis holds.

Before we state the validity results from jointly applying the CI technique and the CRI method, let us formally describe the CI technique on our proposed CRI method in relation to H_0. Suppose

Algorithm 2: Conditional Randomization Algorithm for HTE with Estimated nuisance parameters Using the SS technique

Data: $(\mathbf{Y}^{obs}, \mathbf{T}, \mathbf{A}, \mathbf{X}, \{\Pi_i\}_i^N)$

Result: the estimated p-values: e.g., $\widehat{pval}_k(\mathbf{T}, \mathbf{Y}^{obs})$

1 Randomly split the data into two balanced sub-samples, i.e., $\mathcal{U}^{est}$ and $\mathcal{U}^{inf}$.

2 Estimate the nuisance parameter(s) using $\mathcal{U}^{est}$. e.g., $\hat{\tau}^{est}$ for τ

3 Compute the appropriate test statistic using the original sub-sample $\mathcal{U}^{inf}$ e.g.,$\text{TS}_k(\mathbf{Y}^{obs}, \mathbf{T}; \mathcal{T}_k, F_k)$.

4 **for** *{b = 1 to B}* **do**

5 Choose ϵ and draw $\mathbf{t}^{(b)}$ independently from the appropriate subset of treatment assignment vectors *of the full data* e.g.,$\mathcal{T}_k$.

6 Use $\mathbf{t}^{(b)}$ to obtain the new exposure values.

7 Compute the test statistic using using $\mathbf{t}^{(b)}$ and the focal units that are in $\mathcal{U}^{inf}$ e.g., $\text{TS}_k(\mathbf{Y}_{F_k}(\mathbf{t}^{(b)}), \mathbf{t}^{(b)}; \mathcal{T}_k, F_k)$

8 Compute the empirical p-value, e.g., $\widehat{pval}_k(\mathbf{T}, \mathbf{Y}^{obs}; \mathcal{T}_k, F_k, \hat{\tau}^{est}) = B^{-1} \sum_{b=1}^{B} \mathbb{I}\{\text{TS}_k(\mathbf{Y}_{F_k}(\mathbf{t}^{(b)}), \mathbf{t}^{(b)}; \mathcal{T}_k, F_k) \geq \text{TS}_k(\mathbf{Y}^{obs}, \mathbf{T}; \mathcal{T}_k, F_k)\}$

CI_γ is the $(1-\gamma)$ confidence interval[7] for the unknown τ in the null hypothesis (2.3), then according to Berger and Boos (1994), for each k, the conditional p-value over CI_γ using the test statistic TS_k is

$$pval_{k,\gamma}(\mathbf{T}, \mathbf{Y}^{obs}; \mathcal{T}_k, F_k) := \sup_{\tau' \in \text{CI}_\gamma} pval_k(\tau') + \gamma,$$

where $pval_k(\tau')$ is the p-value when $\tau = \tau'$. On the other hand, the conditional p-value over CI_γ using the combined test statistic TS is

$$pval_\gamma(\mathbf{T}, \mathbf{Y}^{obs}; \mathcal{T}, F) := \sup_{\tau' \in \text{CI}_\gamma} pval(\tau') + \gamma,$$

where $pval(\tau')$ is the p-value when $\tau = \tau'$. We approximate the confidence intervals using the Neyman variance estimator, which is valid under the proposed null hypotheses.

Although this technique is attractive, the resulting p-values are the "worst-case" p-values, and they are conservative. However, the following theorem shows the asymptotic validity of the procedure.

[7]For H_0^{Π} and $H_0^{X,\Pi}$, CI_γ represents Bonferroni-corrected confidence region.

Theorem 2.3.3 (Asymptotic Validity of the CI technique on CRI method) *Suppose Assumptions 2.2.1-2.3.1 holds, and the variation in potential outcomes is finite across the population. i) Given that* $0 \leq \epsilon < 1$, *under the null hypothesis* H_0 *in* (2.3) *, using the test statistic* $\text{TS}_k(\mathbf{Y}^{obs}, \mathbf{T}, \mathbf{A}; \mathcal{T}_k, F_k)$, *and given that* $\text{CI}_\gamma, \gamma \in (0,1)$ *is a* $(1-\gamma)$ *confidence interval for the nuisance parameter* τ, *then, for all* $k = 1, \ldots, K$, $pval_{k,\gamma}$ *is an asymptotically valid p-value at any level* $\alpha \in [0,1]$, *i.e.,*

$$\lim_{N_k \to \infty} \Pr(pval_{k,\gamma}(\mathbf{T}, \mathbf{Y}^{obs}; \mathcal{T}_k, F_k) \leq \alpha) \leq \alpha \text{ for any } \alpha \in [0,1] \tag{2.32}$$

ii) Given that $0 \leq \epsilon < 1$, *under the null hypothesis* H_0 *in* (2.3), *using the test statistic* $\text{TS}(\mathbf{Y}^{obs}, \mathbf{T}, \mathbf{A}; \mathcal{T}, F)$, *and given that* $\text{CI}_\gamma, \gamma \in (0,1)$ *is a* $(1-\gamma)$ *confidence interval for the nuisance parameter* τ, *then* $pval_\gamma$ *is a valid p-value at any level* $\alpha \in [0,1]$, *i.e.,*

$$\lim_{N \to \infty} \Pr(pval_\gamma(\mathbf{T}, \mathbf{Y}^{obs}; \mathcal{T}, F) \leq \alpha) \leq \alpha \text{ for any } \alpha \in [0,1] \tag{2.33}$$

where the probabilities are taken over $\mathbf{T}$.

We defer the proof of (i) to the Appendix 2.6. Theorem 2.3.3 shows the asymptotic validity of applying the CI technique to the CRI approach. Additionally, from Theorems 2.3.2(i) and 2.3.3(i), it is straightforward to show the validity of testing H_0 using a multiple testing procedure which jointly tests the K null hypotheses $H_0^{(k)} : Y_i(1, \pi_k) - Y_i(0, \pi_k) = \tau$ for some τ, for $i = 1, \ldots, N$ where $k = 1 \ldots K$. In addition, Theorem 2.3.3 also applies to H_0^{Π} and $H_0^{X,\Pi}$.

Finally, it may be infeasible to compute the p-value at each point in the confidence interval. Therefore, we calculate the p-values on a finite uniform grid in the estimated confidence interval. Algorithm 3 summarizes how we implement the testing procedure when we apply the CI technique to the CRI method.

Algorithm 3: Conditional Randomization Algorithm for HTE with Estimated Nuisance Parameters Using the CI technique

Data: $(\mathbf{Y}^{obs}, \mathbf{T}, \mathbf{A}, \mathbf{X}, \{\Pi_i\}_i^N)$

Result: the estimated p-values: e.g., $\widehat{pval}_k(\mathbf{T}, \mathbf{Y}^{obs})$

Compute the appropriate test statistic using the observed data: e.g.,$\text{TS}_k(\mathbf{Y}^{obs}, \mathbf{T}, \mathbf{A})$.

Select $\gamma \in (0, 1)$, e.g. $\gamma = 0.001$

Estimate the confidence interval (region) of the nuisance parameter(s), (e.g., $\hat{CI}_\gamma(\tau)$) and obtain a finite uniform grid of points $C = \{\tau_1 \dots \tau_M\}$ in the confidence in interval.

for $\{\tau' = \tau_1 \; to \; \tau_M\}$ **do**

  for $\{b = 1 \; to \; B\}$ **do**

    Draw $\mathbf{t}^{(b)}$ independently from the appropriate subset of treatment assignment vectors, e.g.,$\mathcal{T}_k$.

    Compute the test statistic using $\mathbf{t}^{(b)}$ under the corresponding null hypothesis, e.g., $\text{TS}_k(\mathbf{Y}_{F_k}(\mathbf{t}^{(b)}), \mathbf{t}^{(b)}, \tau')$

  Compute the empirical p-value for each τ', e.g., $\widehat{pval}_k(\mathbf{T}, \mathbf{Y}^{obs}, \tau') = B^{-1} \sum_{b=1}^{B} \mathbb{I}\{\text{TS}_k(\mathbf{Y}_{F_k}(\mathbf{t}^{(b)}), \mathbf{t}^{(b)}, \tau') \geq \text{TS}_k(\mathbf{Y}^{obs}, \mathbf{T}, \tau')\}$

Obtain the maximum empirical p-value across the grid points and add γ. e.g., $\hat{p}_{k,\gamma}(\mathbf{T}, \mathbf{Y}^{obs}) := \max_{\tau' \in C} \widehat{pval}_k(\mathbf{T}, \mathbf{Y}^{obs}, \tau') + \gamma$

2.4 Simulation

In this section, we present the Monte Carlo simulation design and results that assess the performance of the proposed testing procedures. We build on the Monte Carlo experimental designs of Ding et al. (2016). Specifically, we model network interference into their setup using a sparse adjacency matrix $\mathbf{A}$ where the number of edges or degrees of each node is five. For each unit $i = 1, \dots, N$, the relationship between the potential outcomes is

$$
\begin{aligned}
Y_i(1, \pi) &= Y_i(0, \pi) + \tau_i(\pi, x) \\
Y_i(0, \pi) &= u_i(\pi) \\
\tau_i(\pi, x) &= (1 + \psi_0 \pi + \psi_1 x) + \sigma_\tau \cdot Y_i(0, \pi) \; \forall \pi \in \mathbf{\Pi} \; \forall x \in \mathbb{X},
\end{aligned} \tag{2.34}
$$

where ψ_0, ψ_1 and σ_τ controls the different types of treatment effects heterogeneity. For instance, when $\psi_0 = \psi_1 = \sigma_\tau = 0$, H_0 in (2.3) holds, i.e., there are no systematic and idiosyncratic variations in treatment effects.

To assess the finite sample performance of the test statistics of H_0 and H_0^{Π}, we can ignore the pretreatment variables and let $\tau_i(\pi, x) = \tau_i(\pi) = (1 + \psi_0\pi) + \sigma_\tau \cdot Y_i(0, \pi)$. Also, for our simulation exercise, we let $u_i(\pi), i = 1 \dots N$ be i.i.d random variables drawn from either the normal or the log-normal distributions with variance 1, and mean π. Furthermore, we set $B = 150$, and the number of replications to 1000 for all the Monte Carlo exercises. In all the experiments, we use the following binary-valued exposure mapping

$$\pi_i(\mathbf{T}) := \mathbb{I}\bigg(\frac{\sum_{j=1}^{N} T_j A_{ij}}{\sum_{j=1}^{N} A_{ij}} > 0.5\bigg).$$

Focusing on H_0 in (2.3), we first set $\psi_0 = \psi_1 = 0$, $N = 200$ and $\epsilon = 0.20$. We display the rejection probabilities and FWER in Table 2.4 for the test statistics TS_k when H_0 is true and false using the CI and SS techniques. In particular, to compute the individual rejection probability when H_0 is true, we also set $\sigma_\tau = 0$. We report the rejection probabilities and the FWER at a 5% significance level for the normal and log-normal DGPs in the first rows of the two panels in Table 2.4. The other rows of each of the panels in Table 2.4 report the rejection probabilities under fixed alternatives (or when the null is false) by setting $\sigma_\tau \in \{0.5, 1.0, 1.5, 2.0\}$.

The results support the asymptotic validity of the CI and SS techniques on the CRI method under both DGPs. In general, the p-values of the CI technique are lower, which supports our conjecture that the CI technique produces conservative p-values. Note that, under both techniques, any multiple testing procedures that control the FWER are valid at the 5% level since FWER$\leq$ 0.097 when $\sigma_\tau = 0$. The rejection probabilities also tend to one under fixed alternatives that move further away from the null hypothesis. Therefore, the procedures are consistent and have non-negligible power.

In Table 2.5, we also report the rejection probabilities at a 5% significance level using the combined test statistic TS in (2.15) for H_0. We keep the same parameters we used for the individual test statistics. Similar to Table 2.4, we observe that the resulting p-values are valid under both DGPs for both techniques. In addition, both techniques are consistent and have non-negligible power. Overall, comparing the p-values, both testing approaches are competitive.

Next, we focus on the null hypothesis H_0^{Π} in (2.4). Here, we first set $\psi_0 = 1, \psi_1 = 0$, $N = 200$, and $\epsilon = 0.20$. To compute the rejection probability under the null hypothesis, we also set $\sigma_\tau = 0$. Again using the test statistics TS_k defined in (2.6), we report the rejection probabilities at a 5% significance level under the normal and log-normal DGPs in Table 2.6. As in Table 2.4, the p-

Table 2.4: Empirical Rejection Probabilities using the Individual Test Statistics TS_k's when H_0 is True and False with $\alpha = 0.05$

		Normal			Log-normal		
Techniques	σ_τ	$\Pr(\widehat{pval}_0 < \alpha)$	$\Pr(\widehat{pval}_1 < \alpha)$	FWER	$\Pr(\widehat{pval}_0 < \alpha)$	$\Pr(\widehat{pval}_0 < \alpha)$	FWER
	0.0	0.003	0.000	0.003	0.003	0.003	0.006
	0.5	0.291	0.296	0.504	0.024	0.022	0.045
CI	1.0	0.892	0.886	0.990	0.095	0.094	0.182
	1.5	0.994	0.996	1.000	0.224	0.192	0.371
	2.0	1.000	0.999	1.000	0.397	0.342	0.603
	0.0	0.008	0.016	0.024	0.061	0.008	0.068
	0.5	0.246	0.254	0.425	0.216	0.035	0.240
SS	1.0	0.701	0.704	0.916	0.459	0.122	0.518
	1.5	0.925	0.924	0.996	0.660	0.318	0.757
	2.0	0.990	0.976	1.000	0.771	0.445	0.869

Estimates are based on 1000 replications, with a simulation standard error of 0.00689 under H_0.

values using the CI techniques are valid but relatively lower under both DGPs. Both techniques are consistent and have non-negligible power. In Table 2.7, we report the rejection probabilities by applying the combined test statistic to H_0^{Π}. We set the same model parameters we used for the individual test statistics. The results indicate the validity of the p-values under both DGPs and both techniques. Also, the results show that under fixed alternatives, we have non-negligible statistical power for both DGPs and both techniques.

Table 2.6: Empirical Rejection Probabilities using the Individual Test Statistics TS_k's when H_0^{Π} is True and False with $\alpha = 0.05$

		Normal			Log-normal		
Techniques	σ_τ	$\Pr(\widehat{pval}_0 < \alpha)$	$\Pr(\widehat{pval}_1 < \alpha)$	FWER	$\Pr(\widehat{pval}_0 < \alpha)$	$\Pr(\widehat{pval}_0 < \alpha)$	FWER
	0.0	0.003	0.000	0.003	0.003	0.003	0.006
	0.5	0.281	0.286	0.488	0.024	0.022	0.045
CI	1.0	0.885	0.874	0.989	0.093	0.092	0.179
	1.5	0.993	0.995	1.000	0.217	0.191	0.365
	2.0	1.000	0.999	1.000	0.387	0.336	0.595
	0.0	0.013	0.031	0.044	0.042	0.025	0.067
	0.5	0.208	0.405	0.516	0.129	0.082	0.202
SS	1.0	0.596	0.886	0.955	0.312	0.264	0.492
	1.5	0.864	0.989	0.999	0.525	0.449	0.747
	2.0	0.944	0.997	1.000	0.670	0.640	0.884

Estimates are based on 1000 replications, with a simulation standard error of 0.00689 under H_0^{Π}.

Table 2.5: Empirical Rejection Probabilities using the Combined Test Statistic TS when H_0 is True and False with $\alpha = 0.05$

Techniques	σ_τ	Normal	Log-normal
CI	0.0	0.003	0.001
	0.5	0.516	0.012
	1.0	0.998	0.093
	1.5	1.000	0.282
	2.0	1.000	0.453
SS	0.0	0.005	0.018
	0.5	0.309	0.103
	1.0	0.896	0.328
	1.5	0.997	0.566
	2.0	0.999	0.739

Estimates are based on 1000 replications, with a simulation standard error of 0.00689 under H_0.

Table 2.7: Empirical Rejection Probabilities using the Combined Test Statistic TS when H_0^Π is True and False with $\alpha = 0.05$

Techniques	σ_τ	Normal	Log-normal
CI	0.0	0.001	0.001
	0.5	0.509	0.014
	1.0	0.996	0.093
	1.5	1.000	0.271
	2.0	1.000	0.438
SS	0.0	0.000	0.000
	0.5	0.095	0.020
	1.0	0.620	0.090
	1.5	0.933	0.236
	2.0	0.988	0.407

Estimates are based on 1000 replications, with a simulation standard error of 0.00689 under H_0^Π.

Finally, we focus on the null hypothesis $H_0^{X,\Pi}$ in (2.5). For this exercise, we set $\psi_0 = 1$ $\psi_1 = 1$, and $\epsilon = 0.008$. To compute the rejection probability when $H_0^{X,\Pi}$ is true, we also set $\sigma_\tau = 0$. In addition, we set $N = 300$ for the CI technique, and $N = 600$ [8] for the SS technique. Using the test statistics $\text{TS}_{k,l}$, we report the rejection probabilities at a 5% significance level under the normal and log-normal DGPs in Table 2.8. The tests are valid under both DGPs and for both techniques. Also, in Table 2.9, we report the rejection probabilities using the combined test statistic $\text{TS}^{X,\Pi}$ in (2.8),

[8] Using $N = 300$ for the SS technique, we are unable to compute test statistics for each exposure value and arm of pretreatment variable after splitting the sample.

using the same parameters as in Table 2.8. The results indicate the validity of the resulting p-values. Statistical power is non-negligible under fixed alternatives for both testing approaches.

Table 2.8: Empirical Rejection Probabilities using the Individual Test Statistics $TS_{k,l}$'s when $H_0^{X,\Pi}$ is True and False with $\alpha = 0.05$. NB: $\widehat{pval}_{kl} = \hat{p}_{kl}$.

		Normal					Log-normal				
Techniques	σ_τ	$\Pr(\hat{p}_{00} < \alpha)$	$\Pr(\hat{p}_{10} < \alpha)$	$\Pr(\hat{p}_{01} < \alpha)$	$\Pr(\hat{p}_{11} < \alpha)$	FWER	$\Pr(\hat{p}_{00} < \alpha)$	$\Pr(\hat{p}_{10} < \alpha)$	$\Pr(\hat{p}_{01} < \alpha)$	$\Pr(\hat{p}_{11} < \alpha)$	FWER
	0.0	0.002	0.000	0.002	0.001	0.005	0.000	0.000	0.001	0.001	0.002
	0.5	0.137	0.117	0.133	0.125	0.430	0.010	0.009	0.012	0.013	0.041
CI	1.0	0.649	0.613	0.626	0.632	0.984	0.064	0.060	0.055	0.033	0.202
	1.5	0.922	0.893	0.926	0.918	1.000	0.145	0.133	0.128	0.119	0.440
	2.0	0.992	0.982	0.992	0.984	1.000	0.250	0.215	0.251	0.209	0.657
	0.0	0.005	0.003	0.003	0.002	0.013	0.003	0.004	0.004	0.004	0.015
	0.5	0.227	0.219	0.213	0.241	0.646	0.034	0.036	0.026	0.029	0.119
SS	1.0	0.756	0.741	0.768	0.741	0.999	0.120	0.129	0.143	0.136	0.425
	1.5	0.966	0.954	0.956	0.953	1.000	0.306	0.287	0.295	0.301	0.760
	2.0	0.996	0.993	0.994	0.996	1.000	0.440	0.368	0.438	0.433	0.891

Estimates are based on 1000 replications, with a simulation standard error of 0.00689 under $H_0^{X,\Pi}$.

Table 2.9: Empirical Rejection Probabilities using the Combined Test Statistic $TS^{X,\Pi}$ when $H_0^{X,\Pi}$ is True and False with $\alpha = 0.05$

Techniques	σ_τ	Normal	Log-normal
	0.0	0.000	0.005
	0.5	0.238	0.042
CI	1.0	0.951	0.115
	1.5	0.999	0.273
	2.0	1.000	0.375
	0.0	0.000	0.000
	0.5	0.319	0.003
SS	1.0	0.977	0.090
	1.5	1.000	0.292
	2.0	1.000	0.512

Estimates are based on 1000 replications, with a simulation standard error of 0.00689 under $H_0^{X,\Pi}$.

2.5 Conclusion

This chapter proposes randomization tests for heterogeneous treatment effects when units interact on a single network. By incorporating the concept of network exposure mapping, we model network interference in the potential outcomes framework and considers three non-sharp null hypotheses representing different notions of homogeneous treatment effects. The chapter proposes a conditional randomization inference method to deal with the presence of multiple potential outcomes and

two procedures to overcome nuisance parameter issues. Overall, this chapter offers insights for researchers seeking to analyze heterogeneous treatment effects in a networked environment.

We may consider a few possible extensions. It is interesting to investigate goodness-of-fit test statistics like the Kolmogorov-Smirnov (KS) test statistic. In particular, can we extend the martingale transformation method applied to the KS test statistic by Chung and Olivares (2021)? Our immediate conjecture is that it is impossible without modifications due to the dependencies in network data sets. It is also interesting to extend the sample splitting technique to control for the randomness from sample splitting. A natural extension is a cross-fitting routine where we consider multiple independent splits of the data to approximate the distribution. Finally, it is worthwhile to examine the generalizability of the proposed conditioning method. In other words, can we apply the method to other hypotheses? We leave these questions for our future research.

Bibliography

Aronow, P. M. (2012). A general method for detecting interference between units in randomized experiments. *Sociological Methods & Research 41*(1), 3–16.

Aronow, P. M., C. Samii, et al. (2017). Estimating average causal effects under general interference, with application to a social network experiment. *The Annals of Applied Statistics 11*(4), 1912–1947.

Athey, S., D. Eckles, and G. W. Imbens (2018). Exact p-values for network interference. *Journal of the American Statistical Association 113*(521), 230–240.

Baird, S., J. A. Bohren, C. McIntosh, and B. Özler (2018). Optimal design of experiments in the presence of interference. *Review of Economics and Statistics 100*(5), 844–860.

Basse, G. W., A. Feller, and P. Toulis (2019). Randomization tests of causal effects under interference. *Biometrika 106*(2), 487–494.

Berger, R. L. and D. D. Boos (1994). P values maximized over a confidence set for the nuisance parameter. *Journal of the American Statistical Association 89*(427), 1012–1016.

Bitler, M. P., J. B. Gelbach, and H. W. Hoynes (2006). What mean impacts miss: Distributional effects of welfare reform experiments. *American Economic Review 96*(4), 988–1012.

Cai, J., A. De Janvry, and E. Sadoulet (2015). Social networks and the decision to insure. *American Economic Journal: Applied Economics 7*(2), 81–108.

Chung, E. and M. Olivares (2021). Permutation test for heterogeneous treatment effects with a nuisance parameter. *Journal of Econometrics 225*(2), 148–174.

Cox, D. R. (1958). Planning of experiments.

Crump, R. K., V. J. Hotz, G. W. Imbens, and O. A. Mitnik (2008). Nonparametric tests for treatment effect heterogeneity. *The Review of Economics and Statistics 90*(3), 389–405.

Ding, P., A. Feller, and L. Miratrix (2016). Randomization inference for treatment effect variation. *Journal of the Royal Statistical Society: Series B: Statistical Methodology*, 655–671.

Fisher, R. (1925). Statistical methods for research workers (oliver and boyd, edinburgh, scotland), 299.

Hájek, J. (1960). Limiting distributions in simple random sampling from a finite population. *Publications of the Mathematical Institute of the Hungarian Academy of Sciences 5*, 361–374.

Han, S., J. Owusu, and Y. Shin (2022). Statistical treatment rules under social interaction. *arXiv preprint arXiv:2209.09077*.

Lehmann, E. and J. P. Romano (2022). The general decision problem. In *Testing statistical hypotheses*, pp. 3–28. Springer.

Leung, M. P. (2020). Treatment and spillover effects under network interference. *Review of Economics and Statistics 102*(2), 368–380.

Li, X. and P. Ding (2017). General forms of finite population central limit theorems with applications to causal inference. *Journal of the American Statistical Association 112*(520), 1759–1769.

Liu, L. and M. G. Hudgens (2014). Large sample randomization inference of causal effects in the presence of interference. *Journal of the american statistical association 109*(505), 288–301.

Manski, C. F. (2013). Identification of treatment response with social interactions. *The Econometrics Journal 16*(1), S1–S23.

Sant'Anna, P. H. (2021). Nonparametric tests for treatment effect heterogeneity with duration outcomes. *Journal of Business & Economic Statistics 39*(3), 816–832.

Sävje, F., P. Aronow, and M. Hudgens (2021). Average treatment effects in the presence of unknown interference. *Annals of statistics 49*(2), 673.

Viviano, D. (2019). Policy targeting under network interference. *arXiv preprint arXiv:1906.10258*.

Wager, S. and S. Athey (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association 113*(523), 1228–1242.

Zhang, Y. and Q. Zhao (2022). What is a randomization test? *arXiv preprint arXiv:2203.10980*.

2.6 Appendix

To prove Theorem 2.3.1, we need the following Lemmata.

Lemma 2.6.1 (The probability integral transform theorem) *Assume a random variable A has a continuous distribution for which the cumulative distribution function (CDF) is F_A. Then the random variable B defined as $B = F_A(A)$ has a standard uniform distribution.*

The following lemma of Hájek (1960) states the central limit theorem in finite populations.

Lemma 2.6.2 (Central limit theorem in finite populations) *Let $\bar{y}_n$ be the average of a simple random sample of size n from a finite population $\{y_1, y_2, \ldots, y_N\}$. As $N \to \infty$, if*

$$\frac{1}{min(n, N-n)} \cdot \frac{\max_{1\leq i\leq N}(y_i - N^{-1}\sum_{i=1}^{N} y_i)^2}{(N-1)^{-1}\sum_{i=1}^{N}(y_i - N^{-1}\sum_{i=1}^{N} y_i)^2} \to 0 \tag{2.35}$$

then

$$\frac{(\bar{y}_n - N^{-1}\sum_{i=1}^{N} y_i)}{\text{Var}(\bar{y}_n)} \to N(0,1)$$

Note that the finite population asymptotic results are obtained under the hypothetical concept that there is an infinite sequence of finite populations with increasing sizes (Li and Ding, 2017). Finally, note that assumption of finite second order moments of $\{y_1, y_2, \ldots, y_N\}$ is sufficient for the condition in (2.35).

2.6.1 Proof of Theorem 2.3.1

proof 2.6.1 *For any k, if G_k denote the asymptotic null distribution of $\text{TS}_k(\mathbf{Y}_{F_k}(\Lambda(\mathbf{T})), \Lambda(\mathbf{T}), \mathbf{A}; \mathcal{T}_k, F_k)$ for any treatment vector $\Lambda(\mathbf{T}) \in \mathcal{T}_k$, then the p-value and its asymptotic counterpart are respectively,*

$$pval_k(\mathbf{T}, \mathbf{Y}^{obs}; \mathcal{T}_k, F_k) = \Pr(\text{TS}_k(\mathbf{Y}_{F_k}(\Lambda(\mathbf{T})), \Lambda(\mathbf{T}), \mathbf{A}; \mathcal{T}_k, F_k) \geq \text{TS}_k(\mathbf{Y}^{obs}, \mathbf{T}, \mathbf{A}; \mathcal{T}_k, F_k) | H_0) \text{ and,}$$
$$\lim_{N_k\to\infty} pval_k(\mathbf{T}, \mathbf{Y}^{obs}; \mathcal{T}_k, F_k) = G_k(\text{TS}_k(\mathbf{Y}^{obs}, \mathbf{T}^{obs}, \mathbf{A}; \mathcal{T}_k, F_k)).$$

Under Assumptions 2.2.1-2.3.1 and finite conditional variances of the potential outcomes, we can apply the central limit theorem in Lemma 2.6.2. Note that conditional on treatment and exposure values, realized outcomes are the same as the potential outcomes hence they are fixed (and no cross-sectional dependency exists).

Therefore, using Lemma 2.6.2, we can deduce that the sample mean for a given treatment and exposure value converges to a normal distribution. The sample variance for a given treatment

and exposure value will also weakly converge. Since the test statistic depend on the conditional variances, note that the asymptotic distribution of $\text{TS}_k(\mathbf{Y}_{F_k}(\Lambda(\mathbf{T})), \Lambda(\mathbf{T}), \mathbf{A}; \mathcal{T}_k, F_k)$ *coincides with that of the observed test statistic.*

Hence, for a large sample size, the observed test statistic $\text{TS}_k(\mathbf{Y}^{obs}, \mathbf{T}^{obs}, \mathbf{A}; \mathcal{T}_k, F_k)$ *also has the asymptotic null distribution* G_k.

Now, from Lemma 2.6.1, $pval_k(\mathbf{T}, \mathbf{Y}^{obs}; \mathcal{T}_k, F_k)$ *has a standard uniform distribution in the limit and as such,*
$\lim_{N_k \to \infty} \Pr(pval_k(\mathbf{T}, \mathbf{Y}^{obs}; \mathcal{T}_k, F_k) \leq \alpha) \leq \alpha$ *for any* $\alpha \in [0, 1]$ *as required.*

2.6.2 Proof of Theorem 2.3.2

proof 2.6.2 *Let* $\mathcal{U}^{est}$ *and* $\mathcal{U}^{inf}$ *be a random equal split of the data into estimation and inference sub-samples. Now, let* $\hat{\tau}^{est}$ *be an estimator of the sample average treatment effect conditional on the sub-sample* $\mathcal{U}^{est}$. *Assume that* $\hat{\tau}^{est}$ *is a consistent estimator*[9] *of the unconditional sample average treatment effect* τ. *Specifically,*

$$|\hat{\tau}^{est} - \tau| = o_p(1) \tag{2.36}$$

Next, for any k, *let* $G_{k|\hat{\tau}^{est}}$ *denote the asymptotic distribution of* $\text{TS}_k(\mathbf{Y}_{F_k}(\Lambda(\mathbf{T})), \Lambda(\mathbf{T}), \mathbf{A}; \mathcal{T}_k, F_k, \hat{\tau}^{est})$. *Note that this distribution is conditional on a given balanced sample split and the conditioning mechanism. Therefore, the probability is with respect to the randomness of* $\Lambda(\mathbf{T}) \in \mathcal{T}_k^{\mathcal{U}^{inf}}$, *where* $\mathcal{T}_k^{\mathcal{U}^{inf}}$ *is the subset of treatment assignment vectors of the inference sub-sample that meet the conditioning mechanism on the full sample.*

Under the null hypothesis (2.3), *as* $N \to \infty$ *(which by assumption implies that* $N_k \to \infty$, *for all* k*), the randomization hypotheses also hold. From Theorem 2.3.1, the limit laws of the test statistics are invariant to randomized treatments, i.e.,* $\text{TS}_k(\mathbf{Y}^{obs}, \mathbf{T}, \mathbf{A}; \mathcal{T}_k, F_k, \hat{\tau}^{est})$ *also has an asymptotic distribution* $G_{k|\hat{\tau}^{est}}$ *since by assumption,* $\mathbf{T} \in \mathcal{T}_k^{\mathcal{U}^{inf}}$. *Therefore, the p-value can be written as*

$$pval_k(\mathbf{T}, \mathbf{Y}^{obs}; \mathcal{T}_k, F_k, \hat{\tau}^{est}) = \Pr(\text{TS}_k(\mathbf{Y}_{F_k}(\Lambda(\mathbf{T})), \Lambda(\mathbf{T}), \mathbf{A}; \mathcal{T}_k, F_k, \hat{\tau}^{est}) \geq \text{TS}_k(\mathbf{Y}^{obs}, \mathbf{T}, \mathbf{A}; \mathcal{T}_k, F_k, \hat{\tau}^{est}) | H_0)$$
$$pval_k(\mathbf{T}, \mathbf{Y}^{obs}; \mathcal{T}_k, F_k, \hat{\tau}^{est}) = G_{k|\hat{\tau}^{est}}(\text{TS}_k(\mathbf{Y}^{obs}, \mathbf{T}^{obs}, \mathbf{A}; \mathcal{T}_k, F_k, \hat{\tau}^{est}))$$

From Lemma 2.6.1, $pval_k(\mathbf{T}, \mathbf{Y}^{obs}; \mathcal{T}_k, F_k, \hat{\tau}^{est})$ *has a standard uniform distribution in the limit. Combining this fact with the consistency assumption in* (2.36), *we have* $\lim_{N \to \infty} \Pr(pval_k(\mathbf{T}, \mathbf{Y}^{obs}; \mathcal{T}_k, F_k, \hat{\tau}^{est}) \leq \alpha) \leq \alpha$ *for any* $\alpha \in [0, 1]$ *as required.*

[9] We can show that the difference-in-means estimator is consistent in a design-based setting.

2.6.3 Proof of Theorem 2.3.3

proof 2.6.3 *Let τ_0 be the true value of the nuisance parameter τ, and let N_k be the subset of units in the population with observed exposure value equal to π_k. Then*

$$\begin{aligned}
\lim_{N_k\to\infty} \Pr(pval_{k,\gamma}(\mathbf{T}, \mathbf{Y}^{obs}; \mathcal{T}_k, F_k) \leq \alpha | H_0) = &\lim_{N_k\to\infty} \Pr(pval_{k,\gamma}(\mathbf{T}, \mathbf{Y}^{obs}; \mathcal{T}_k, F_k) \leq \alpha, \tau_0 \in \text{CI}_\gamma | H_0) \\
&+ \lim_{N_k\to\infty} \Pr(pval_{k,\gamma}(\mathbf{T}, \mathbf{Y}^{obs}; \mathcal{T}_k, F_k) \leq \alpha, \tau_0 \notin \text{CI}_\gamma | H_0) \\
\leq &\lim_{N_k\to\infty} \Pr\left(\sup_{\tau' \in \text{CI}_\gamma} pval_k(\tau') + \gamma \leq \alpha, \tau_0 \in \text{CI}_\gamma | H_0 \right) \\
&+ \lim_{N_k\to\infty} \Pr(\tau_0 \notin \text{CI}_\gamma | H_0) \\
\leq &\lim_{N_k\to\infty} \Pr(pval_k(\tau_0) \leq \alpha - \gamma, \tau_0 \in \text{CI}_\gamma | H_0) \\
&+ \lim_{N_k\to\infty} \Pr(\tau_0 \notin \text{CI}_\gamma | H_0) \\
\leq &\alpha - \gamma + \gamma = \alpha
\end{aligned}$$

Note that a large N_k is only a requirement for the randomization hypothesis. In other words, by construction, as $N_k \to \infty$, the number of focal units uniformly increases for each treatment vector which implies that the randomization hypothesis will hold. The first equality is an application of the law of total probability. The first inequality (second line) is also a straightforward use of the relationship between marginal and joint probabilities for the second term. The second inequality (third line) uses the fact that $\sup_{\tau' \in \text{CI}_\gamma} pval_k(\tau')$ stochastically dominates $pval_k(\tau_0)$. The last inequality uses Theorem 2.3.1 and the fact that CI_γ is a valid confidence interval.

Chapter 3

Statistical Treatment Rules under Social Interaction

3.1 Introduction

One of the most crucial questions for a policy maker is how to assign a treatment to an individual or a group. For example, during the COVID-19 pandemic, each government has tried to find an effective order of vaccination. Recently, statistical treatment rules based on the decision theoretic framework have received much attention in treatment evaluation studies (for a general review, see Manski (2004, 2021) and Hirano and Porter (2020)). Compared to the conventional approaches based on the point estimation and inference procedures, statistical treatment rules make it possible to evaluate a broader range of treatment rules, which includes a direct map from data to an action. Despite active research in this area, most studies focus on the individualistic treatment response and we have limited results for the case where treatment outcomes depend on each other. As we can see from the vaccination example, it is important in many empirical settings to consider dependent treatment outcomes

In this chapter, we study a treatment assignment rule in the presence of treatment outcome dependency. In addition to the problem of vaccination, there are many applications that a policy maker has to weigh dependent treatment outcomes. Heckman, Lochner, and Taber (1999) evaluate the effect of a tuition reduction policy in the UK in a general equilibrium framework. They show that ignoring the outcome dependency over-estimates the effect of the policy on college enrollment more than 10 times. Duflo (2004) also argues that even a randomized control trial faces a challenge in scaling up to a larger level because of the general equilibrium effects or, more generally, dependent treatment outcomes. Using Danish data on a large job assistance program, Gautier et al. (2018) show that the unemployed who are not selected in the program spend more time in job search than those who look for a job in provinces without such a program. Thus, the outcome of the untreated depends on that of the treated, and the treatment evaluations assuming independent treatment outcomes can mislead a policy maker.[1]

We investigate this problem in the framework of the statistical decision theory. Treatment outcomes are allowed to depend on each other in a flexible way. We aim to construct a treatment assignment rule under the minimax regret approach and to characterize it. Thus, a treatment choice using sample data, i.e. a statistical decision rule, is the main object of interest in this chapter. Having in mind a large-scale policy implementation, we assume that the links between individuals (network information) is unknown. Instead, we impose a shape restriction on treatment response functions following Manski (2013). Specifically, we assume *anonymous interactions*, which implies that the treatment response of an individual does depend on the treatment status of others but is invariant of

[1] See also Beaman (2012), Bursztyn et al. (2014), and Duflo and Saez (2003) for additional examples.

the identity of other individuals. In other words, it is independent of the permutation of the treatment assignments on others. In the job assistance program above, for instance, this condition implies that the negative effect of the policy on the untreated only depends on the total size of people who receive the benefit of the job assistance program. This assumption provides a good approximation of the world with a large-scale policy implementation, and it makes both theoretical and empirical analyses feasible by reducing the domain of the response function substantially.

We define the sampling process carefully following the statistical decision theory framework. It contrasts to the standard *individualistic* treatment effect model in that our process represents both the treatment status variable and the outcome variables as a vector. The dimension of the vector is the same as the number of different treatment ratios in the target population. We adopt the minimax regret approach to handle the underlying ambiguity of the data generating process. We propose an intuitive decision rule called the multinomial empirical success (MES) rule that extends the empirical success rule in Manski (2004) to the current setup. We investigate the properties of the MES rule followed by the possible applications.

The main contributions of this chapter are summarized as follows. First, we prove that the MES rule achieves the asymptotic optimality for the minimax regret criterion. Using the structure of the finite action problem in statistics literature, it extends the seminal optimality result in Hirano and Porter (2009) to multiple treatments. Second, we derive the non-asymptotic bounds of the expected welfare and the maximum regret under the MES rule. We also provide two applications on how these bounds can be used: (i) designing an optimal sampling procedure, and (ii) computing the sufficient sample size to allow additional covariates in the treatment rule.

The rest of the chapter is organized as follows. We finish this section by reviewing related literature. In section 3.2 we provide the main framework of the analysis. In section 3.3 we define the MES rule and derive the upper bounds of the maximum regret. We also provide two applications of these bounds. We provide some concluding remarks in section 3.4. All proofs and technical details are deferred to the appendix.

3.1.1 Related Literature

In the seminal work of Manski (2004), he considers the statistical decision theory in the context of heterogeneous treatment rules. He proposes the empirical success rule and derives the finite sample bounds of the minimax regret. Stoye (2009) characterizes the minimax regret rule using the game theoretic approach and shows that the empirical success rule is a good approximation of the minimax regret rule under certain sampling processes. Hirano and Porter (2009) apply the limit experiment framework to develop large sample approximations to the statistical treatment rules.

Kitagawa and Tetenov (2018) propose the empirical welfare maximization (EWM) method that selects the treatment rule maximizing the sample analogue of the social average welfare. Athey and Wager (2021) propose a doubly robust estimation procedure for the EWM problem and show the rate-optimal regret bounds. Mbakop and Tabord-Meehan (2021) consider a large class of admissible rules and propose a penalized EWM method that chooses the optimal size of the policy class. Manski and Tetenov (2016, 2019) argue to design clinical trials based on the goal of statistical treatment rules rather than on the statistical power of a hypothesis test. Motivated by a risk-averse policy maker, Manski and Tetenov (2007) and Kitagawa, Lee, and Qiu (2022) propose nonlinear transformations of welfare and regret.

Manski (2013) studies identification of treatment effects with social interaction. To make the problem feasible, he proposes possible approximation methods including *anonymous interaction*, which will be explained in detail later. Manski (2009) analyzes statistical treatment rules under the anonymous interaction assumption and the shape restriction on the mean welfare function. Viviano (2019) proposes the network empirical welfare maximization method under the anonymous interaction assumption among those in the first-degree neighbor. However, our approach is different from his since it does not require heavy computation to solve an empirical optimization problem.

3.2 Framework

We consider the following framework based on Manski (2004) and Stoye (2009). Consider a social planner who assigns a binary treatment $T \in \{0, 1\}$ to each individual j in a heterogeneous population J. The population is divided into mutually exclusive and exhaustive groups based on observed characteristics (e.g. high school graduate vs. college graduate). Let $g \in \{1, 2, \ldots, G\}$ be the index of a group and n_g be the (population) size of group g. Individual j in group g has a response function $y_{jg} : \{0, 1\} \times \{0, 1\}^{n_g - 1} \mapsto [0, 1]$ that maps each possible group treatment vector $\mathbf{t} = (t_1, \ldots, t_{n_g}) \in \{0, 1\}^{n_g}$ into an outcome in $[0, 1]$. Thus, we can write $y_{jg}(\mathbf{t}) = y_{jg}(t_j, \mathbf{t}_{-j})$, where t_j is the treatment assigned to individual j and $\mathbf{t}_{-j}$ represents the treatment vector for individuals in the same group excluding person j's treatment assignment. This response function generalizes the individualistic treatment in a way that the spillover effect is allowed inside the same group (e.g. segmented labor markets). Note that the model allows the most flexible interactions when the whole population is categorized as a single group. The range of $[0, 1]$ is a simple normalization and any bounded outcome space can be allowed. For notational simplicity, we consider a single group from now on and drop the subscript g unless it causes any confusion.

We consider a probability space (J, Σ, P_J). The population J is dense in the sense that $P_J(\{j\}) = 0$,

for all $j \in J$. The social planner cannot distinguish members of J. Therefore, we can consider the model as an induced random process, $Y(\mathbf{t})$, which is a potential outcome depending not only on individual treatment status, t_j, but on possible treatments of other members, $\mathbf{t}_{-j}$. Given the large size of the population J, this random process in the most general structure is intractable. Following the social interaction literature, we impose the following assumption.

Assumption 3.2.1 (Anonymous Interactions, Manski (2013)) *The outcome of individual j is invariant with respect to permutations of the treatments received by other members of the group.*

Assumption 3.2.1 implies that a treatment ratio is a sufficient statistic for $\mathbf{t}_{-j}$. Let $\pi(\mathbf{t})$ be a treatment ratio of treatment vector $\mathbf{t}$. Then, for two treatment vectors $\mathbf{t} \neq \mathbf{t}'$ such that $\mathbf{t} = (t_j, \mathbf{t}_{-j})$ and $\mathbf{t}' = (t_j, \mathbf{t}'_{-j})$, Assumption 3.2.1 implies that

$$y_j(\mathbf{t}) = y_j(\mathbf{t}') \text{ if } \pi(\mathbf{t}) = \pi(\mathbf{t}').$$

Therefore, the outcome of a treatment $\mathbf{t}$ depends on individual's treatment status t_j and $\pi(\mathbf{t})$, and we can rewrite the response function $y_j(\mathbf{t})$ as $y_j(t_j, \pi(\mathbf{t}_{-j})) : \{0,1\} \times \Pi \mapsto [0,1]$, where $\Pi := [0,1]$. The potential outcome processes now become $(Y_0(\pi), Y_1(\pi))$ whose distribution is $P_Y(Y_0(\pi), Y_1(\pi))$. Note that the distribution P_Y can be constructed from P_J given the response function $y_j(\cdot)$.

The distribution P_Y is identified with a state of the world $\theta \in \Theta$ that is unknown to the policy maker. Note that $\{P_{Y,\theta}(Y_0(\pi), Y_1(\pi)) : \theta \in \Theta\}$ is composed of all possible distributions on the outcome space $[0,1]^2$ for each $\pi \in \Pi$. To make the main arguments clear, we impose an additional assumption that the set Π is discrete.

Assumption 3.2.2 (Discrete Choice Set) *Let π be the fraction of treated individuals in a group. The support of π denoted by $\mathbf{\Pi}$ is a discrete set of finite elements.*

Assumption 3.2.2 is suitable to many applied settings since the treatment ratio set may be constrained exogenously for ethical, budgetary, equity, legislative or political reasons. In addition, this is a practical assumption when experiments are costly to implement at all feasible treatment ratios. The assumption could also provide a good approximation if $\mathbf{\Pi}$ is a continuous interval, but outcome function y_j is smooth in π

We provide the following examples.

Example 3.2.1 (Job placement assistantship program) *Crépon et al. (2013) design a two-stage randomized experiment to evaluate the direct and displacement impacts of job placement assistance (JPA) on the labor market outcomes of young, educated job seekers in France. Individuals are*

organized in segmented labor markets (e.g. cities) and five treatment ratios (0%, 25%, 50%, 75%, and 100%) are considered. An individual's labor market outcome depends not only on his/her treatment status but on the treatment ratio (fraction of individuals who received the JPA in their labor market).

Example 3.2.2 (Cholera vaccine coverage) *Root et al. (2011) analyze data from a field trial in Bangladesh to assess the evidence of indirect protection from cholera vaccines when vaccination coverage rates varies according to the social network. Households are organized into independent groups using kinship connections. Vaccine coverage rate is discretized into the following ranges:* $(0, 27.2\%]$*,* $(27.2, 40.0\%]$*,* $(40.0 - 50.0\%]$*,* $(50.0\% - 62.5\%]$*, and* $(62.5\%, 100\%]$.

We now turn our attention to a random sample that helps the policy maker infer the state of the world θ. Let $\mathbf{\Pi} = \{\pi_1, \pi_2, \ldots, \pi_K\}$. The experiment generates a sample space $\Omega := (\{0,1\} \times [0,1])^N$, where $N := \sum_{k=1}^{K} N_k$ and N_k is the subgroup size of an experiment with a treatment ratio π_k. A typical element of Ω is represented by

$$\omega^n := \{(t_{n_1}(\pi_1), y_{n_1}(\pi_1))_{n_1=1}^{N_1}, (t_{n_2}(\pi_2), y_{n_2}(\pi_2))_{n_2=1}^{N_2}, \ldots, (t_{n_K}(\pi_K), y_{n_K}(\pi_K))_{n_K=1}^{N_K}\}.$$

Conditional on the treatment realization $t_{n_k}(\pi_k)$, $y_{n_k}(\pi_k)$ is an independent realization of $Y_t(\pi_k)$ for $t = 0, 1$. Therefore, it helps a policy maker to infer the state of the world θ. To make notation simple, we assume the equal subgroup size, $N_1 = \cdots = N_K = N/K$, and ω^n is composed with N/K-copies of $\omega_i := \{(t_i(\pi_1), y_i(\pi_1)), \ldots, (t_i(\pi_K), y_i(\pi_K))\}$.

The policy maker constructs a statistical treatment rule $\delta : \Omega \mapsto \mathbf{\Pi}$ that maps a sample realization ω^n onto a treatment assignment ratio $\pi \in \mathbf{\Pi}$. Recall that we restrict our attention to a single group in this framework but the statistical treatment rule can be group-specific when there are multiple groups.

The expected outcome (or social welfare) given the statistical treatment rule δ and the state θ is

$$u(\delta, \theta) := \int U(\delta(\omega^n), \theta) dQ_\theta^n \tag{3.1}$$

$$= \sum_{k=1}^{K} U(\pi_k, \theta) \Pr(\delta(\omega^n) = \pi_k; \theta), \tag{3.2}$$

where Q_θ is a distribution of ω^n given state θ, $U(\pi, \theta) := (1 - \pi) \cdot E_\theta[Y_0(\pi)] + \pi \cdot E_\theta[Y_1(\pi)]$ is the expected outcome (or social welfare) for any given treatment ratio π in state θ, and $E_\theta[Y_t(\pi)]$ is the mean potential outcome of treatment status t given θ and π. Note that the potential outcome variable $Y_t(\pi)$ depends on the treatment of others through π. This point becomes clearer if we compare the

expected outcome in (3.1) with that of the individualistic treatment model (e.g. Stoye (2009)). When there is no social interaction, the mean potential outcome is independent of the group treatment ratio π, i.e. $E_\theta[Y_t(\pi)] = E_\theta[Y_t]$. Then, the expected outcome in (3.1) becomes

$$\begin{aligned}\int U(\delta(\omega^n), \theta) dQ_\theta^n &= \int \{(1 - \delta(\omega^n))E_\theta[Y_0] + \delta(\omega^n)E_\theta[Y_1]\} \, dQ_\theta^n \\ &= E_\theta[Y_0]\left(1 - \int \delta(\omega^n) dQ_\theta^n\right) + E_\theta[Y_1] \int \delta(\omega^n) dQ_\theta^n \\ &\equiv \mu_0(1 - E_\theta[\delta(\omega)]) + \mu_1 E_\theta[\delta(\omega)],\end{aligned}$$

where the last line is equal to the expected outcome in Stoye (2009) using his notation.

It is interesting to compare our framework to the individualistic multiple-treatment design. Given the finite number of treatment ratios, one might want to interpret the framework in terms of K different individual treatments without any social interaction: e.g. define $Y_1 := Y_1(\pi_1), Y_2 := Y_1(\pi_2), \ldots, Y_K := Y_1(\pi_K)$ and set $(Y_0, Y_1, \ldots, Y_K)$ as a vector of potential outcomes. However, this multiple-treatment design does not capture the feedback effect of the social interaction for any non-treated individual. Note that $Y_0(\pi)$ still depends on the treatment ratio π in our framework, which is not embedded in the potential outcome vector $(Y_0, Y_1, \ldots, Y_K)$ of the standard multiple-treatment design.

The decision problem is to find a statistical treatment rule that maximizes the expected outcome function $u(\delta, \theta)$. However, there exists ambiguity about the state of the world, and we need some decision criteria for unknown θ. In this chapter we adopt the minimax regret rule following Manski (2004) and Stoye (2009). The regret function of δ given state θ is defined as

$$R(\delta, \theta) := \max_{d \in D} u(d, \theta) - u(\delta, \theta), \tag{3.3}$$

where D is a set of all possible statistical treatment rule. The minimax regret solution of the decision problem is defined as

$$\delta^* := \arg\min_{\delta \in D} \sup_{\theta \in \Theta} R(\delta, \theta). \tag{3.4}$$

3.3 Multinomial Empirical Success Rule

In this section we propose a feasible statistical decision rule and characterize it by the non-asymptotic bounds on the maximum regret. To show the main idea, we keep focusing on a homogeneous population in a single group. The results are extended into the heterogeneous population in a single

group case in section 3.3.2 and we show how they can be used to determine the proper conditioning variables set.

It is difficult to attain the optimal statistical treatment rule by solving (3.4) directly since $R(\delta, \theta)$ involves integration over finite sample distributions. As an alternative, researchers may propose possible statistical treatment rules and analyze whether they achieve the optimal regret level. One of the popular rules is an empirical success rule, which substitutes empirical success rates for the population counterparts.

We propose such an empirical success rule suitable for the proposed setup. To focus on our main arguments, we restrict our attention to continuous outcomes which implies a strict ordering of the estimates for $U(\pi, s)$ for all $\pi \in \mathbf{\Pi}$. We define our multinomial empirical success (MES) rule as follows:

$$\delta^{MES}(\omega) := \sum_{k=1}^{K} \pi_k \cdot \mathbb{1}\left(\hat{U}(\pi_k) > \max_{\pi \in \mathbf{\Pi}_{-k}} \hat{U}(\pi)\right), \tag{3.5}$$

where $\mathbf{\Pi}_{-k} := \mathbf{\Pi} \setminus \{\pi_k\}$ and

$$\begin{aligned}\hat{U}(\pi_k) &:= (1 - \pi_k) \cdot \hat{E}_{P_\theta}[Y_0(\pi_k)] + \pi_k \cdot \hat{E}_{P_\theta}[Y_1(\pi_k)] \\ &= (1 - \pi_k) \cdot \frac{\sum_{n_k=1}^{N_k} Y_{n_k}(\pi_k) \cdot \mathbb{1}(T_{n_k}(\pi_k) = 0)}{\sum_{n_k=1}^{N_k} \mathbb{1}(T_{n_k}(\pi_k) = 0)} + \pi_k \cdot \frac{\sum_{n_k=1}^{N_k} Y_{n_k}(\pi_k) \cdot \mathbb{1}(T_{n_k}(\pi_k) = 1)}{\sum_{n_k=1}^{N_k} \mathbb{1}(T_{n_k}(\pi_k) = 1)}.\end{aligned} \tag{3.6}$$

Note that, using the convention $0 \cdot \infty = 0$, we define $\hat{U}(0) = N_1^{-1} \sum_{n_1=1}^{N_1} Y_{n_1}(0)$ when $\pi_1 = 0$. Similarly, $\hat{U}(1) = N_K^{-1} \sum_{n_K=1}^{N_K} Y_{n_1}(1)$ when $\pi_K = 1$.

We have a few remarks on the proposed statistical decision rule. First, we call the rule in (3.5) as a Multinomial Empirical Success (MES) rule to emphasize the multinomial choice set in the setting. Second, we estimate $E_{P_\theta}[Y_t(\pi)]$ by using the empirical measure that depends on the unknown state θ of the world. Thus, both $\hat{U}(\pi_k)$ and the outcome of $\delta^{MES}(\cdot)$ depend on θ although it is not included as an argument explicitly. Third, the MES rule encompasses the (unconditional) empirical success rule in Manski (2004). Let $\Pi = \{0, 1\}$ with $\pi_1 = 0$ and $\pi_2 = 1$. Then, the MES rule becomes

$$\begin{aligned}\delta^{MES}(\omega) &= 0 \cdot \mathbb{1}\left(\hat{U}(0, \theta) > \hat{U}(1, \theta)\right) + 1 \cdot \mathbb{1}\left(\hat{U}(0, \theta) < \hat{U}(1, \theta)\right) \\ &= \mathbb{1}\left(\frac{1}{N_1} \sum_{n_1=1}^{N_1} Y_{n_1}(0) < \frac{1}{N_2} \sum_{n_2=1}^{N_2} Y_{n_1}(1)\right),\end{aligned}$$

which is the empirical success rule in Manski (2004).

We next evaluate the expected outcome in (3.2) using the MES rule in (3.5):

$$\begin{aligned} u(\delta^{MES}, \theta) &= \sum_{k=1}^{K} \Pr(\delta^{MES}(\omega^n) = \pi_k) \cdot U(\pi_k, \theta) \\ &= \sum_{k=1}^{K} \Pr\left(\hat{U}(\pi_k) > \max_{\pi \in \mathbf{\Pi}_{-k}} \hat{U}(\pi) \right) \cdot U(\pi_k, \theta) \\ &= \sum_{k=1}^{K} \Pr\left(\bigcap_{j=1, j\neq k}^{K} \{\hat{U}(\pi_k) > \hat{U}(\pi_j)\} \right) \cdot U(\pi_k, \theta). \end{aligned}$$

As we discussed above, $u(\delta^{MES}, \theta)$ is intractable since it involves all possible finite sample distributions. However, building on Manski (2004), we can construct bounds for the expected outcome with the MES rule as follows.

Theorem 3.3.1 *Fix $\theta \in \Theta$. Let $\mathbf{\Pi} = \{\pi_1, \ldots, \pi_K\}$, $\Delta_{kl} := |U(\pi_k, \theta) - U(\pi_l, \theta)|$ for $k, l = 1, \ldots, K$, and $\pi_{M^*} := \arg\max_{\pi \in \mathbf{\Pi}} U(\pi, \theta)$. Then, the following inequality holds:*

$$U(\pi_{M^*}, \theta) - \sum_{k=1}^{K} \exp\left[-2\Delta_{M^*k}^2 \{A_k + A_{M^*}\}^{-1} \right] \cdot \Delta_{M^*k} \leq u(\delta^{MES}, \theta) \leq U(\pi_{M^*}, \theta), \tag{3.7}$$

where $A_k := (1 - \pi_k)^2 N_{k0}^{-1} + \pi_k^2 N_{k1}^{-1}$, $A_{M^*} := (1 - \pi_{M^*})^2 N_{M^*0}^{-1} + \pi_{M^*}^2 N_{M^*1}^{-1}$, and N_{kt} denotes the number of individuals in the sample with $\pi = \pi_k$ and $T = t$.

It is worth comparing these bounds with those in Proposition 1 of Manski (2004). Note that both frameworks allow the potential outcome distributions to vary across some indexing variables. For example, the potential outcomes in Manski (2004) depend on exogenous conditioning variables X, i.e. heterogeneous treatment effects over X. However, we focus on the dependence of the potential outcomes on the choice variable π. They look similar from the mathematical perspective, but the implications are quite different since the result in this chapter allows the effect of social interaction. This point becomes clearer when we extend the model to the case that includes additional conditioning variables.

We further investigate the finite sample penalty of the lower bound in (3.7), which measures the possible difference of $u(\delta^{MES}, \theta)$ from the ideal solution $U(\pi_{\pi_{M^*}}, \theta)$. First, the penalty converges to zero at the exponential rate as N_{tk} increases uniformly for all $t \in \{0, 1\}$ and $k \in \{1, \ldots, K\}$. Second, the penalty is maximized when $\Delta_{M^*k} = \{A_k + A_{M^*}\}^{1/2}/2$ for each $k \neq M^*$. Thus, we can compute the

upper bound of the penalty as follows:

$$\sum_{k=1}^{K} \exp\left[-2\Delta^2_{M^*k}\{A_k + A_{M^*}\}^{-1}\right] \cdot \Delta_{M^*k} \leq \frac{1}{2} \cdot e^{-\frac{1}{2}} \sum_{k=1, k\neq M^*}^{K} \{A_k + A_{M^*}\}^{\frac{1}{2}}. \tag{3.8}$$

Third, it is interesting to investigate the relationship between the cardinality of Π denoted by K and the penalty size. Consider the following example of two possible choice sets Π_1 and Π_2 such that $\Pi_1 \subset \Pi_2$. Let π_{M^*} be the optimal solution of Π_1. If π_{M^*} is also the optimal solution of Π_2, then Π_2 has a larger penalty than Π_1. However, if the optimal solution of Π_2 denoted by $\pi_{M^{**}}$ is different from π_{M^*}, then Π_2 may have a smaller penalty than Π_1. Note that $\Delta_{M^{**}k} > \Delta_{M^*k}$ for all $k \in \Pi_1$ and that there may exits some $k \in \Pi_1$ such that $\exp[-2\Delta^2_{M^{**}k}\{A_k + A_{M^{**}}\}^{-1}] < \exp[-2\Delta^2_{M^*k}\{A_k + A_{M^*}\}^{-1}]$. Therefore, a larger choice set may improve the finite sample lower bound only if it contains a better welfare outcome. Finally, we investigate the uniform bound of the regret function over θ. The upper bounds of the regret function with δ^{MES} is represented in terms of the penalty:

$$\begin{aligned} 0 \leq \sup_{\theta\in\Theta} R(\delta^{MES}, \theta) &\leq \sup_{\theta\in\Theta} \left\{ \sum_{k=1}^{K} \exp\left[-2\Delta^2_{M^*k}\{A_k + A_{M^*}\}^{-1}\right] \cdot \Delta_{M^*k} \right\} \\ &\leq \sup_{\theta\in\Theta} \left\{ \frac{1}{2} \cdot e^{-\frac{1}{2}} \sum_{k=1, k\neq M^*}^{K} \{A_k + A_{M^*}\}^{\frac{1}{2}} \right\}. \end{aligned}$$

Different from the result in Manski (2004), A_{M^*} in the right hand side depends on θ since π_{M^*} is defined in terms of $U(\pi, \theta)$. Therefore, we need an additional step to achieve the uniform bound. Let $\overline{A} := \max_{k\in\{1,\ldots,K\}} A_k$. Note that $\overline{A} \geq A_{M^*}$ and that $\overline{A}$ is independent of θ. Then, the desired uniform bound is achieved as follows:

$$0 \leq \sup_{\theta\in\Theta} R(\delta^{MES}, \theta) \leq \frac{1}{2} \cdot e^{-\frac{1}{2}} \sum_{k=1, A_k\neq\overline{A}}^{K} \{A_k + \overline{A}\}^{\frac{1}{2}}. \tag{3.9}$$

These finite sample bounds give us two useful applications. First, we apply this bound to solve the quasi-optimal experiment design problem. Second, we can extend the bound to the covariate dependent treatment rule and determine the minimum sample size to adopt a finer covariate set as in Manski (2004). We provide these applications in the following two subsections.

3.3.1 Application 1: Quasi-optimal Experiment Design

We study the optimal experiment design problem under interference using the upper bound of the maximum regret. Specifically, we focus on the randomized saturation design which is composed of two-stage randomized experiments (for example, see Baird et al. (2018)). Suppose that we are given many clusters. In the first stage, we assign different treatment ratios in Π to each cluster *randomly* according to a probability distribution f. In the second stage, a binary treatment is assigned to each member of a cluster according to a treatment ratio assigned in the previous stage. Therefore, the randomized saturation design is fully characterized by a pair (Π, f) and it encompasses other designs like clustered, block, and partial population designs commonly employed under interference.

We now consider an experiment design problem that minimizes the maximum regret. We cannot compute the exact regret function because of the ambiguity in θ. Instead, we reformulate the problem as minimizing the feasible upper bound of the regret in (3.9).

Recall that N denotes the total sample size over all clusters and $\Pi = \{\pi_1, \pi_2, \cdots, \pi_K\}$ be a finite set of treatment ratios. Since Π is a finite set, we can write $f = \{(\alpha_1, \alpha_2, \cdots \alpha_K) : \sum_{k=1}^{K} \alpha_k = 1\}$, where α_k is a probability mass of assigning π_k. The subsample sizes can be written in terms of the treatment ratios and their corresponding probabilities: $N_{k0} = (1 - \pi_k)\alpha_k N$ and $N_{k1} = \pi_k \alpha_k N$ for all $k = 1, 2, \ldots, K$. Then, for each A_k, we have

$$\begin{aligned} A_k &= \frac{(1-\pi_k)^2}{N_{k0}} + \frac{\pi_k^2}{N_{k1}} \\ &= \frac{(1-\pi_k)}{\alpha_k N} + \frac{\pi_k}{\alpha_k N} \\ &= \frac{1}{\alpha_k N}, \end{aligned}$$

which makes the optimization problem simple. Without loss of generality, let $\overline{A} = A_1$. We substitute A_k in (3.9) and drop all irrelevant variables to get

$$\begin{aligned} &\min_{\{\alpha_k\}_{k=1}^K} \sum_{k=2}^{K} \left(\frac{1}{\alpha_1 N} + \frac{1}{\alpha_k N} \right)^{1/2} \\ &\text{subject to} \quad \sum_{k=1}^{K} \alpha_k = 1 \\ &\qquad\qquad \alpha_1 \leq \alpha_k \text{ for } k = 2, \ldots, K. \end{aligned}$$

Solving this optimization problem, we derive the quasi-optimal design of equal α_k^* ($\alpha_k^* = 1/K$)

only when $K = 2$. It is worthwhile to note that Baird et al. (2018) derive the optimal randomized saturation design based on the statistical power but we focus on the maximum regret directly (see Manski and Tetenov (2016) for further discussion).

3.3.2 Application 2: Covariate-dependent Treatment Rules

In this section, we extend the model and consider covariate-dependent treatment rules. We first introduce new notation. Let X be a vector of covariates. In the similar spirit of Assumption 3.2.2, we restrict our attention to discrete and finite covariates. Then, we can vectorize the possible outcomes of X and partition the population into L different subsets denoted by $\mathcal{X} := \{x_1, \ldots, x_L\}$. To make notation simple, we assume a common domain of treatment ratios Π for each x_l[2]. We define a statistical treatment rule as $\delta(x, \omega^n) : \mathcal{X} \times \Omega \mapsto \Pi$. Let $\boldsymbol{\pi} := (\pi_1, \ldots, \pi_L)'$ be a vector of treatment assignment ratios, where π_l is applied to subgroup $x_l \in \mathcal{X}$. Let $\boldsymbol{p}$ be a vector of population subgroup proportions. Then, $\bar{\pi} := \boldsymbol{p}'\boldsymbol{\pi}$ becomes the unconditional treatment ratio. Under Assumption 3.2.1, the response function can be rewritten as $y_j(t_j, \bar{\pi})$.

Given $\boldsymbol{\pi}$ and θ, the outcome of the subgroup with covariate x_l is

$$U_l(\boldsymbol{\pi}, \theta) := (1 - \pi_l) \cdot E_\theta \left[Y_0(\bar{\pi}) | X = x_l \right] + \pi_l \cdot E_\theta \left[Y_1(\bar{\pi}) | X = x_l \right]. \tag{3.10}$$

Note that U_l is affected by the treatment ratios of other covariate types through $\bar{\pi}$ as well as its own ratio π_l. Let $\boldsymbol{\delta}(\omega^n) := (\delta(x_1, \omega^n), \ldots, \delta(x_L, \omega^n))$ be a vector of statistical treatment rules over $\mathcal{X}$ when sample ω^n is realized, i.e. $\boldsymbol{\delta}(\omega^n) : \Omega \mapsto \Pi^L$. The expected outcome of the whole population is defined by the weighted sum of U_l:

$$u(\boldsymbol{\delta}, \theta) := \sum_{l=1}^{L} \left[\int U_l(\boldsymbol{\delta}(\omega^n), \theta) dQ_\theta^n \right] \Pr(X = x_l). \tag{3.11}$$

If $\Pr(X = x_l; \theta) = 1$ for some l, then $\pi_l = \bar{\pi} \equiv \pi$, $L = 1$ and $u(\delta, \theta) = \int U(\delta(\omega^n), \theta) dQ_\theta^n$. Therefore, the expected outcome becomes equation (3.1), where there exists a single type of population.

Similar to (3.4), we can define the minimax regret solution of the decision problem as

$$\boldsymbol{\delta}^* := \arg\min_{\boldsymbol{\delta} \in \boldsymbol{D}} \sup_{\theta \in \Theta} R(\boldsymbol{\delta}, \theta),$$

where $R(\boldsymbol{\delta}, \theta) := \max_{\boldsymbol{d} \in \boldsymbol{D}} u(\boldsymbol{d}, \theta) - u(\boldsymbol{\delta}, \theta)$ is a regret function. Since the expected welfare with

[2]We can allow different assignment ratio sets at the cost of extra notation, e.g. $\Pi := \cup_{l=1}^{L} \Pi_l$, where $\Pi_l := \{\pi_1, \ldots, \pi_{K_l}\}$ is the set of assignment ratios for x_l.

covariate x_l is affected by the treatment assignment ratios of other covariates $x_m \neq x_l$, we need to find the decision rule simultaneously over all elements in $\mathcal{X}$, i.e. the decision rule vector $\boldsymbol{\delta}$. It is worth noting that, when we consider x_l as a single group, this extension can be interpreted as multiple groups with interaction between groups via $\bar{\pi}$.

We now construct the multinomial empirical success rule conditional on covariate x_l. Note that Π^L contains at most K^L elements, $|\Pi^L| = K^L < \infty$. Let $\boldsymbol{\pi}_k$ be a generic element of Π^L. Then, the population (unconditional) treatment ratio is $\bar{\pi}_k = \boldsymbol{p}'\boldsymbol{\pi}_k$. The empirical mean of $Y_t(\bar{\pi}_k)$ conditional on x_l is

$$\hat{E}_\theta[Y_t(\bar{\pi}_k)|X = x_l] := \frac{\sum_{n_k=1}^{N_k} Y_{n_k}(\bar{\pi}_k) \cdot \mathbb{1}(T_{n_k}(\bar{\pi}_k) = t, X = x_l)}{\sum_{n_k=1}^{N_k} \mathbb{1}(T_{n_k}(\bar{\pi}_k) = t, X = x_l)} \text{ for } k = 1, \ldots, K^L \text{ and } t = 0, 1.$$

Finally, the conditional multinomial empirical success rule (CMES) is defined as follows:

$$\boldsymbol{\delta}^{CMES}(\omega^n) := \sum_{k=1}^{K^L} \boldsymbol{\pi}_k \cdot \mathbb{1}[\hat{U}(\boldsymbol{\pi}_k) > \max_{\boldsymbol{\pi} \in \Pi^L_{-k}} \hat{U}(\boldsymbol{\pi})] \tag{3.12}$$

where $\Pi^L_{-k} := \Pi^L \setminus \{\boldsymbol{\pi}_k\}$ and

$$\begin{aligned}\hat{U}(\boldsymbol{\pi}_k) &:= \sum_{l=1}^{L} \Pr(X = x_l) \cdot \hat{U}_l(\boldsymbol{\pi}_k) \\ &= \sum_{l=1}^{L} \Pr(X = x_l)\Big[(1 - \pi_{kl}) \cdot \hat{E}_\theta[Y_0(\bar{\pi}_k)|X = x_l] + \pi_{kl} \cdot \hat{E}_\theta[Y_1(\bar{\pi}_k)|X = x_l]\Big],\end{aligned}$$

where π_{kl} is the l-th element of the L-dimensional vector $\boldsymbol{\pi}_k$. The CMES rule $\delta^{CMES}(\omega)$ in (3.12) looks similar to the (unconditional) MES rule in Section 3.3. However, $\boldsymbol{\pi}_k$ is now an L-dimensional vector and the rule itself is an L-dimensional vector-valued function. Let $U(\boldsymbol{\pi}_k, \theta)$ be the population counterpart of $\hat{U}(\boldsymbol{\pi})$ by replacing $\hat{E}_\theta$ with E_θ. Then, we can define the expected outcome given the CMES rule $\boldsymbol{\delta}^{CMES}$ as follows:

$$\begin{aligned}u(\boldsymbol{\delta}^{CMES}, \theta) &= \sum_{k=1}^{K^L} \Pr(\boldsymbol{\delta}^{CMES}(\omega^n) = \boldsymbol{\pi}_k; \theta) \cdot U(\boldsymbol{\pi}_k, \theta) \\ &= \sum_{k=1}^{K^L} \Pr\left(\bigcap_{j=1, j\neq k}^{K^L} \{\hat{U}(\boldsymbol{\pi}_k) > \hat{U}(\boldsymbol{\pi}_j)\}; \theta\right) \cdot U(\boldsymbol{\pi}_k, \theta).\end{aligned}$$

We are now ready to extend the bounds of the expected outcome in (3.7) to the CMES rule.

Theorem 3.3.2 *Fix $\theta \in \Theta$. Let $\Pi^L = \{\boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_{K^L}\}$, $\Delta_{kj} := |U(\boldsymbol{\pi}_k, \theta) - U(\boldsymbol{\pi}_j, \theta)|$ for $k, j = 1 \ldots, K^L$, and $\boldsymbol{\pi}_{M^*} := \arg\max_{\boldsymbol{\pi} \in \Pi^L} U(\boldsymbol{\pi}, \theta)$. Then, the following inequality holds:*

$$U(\boldsymbol{\pi}_{M^*}, \theta) - \sum_{k=1}^{K^L} \exp\left(-2\Delta_{M^*k}^2 \cdot \left\{\sum_{l=1}^{L} \Pr(X = x_l)^2 (A_{kl} + A_{M^*l})\right\}^{-1}\right) \cdot \Delta_{M^*k} \leq u(\boldsymbol{\delta}^{CMES}, \theta) \leq U(\boldsymbol{\pi}_{M^*}, \theta),$$

where $A_{kl} := (1 - \pi_{kl})^2 N_{k0l}^{-1} + \pi_{kl}^2 N_{k1l}^{-1}$ and $A_{M^*l} := (1 - \boldsymbol{\pi}_{M^*l})^2 N_{M^*0l}^{-1} + \pi_{l1}^2 N_{M^*1l}^{-1}$, with N_{ktl} representing the number of individuals with $\boldsymbol{\pi}_k$, $T = t$, and $X = x_l$.

Using the similar arguments in Section 3.3, we define the non-negative finite sample penalty:

$$D(\boldsymbol{\delta}^{CMES}, \theta) := \sum_{k=1}^{K^L} \exp\left(-2\Delta_{M^*k}^2 \cdot \left\{\sum_{l=1}^{L} \Pr(X = x_l)^2 (A_{kl} + A_{M^*l})\right\}^{-1}\right) \cdot \Delta_{M^*k},$$

and derive the following inequality:

$$D(\boldsymbol{\delta}^{CMES}, \theta) \leq \frac{1}{2} \cdot e^{-\frac{1}{2}} \sum_{k=1, k \neq M^*}^{K^L} \left\{\sum_{l=1}^{L} \Pr(X = x_l)^2 (A_{kl} + A_{M^*l})\right\}^{\frac{1}{2}}. \tag{3.13}$$

Then, we can derive the uniform bound of the regret function, which can be recovered from the observable:

$$0 \leq \sup_{\theta \in \Theta} R(\boldsymbol{\delta}^{CMES}, \theta) \leq \frac{1}{2} \cdot e^{-\frac{1}{2}} \sum_{k=1, A_{kl} \neq \bar{A}_l}^{K} \left\{\sum_{l=1}^{L} \Pr(X = x_l)^2 (A_{kl} + \bar{A}_l)\right\}^{\frac{1}{2}}, \tag{3.14}$$

where $\bar{A}_l :- \max_{k \in \{1, \ldots, K^L\}} A_{kl}$ $\forall l \in \mathcal{L}$.

We next investigate the relationship between the sample size and the proper conditioning level of covariates. Recall that given a fixed sample size using all available covariates may reduce the statistical precision in practice. Let $\mathcal{Z} := \{z_1, \ldots z_{L'}\}$ be a partitioning of the covariate space that is coarser than $\mathcal{X}$. Thus $L' < L$ and there exists a mapping $z(\cdot) : \mathcal{X} \mapsto \mathcal{Z}$. Slightly abusing notation, we use the same $\boldsymbol{\pi}$ and $\boldsymbol{p}$ for assignment ratios and proportions whose dimension is L'. Finally, if $\boldsymbol{\pi}_{k'}$ is a generic element of $\Pi^{L'}$ and $\boldsymbol{\delta}_Z^{CMES}$ is the MES rule conditional on Z, then the population expected

outcome becomes:

$$u(\boldsymbol{\delta}_Z^{CMES}, \theta) = \sum_{k'=1}^{K^{L'}} \Pr\left(\bigcap_{j=1, j\neq k'}^{K^{L'}} \{\hat{U}(\boldsymbol{\pi}_{k'}) > \hat{U}(\boldsymbol{\pi}_j)\} \right) \cdot U(\boldsymbol{\pi}_{k'}, \theta),$$

where $U(\boldsymbol{\pi}_{k'}, \theta) := \sum_{l'=1}^{L'} \Pr(Z = z_{l'}) \cdot U_{l'}(\boldsymbol{\pi}_{k'}, \theta) \equiv \sum_{l'=1}^{L'} \Pr(Z = z_{l'}) \cdot [(1 - \pi_{k'l'}) \cdot E_{P_\theta}[Y_0(\bar{\pi}_{k'})|Z = z_{l'}] + \pi_{k'l'} \cdot E_{P_\theta}[Y_1(\bar{\pi}_{k'})|Z = z_{l'}]]$ and $\bar{\pi}_{k'} := \boldsymbol{p}'\boldsymbol{\pi}_{k'}$, $k' \in \{1, \ldots, K^{L'}\}$ and $l' \in \{1, \ldots, L'\}$.

Similar to the results in Theorem 3.3.2, we can bound the expected outcome in the following corollary.

Corollary 3.3.1 *Fix $\theta \in \Theta$. Let $\Pi^{L'} = \{\boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_{K^{L'}}\}$, $\Delta_{k'j} := |U(\boldsymbol{\pi}_{k'}, \theta) - U(\boldsymbol{\pi}_j, \theta)|$ for $k', j \in 1, \ldots, K^{L'}$ and $\boldsymbol{\pi}_{M^{**}} := \arg\max_{\boldsymbol{\pi} \in \Pi^{L'}} U(\boldsymbol{\pi}, \theta)$. Then, the following inequality holds:*

$$\sum_{l'=1}^{L'} \Pr(Z = z_{l'}) \cdot U_{l'}(\boldsymbol{\pi}_{M^{**}}, \theta) - \sum_{k'=1}^{K^{L'}} \exp\left(-2\Delta_{M^{**}k'}^2 \cdot \left\{ \sum_{l'=1}^{L'} \Pr(Z = z_{l'})^2 (A_{k'l'} + A_{M^{**}l'}) \right\}^{-1} \right) \cdot \Delta_{M^{**}k'}$$
$$\leq u(\boldsymbol{\delta}_Z^{CMES}, \theta) \leq \sum_{l'=1}^{L'} \Pr(Z = z_{l'}) \cdot U_{l'}(\boldsymbol{\pi}_{M^{**}}, \theta) \tag{3.15}$$

where $A_{k'l'} := (1 - \pi_{k'l'})^2 N_{k'0l'}^{-1} + \pi_{k'l'}^2 N_{k'1l'}^{-1}$ and $A_{M^{**}l'} := (1 - \boldsymbol{\pi}_{l1})^2 N_{M^{**}0l'}^{-1} + \pi_{l1}^2 N_{M^{**}1l'}^{-1}$, with $N_{k'tl'}$ representing the number of individuals with $\boldsymbol{\pi}_{k'}$, $Z = z_{l'}$, and $T = t$.

We now suppose that the decision maker need to choose the conditioning level between X and Z. The idealized bounds for the regret function is as follows.

$$\sum_{l=1}^{L} \Pr(X = x_l) \cdot U_l(\boldsymbol{\pi}_{M^*}, \theta) - \sum_{l'=1}^{L'} \Pr(Z = z_{l'}) \cdot U_{l'}(\boldsymbol{\pi}_{M^{**}}, \theta) \leq R(\boldsymbol{\delta}_Z^{CMES}, \theta)$$
$$\leq \sum_{l'=1}^{L} \Pr(X = x_l) \cdot U_l(\boldsymbol{\pi}_{M^*}, \theta) - \sum_{l'=1}^{L'} \Pr(Z = z_{l'}) \cdot U_{l'}(\boldsymbol{\pi}_{M^{**}}, \theta) + D(\theta). \tag{3.16}$$

where

$$D(\theta) := \sum_{k'=1}^{K'} \exp\left(-2\Delta_{M^{**}k'}^2 \cdot \left\{ \sum_{l'=1}^{L'} \Pr(Z = z_{l'})^2 (A_{k'l'} + A_{M^{**}l'}) \right\}^{-1} \right) \cdot \Delta_{M^{**}k'}.$$

Finally, we achieve a uniform bound on the maximum regret function as follow.

$$L \leq \sup_{\theta \in \Theta} R(\boldsymbol{\delta}_Z^{CMES}, \theta) \leq H, \tag{3.17}$$

where

$$L :- \sup_{\theta\in\Theta}\left\{\sum_{l=1}^{L}\Pr(X=x_l)\cdot U_l(\boldsymbol{\pi}_{M^*},\theta)-\sum_{l'=1}^{L'}\Pr(Z=z_{l'})\cdot U_{l'}(\boldsymbol{\pi}_{M^{**}},\theta)\right\},$$

and

$$H :- \sup_{\theta\in\Theta}\left\{\sum_{l=1}^{L}\Pr(X=x_l)\cdot U_l(\boldsymbol{\pi}_{M^*},\theta)-\sum_{l'=1}^{L'}\Pr(Z=z_{l'})\cdot U_{l'}(\boldsymbol{\pi}_{M^{**}},\theta)+D(\theta)\right\}.$$

Using these bounds, we can compute the minimum sample size to test the proper level of conditioning variables. Let $\boldsymbol{N}_{KTL} := (N_{ktl} : k = 1,\ldots,K, t = 0, 1, \text{ and } l = 1,\ldots,L)$ be a 3-dimensional array of stratum sample sizes. Recall that the upper bound of the maximum regret conditional on X decreases as each N_{ktl} increases. Therefore, we can find a sufficient sample size that justifies conditioning on X rather than conditioning on Z:

$$\begin{aligned}&\min \boldsymbol{N}_{KTL}\\ &\text{subject to } L > \frac{1}{2}\cdot e^{-\frac{1}{2}}\sum_{k=1,A_{kl}\neq\bar{A}_l}^{K}\left\{\sum_{l=1}^{L}\Pr(X=x_l)^2(A_{kl}+\bar{A}_l)\right\}^{\frac{1}{2}},\end{aligned}$$

where we minimize each component of vector $\mathbf{N}_{KTL}$. Similar to the results in Manski (2004), it requires additional bound conditions on $U_l(\pi_{M^*},\theta)$ and $U_{l'}(\pi_{M^{**}},\theta)$ to solve for $\boldsymbol{N}_{KTL}$. Note also that the solution may not be unique since $\boldsymbol{N}_{KTL}$ is a tensor.

3.3.3 Numerical Experiments

In this subsection, we conduct some numerical experiments, where we determine a sufficient sample size to use covariate-dependent treatment rules. Suppose that we have a binary covariate $X = \{low, high\}$ available in a sample. We now construct a treatment rule with or without the covariate. The sufficient sample size guarantees that the maximum regret from a covariate-dependent rule is smaller than that from a rule without considering any covariate. Thus, we can focus on covariate-dependent rules if the sample size is bigger than the sufficient one.

In this experiment, a sample is partitioned into 2 groups ($X = low$, $X = high$), and $L = |\mathcal{X}| = 2$. Therefore, covariate-dependent rules $\boldsymbol{\pi}$ also becomes a 2-dimensional vector $\boldsymbol{\pi} = (\pi_{low}, \pi_{high})$. Suppose that we consider two possible treatment rules, $\{\boldsymbol{\pi}_1 = (0.5, 0.5), \boldsymbol{\pi}_2 = (0.7, 0.3)\}$. Unconditional treatment ratios for them becomes:

$$\begin{aligned}\bar{\pi}_1 &= \Pr(X = low)\cdot 0.50 + \Pr(X = high)\cdot 0.50,\\ \bar{\pi}_2 &= \Pr(X = low)\cdot 0.70 + \Pr(X = high)\cdot 0.30.\end{aligned}$$

We set that $\Pr(X = low)$ varies in $\{0.1, 0.5, 0.9, 0.99\}$ and that $\Pr(X = high) := 1 - \Pr(X = low)$ varies in $\{0.9, 0.5, 0.1, 0.01\}$. Recall that $(N_{ktx} : k = 1, 2, t = 0, 1, \text{ and } x = low, high)$ denotes the sample size of each partition separated by treatment rule k, treatment status t, and covariate x. In addition, N denote the total sample size. N_1 and N_2 denote the sample sizes of each cluster, where we apply $\boldsymbol{\pi_1}$ and $\boldsymbol{\pi_2}$, respectively. Assuming that all states of the nature are feasible, we compute the lower bound of maximum regret for the MES rule that does not depend on covariate X. We also compute the upper bounds of maximum regret for the covariate-dependent MES rule as the sample size increases. We then check when this upper bound with covariates becomes smaller than the lower bound without covariates.

In Tables 3.1–3.4, we summarize the experiment results. We denote the upper bound with X in bold when it becomes smaller than the lower bound without X. The sufficient sample size is as low as $N = 21$ when $\Pr(X = low) = 0.1$, $N = 18$ when $\Pr(X = low) = 0.5$, $N = 68$ when $\Pr(X = low) = 0.9$, and $N = 5875$ when $\Pr(X = low) = 0.99$. In each table, We also provide a breakdown of the sample sizes in each partition. This numerical study shows that covariate-dependent treatment rules can be justified with relatively small sample sizes unless the sizes of heterogeneous groups are quite uneven, e.g. $\Pr(X = low) = 0.99$.

Table 3.1: Sufficient Sample Sizes: $\Pr(X = low) = 0.10$

N	N_1	N_2	N_{10low}	N_{11low}	N_{20low}	N_{21low}	N_{10high}	N_{11high}	N_{20high}	N_{21high}	Upper bound with X	Lower bound without X
21	10	11	1	1	1	1	4	4	6	3	**0.144**	0.450
37	18	19	1	1	1	2	8	8	11	5	**0.100**	0.450
52	26	26	2	2	1	2	11	11	16	7	**0.085**	0.450
68	34	34	2	2	1	3	15	15	21	9	**0.074**	0.450
82	40	42	2	2	2	3	18	18	26	11	**0.067**	0.450
100	50	50	3	3	2	4	22	22	31	13	**0.061**	0.450
116	58	58	3	3	2	4	26	26	36	16	**0.056**	0.450
132	66	66	4	4	2	5	29	29	41	18	**0.053**	0.450
149	74	75	4	4	3	6	33	33	46	20	**0.050**	0.450
162	80	82	4	4	3	6	36	36	51	22	**0.048**	0.450

Table 3.2: Sufficient Sample Sizes: Pr($X = low$) = 0.50

N	N_1	N_2	N_{10low}	N_{11low}	N_{20low}	N_{21low}	N_{10high}	N_{11high}	N_{20high}	N_{21high}	Upper bound with X	Lower bound without X
18	8	10	2	2	2	3	2	2	3	2	**0.145**	0.250
34	16	18	4	4	3	6	4	4	6	3	**0.104**	0.250
50	24	26	6	6	4	9	6	6	9	4	**0.086**	0.250
66	32	34	8	8	5	12	8	8	12	5	**0.075**	0.250
81	40	41	10	10	7	14	10	10	14	6	**0.067**	0.250
98	48	50	12	12	8	17	12	12	17	8	**0.061**	0.250
114	56	58	14	14	9	20	14	14	20	9	**0.057**	0.250
130	64	66	16	16	10	23	16	16	23	10	**0.053**	0.250
146	72	74	18	18	11	26	18	18	26	11	**0.050**	0.250
161	80	81	20	20	13	28	20	20	28	12	**0.048**	0.250

Table 3.3: Sufficient Sample Sizes: Pr($X = low$) = 0.90

N	N_1	N_2	N_{10low}	N_{11low}	N_{20low}	N_{21low}	N_{10high}	N_{11high}	N_{20high}	N_{21high}	Upper bound with X	Lower bound without X
21	10	11	4	4	3	6	1	1	1	1	0.136	0.072
37	18	19	8	8	5	11	1	1	2	1	0.100	0.072
52	26	26	11	11	7	16	2	2	2	1	0.085	0.072
68	34	34	15	15	9	21	2	2	3	1	0.074	0.072
82	40	42	18	18	11	26	2	2	3	2	**0.067**	0.072
100	50	50	22	22	13	31	3	3	4	2	**0.061**	0.072
116	58	58	26	26	16	36	3	3	4	2	**0.056**	0.072
132	66	66	29	29	18	41	4	4	5	2	**0.053**	0.072
149	74	75	33	33	20	46	4	4	6	3	**0.050**	0.072
162	80	82	36	36	22	51	4	4	6	3	**0.048**	0.072

Table 3.4: Sufficient Sample Sizes: $\Pr(X = low) = 0.99$

N	N_1	N_2	N_{10low}	N_{11low}	N_{20low}	N_{21low}	N_{10high}	N_{11high}	N_{20high}	N_{21high}	Upper bound with X	Lower bound without X
21	10	11	4	4	3	6	1	1	1	1	0.14609	0.00792
37	18	19	8	8	5	12	1	1	1	1	0.10463	0.00792
53	26	27	12	12	8	17	1	1	1	1	0.08590	0.00792
69	34	35	16	16	10	23	1	1	1	1	0.07455	0.00792
84	42	42	20	20	12	28	1	1	1	1	0.06721	0.00792
5764	2882	2882	1426	1426	856	1996	15	15	21	9	0.00799	0.00792
5780	2890	2890	1430	1430	858	2002	15	15	21	9	0.00798	0.00792
5796	2898	2898	1434	1434	861	2007	15	15	21	9	0.00797	0.00792
5812	2906	2906	1438	1438	863	2013	15	15	21	9	0.00796	0.00792
5828	2914	2914	1442	1442	865	2019	15	15	21	9	0.00795	0.00792
5844	2922	2922	1446	1446	868	2024	15	15	21	9	0.00793	0.00792
5860	2930	2930	1450	1450	870	2030	15	15	21	9	0.00792	0.00792
5875	2938	2937	1454	1454	872	2035	15	15	21	9	**0.00791**	0.00792
5892	2946	2946	1458	1458	875	2041	15	15	21	9	**0.00790**	0.00792
5907	2954	2953	1462	1462	877	2046	15	15	21	9	**0.00789**	0.00792
5924	2962	2962	1466	1466	880	2052	15	15	21	9	**0.00788**	0.00792

3.4 Conclusion

In this chapter we study statistical treatment rules under social interaction. We impose the anonymous interaction assumption, and consider a treatment decision problem, where we choose the treatment ratio for each cluster. We propose a simple but intuitive rule called the multinomial empirical success (MES) rule. We construct the finite sample regret bound of the MES rule and show how it can be applied in the treatment decision problems.

We may consider a few possible extensions. It is interesting to investigate the finite sample optimality of the MES rule. The solution does not follow immediately if we apply the finite action problem framework, which we adopt in the asymptotic optimality analysis, and the game theoretic approach in Stoye (2009) in the finite sample case. It is also interesting to relax the anonymous interaction assumption. Then, we have to ask what kind of additional information can help reduce the dimension of the action space. The network information can be such an example. We leave these questions for our future research.

Bibliography

Athey, S. and S. Wager (2021). Policy learning with observational data. *Econometrica 89*(1), 133–161.

Baird, S., J. A. Bohren, C. McIntosh, and B. Özler (2018). Optimal design of experiments in the presence of interference. *Review of Economics and Statistics 100*(5), 844–860.

Beaman, L. A. (2012). Social networks and the dynamics of labour market outcomes: Evidence from refugees resettled in the us. *The Review of Economic Studies 79*(1), 128–161.

Bursztyn, L., F. Ederer, B. Ferman, and N. Yuchtman (2014). Understanding mechanisms underlying peer effects: Evidence from a field experiment on financial decisions. *Econometrica 82*(4), 1273–1301.

Crépon, B., E. Duflo, M. Gurgand, R. Rathelot, and P. Zamora (2013). Do labor market policies have displacement effects? evidence from a clustered randomized experiment. *The quarterly journal of economics 128*(2), 531–580.

Duflo, E. (2004). Scaling up and evaluation. In *Annual World Bank Conference on Development Economics*, pp. 341–369.

Duflo, E. and E. Saez (2003). The role of information and social interactions in retirement plan decisions: Evidence from a randomized experiment. *The Quarterly journal of economics 118*(3), 815–842.

Gautier, P., P. Muller, B. van der Klaauw, M. Rosholm, and M. Svarer (2018). Estimating equilibrium effects of job search assistance. *Journal of Labor Economics 36*(4), 1073–1125.

Heckman, J. J., L. Lochner, and C. Taber (1999). Human capital formation and general equilibrium treatment effects: a study of tax and tuition policy. *Fiscal Studies 20*(1), 25–40.

Hirano, K. and J. R. Porter (2009). Asymptotics for statistical treatment rules. *Econometrica 77*(5), 1683–1701.

Hirano, K. and J. R. Porter (2020). Asymptotic analysis of statistical decision rules in econometrics. In *Handbook of Econometrics*, Volume 7, pp. 283–354. Elsevier.

Kitagawa, T., S. Lee, and C. Qiu (2022). Treatment choice with nonlinear regret. *arXiv preprint arXiv:2205.08586*.

Kitagawa, T. and A. Tetenov (2018). Who should be treated? empirical welfare maximization methods for treatment choice. *Econometrica 86*(2), 591–616.

Manski, C. F. (2004). Statistical treatment rules for heterogeneous populations. *Econometrica 72*(4), 1221–1246.

Manski, C. F. (2009). *Identification for prediction and decision*. Harvard University Press.

Manski, C. F. (2013). Identification of treatment response with social interactions. *The Econometrics Journal 16*(1), S1–S23.

Manski, C. F. (2021). Econometrics for decision making: Building foundations sketched by haavelmo and wald. *Econometrica 89*(6), 2827–2853.

Manski, C. F. and A. Tetenov (2007). Admissible treatment rules for a risk-averse planner with experimental data on an innovation. *Journal of Statistical Planning and Inference 137*(6), 1998–2010.

Manski, C. F. and A. Tetenov (2016). Sufficient trial size to inform clinical practice. *Proceedings of the National Academy of Sciences 113*(38), 10518–10523.

Manski, C. F. and A. Tetenov (2019). Trial size for near-optimal choice between surveillance and aggressive treatment: Reconsidering mslt-ii. *The American Statistician 73*(sup1), 305–311.

Mbakop, E. and M. Tabord-Meehan (2021). Model selection for treatment choice: Penalized welfare maximization. *Econometrica 89*(2), 825–848.

Root, E. D., S. Giebultowicz, M. Ali, M. Yunus, and M. Emch (2011). The role of vaccine coverage within social networks in cholera vaccine efficacy. *PLoS One 6*(7), e22971.

Stoye, J. (2009). Minimax regret treatment choice with finite samples. *Journal of Econometrics 151*(1), 70–81.

Viviano, D. (2019). Policy targeting under network interference. *arXiv preprint arXiv:1906.10258*.

3.5 Appendix

3.5.1 Proof of Theorem 3.3.1

We first show the bounds in the main text. Without loss of generality, suppose $\max_{\pi\in\mathbf{\Pi}} U(\pi,\theta) = U(\pi_1,\theta)$. By definition, the upper bound of $u(\delta^{MES},\theta)$ is $U(\pi_1,\theta)$. We now restate the expected welfare under the MES rule as

$$\begin{aligned} u(\delta^{MES},\theta) &= \Pr\left(\hat{U}(\pi_1,\theta) > \max_{\pi\in\mathbf{\Pi}_{-1}} \hat{U}(\pi,\theta)\right)\cdot U(\pi_1,\theta) + \sum_{k=2}^{K}\Pr\left(\hat{U}(\pi_k,\theta) > \max_{\pi\in\mathbf{\Pi}_{-k}} \hat{U}(\pi,\theta)\right)\cdot U(\pi_k,\theta) \\ &= \left[1 - \sum_{k=2}^{K}\Pr\left(\hat{U}(\pi_k,\theta) > \max_{\pi\in\mathbf{\Pi}_{-k}} \hat{U}(\pi,\theta)\right)\right]\cdot U(\pi_1,\theta) \\ &\quad + \sum_{k=2}^{K}\Pr\left(\hat{U}(\pi_k,\theta) > \max_{\pi\in\mathbf{\Pi}_{-k}} \hat{U}(\pi,\theta)\right)\cdot U(\pi_k,\theta) \end{aligned} \tag{3.18}$$

Define

$$\begin{aligned} w(\delta^{MES},\theta) := u(\delta^{MES},\theta) - \Bigg\{&\left[1 - \sum_{k=2}^{K}\Pr(\hat{U}(\pi_k,\theta) - \hat{U}(\pi_1,\theta) > 0)\right]\cdot U(\pi_1,\theta) \\ &+ \sum_{k=2}^{K}\left[\Pr(\hat{U}(\pi_k,\theta) - \hat{U}(\pi_1,\theta) > 0)\right]\cdot U(\pi_k,\theta)\Bigg\}. \end{aligned}$$

Plugging in (3.18) into w, we have

$$\begin{aligned} w(\delta^{MEG},\theta) &= \left\{\left[1 - \sum_{k=2}^{K}\Pr\left(\hat{U}(\pi_k,\theta) > \max_{\pi\in\mathbf{\Pi}_{-k}} \hat{U}(\pi,\theta)\right)\right] - \left[1 - \sum_{k=2}^{K}\Pr\left(\hat{U}(\pi_k,\theta) - \hat{U}(\pi_1,\theta) > 0\right)\right]\right\}\cdot U(\pi_1,\theta) \\ &\quad + \sum_{k=2}^{K}\left[\Pr\left(\hat{U}(\pi_k,\theta) > \max_{\pi\in\mathbf{\Pi}_{-k}} \hat{U}(\pi,\theta)\right) - \Pr\left(\hat{U}(\pi_k,\theta) - \hat{U}(\pi_1,\theta) > 0\right)\right]\cdot U(\pi_k,\theta) \\ &= \sum_{k=2}^{K}\left[\Pr\left(\hat{U}(\pi_k,\theta) - \hat{U}(\pi_1,\theta) > 0\right) - \Pr\left(\hat{U}(\pi_k,\theta) > \max_{\pi\in\mathbf{\Pi}_{-k}} \hat{U}(\pi,\theta)\right)\right]\cdot U(\pi_1,\theta) \\ &\quad + \sum_{k=2}^{K}\left[\Pr\left(\hat{U}(\pi_k,\theta) > \max_{\pi\in\mathbf{\Pi}_{-k}} \hat{U}(\pi,\theta)\right) - \Pr\left(\hat{U}(\pi_k,\theta) - \hat{U}(\pi_1,\theta) > 0\right)\right]\cdot U(\pi_k,\theta) \\ &= \sum_{k=2}^{K}\left[\Pr\left(\hat{U}(\pi_k,\theta) - \hat{U}(\pi_1,\theta) > 0\right) - \Pr\left(\hat{U}(\pi_k,\theta) > \max_{\pi\in\mathbf{\Pi}_{-k}} \hat{U}(\pi,\theta)\right)\right]\cdot U(\pi_1,\theta) \end{aligned}$$

$$
\begin{aligned}
&- \sum_{k=2}^{K} \left[\Pr\left(\hat{U}(\pi_k,\theta) - \hat{U}(\pi_1,\theta) > 0\right) - \Pr\left(\hat{U}(\pi_k,\theta) > \max_{\pi \in \mathbf{\Pi}_{-k}} \hat{U}(\pi,\theta)\right) \right] \cdot U(\pi_k,\theta) \\
&= \sum_{k=2}^{K} \left[\Pr\left(\hat{U}(\pi_k,\theta) - \hat{U}(\pi_1,\theta) > 0\right) - \Pr\left(\hat{U}(\pi_k,\theta) > \max_{\pi \in \mathbf{\Pi}_{-k}} \hat{U}(\pi,\theta)\right) \right] \cdot (U(\pi_1,\theta) - U(\pi_k,\theta)) \\
&\geq 0.
\end{aligned}
$$

Note that the last inequality holds since

$$
\Pr\left(\hat{U}(\pi_k,\theta) - \hat{U}(\pi_1,\theta) \geq 0\right) \geq \Pr\left(\hat{U}(\pi_k,\theta) > \max_{\pi \in \mathbf{\Pi}_{-k}} \hat{U}(\pi,\theta)\right), \ \forall k.
$$

Therefore, we have

$$
\begin{aligned}
u(\delta^{MEG},\theta) \geq &\left[1 - \sum_{k=2}^{K} Pr(\hat{U}(\pi_k,\theta) - \hat{U}(\pi_1,\theta) \geq 0)\right] \cdot U(\pi_1,\theta) \\
&+ \sum_{k=2}^{K} \left[Pr(\hat{U}(\pi_k,\theta) - \hat{U}(\pi_1,\theta) \geq 0) \right] \cdot U(\pi_k,\theta) \qquad (3.19)
\end{aligned}
$$

To proceed, we use the Hoeffding inequality to derive bounds for the probabilities in (3.19). For $k = 2, \ldots, K$,

$$
\begin{aligned}
\hat{U}(\pi_k,\theta) - \hat{U}(\pi_1,\theta) = \frac{1}{N_k + N_1} \Bigg\{ &\sum_{n \in N(0,\pi_k)} (1-\pi_k) y_n \frac{N_k + N_1}{N_{k0}} + \sum_{n \in N(1,\pi_k)} \pi_k y_n \frac{N_k + N_1}{N_{k1}} + \\
&\sum_{n \in N(0,\pi_1)} -(1-\pi_1) y_n \frac{N_k + N_1}{N_{10}} + \sum_{n \in N(1,\pi_1)} -\pi_1 y_n \frac{N_k + N_1}{N_{11}} \Bigg\}
\end{aligned}
$$

Thus, $\hat{U}(\pi_k,\theta) - \hat{U}(\pi_1,\theta)$ is the average of $(N_1 + N_k)$ independent random variables whose ranges are $[0, (1-\pi_k)(N_1+N_k)/N_{k0}]$, $[0, \pi_k(N_1+N_k)/N_{k1}]$, $[-(1-\pi_1)(N_1+N_k)/N_{10}, 0]$, and $[-\pi_1(N_1+N_k)/N_{11}, 0]$. Since $\max_{\pi \in \mathbf{\Pi}} U(\pi,\theta) = U(\pi_1,\theta)$, $\mathbb{E}[\hat{U}(\pi_k,\theta) - \hat{U}(\pi_1,\theta)] = -|U(\pi_k,\theta) - U(\pi_1,\theta)| = -\Delta_{1k}$. Applying the Hoeffding inequality for all $k \neq 1$, we have

$$
\begin{aligned}
&\Pr(\hat{U}(\pi_k,\theta) - \hat{U}(\pi_1,\theta) \geq 0) \\
&\qquad = \Pr(\hat{U}(\pi_k,\theta) - \hat{U}(\pi_1,\theta) + \Delta_{k1} \geq \Delta_{k1}) \\
&\qquad \leq \exp\left(-2\Delta_{k1}^2 \cdot \left\{ (1-\pi_k)^2 N_{k0}^{-1} + \pi_k^2 N_{k1}^{-1} + (1-\pi_1)^2 N_{10}^{-1} + \pi_1^2 N_{11}^{-1} \right\}^{-1} \right) \\
&\qquad \equiv \exp\left(-2\Delta_{k1}^2 \cdot (A_k + A_1)^{-1} \right) \qquad (3.20)
\end{aligned}
$$

Substituting (3.20) into the last inequality of (3.19), we obtain;

$$\begin{aligned} u(\delta^{MEG}, \theta) &\geq \left(1 - \sum_{k=2}^{K} \exp\left(-2\Delta_{k1}^2 \cdot (A_k + A_1)^{-1}\right)\right) \cdot U(\pi_1, \theta) \\ &\quad + \sum_{k=2}^{K} \exp\left(-2\Delta_{k1}^2 \cdot (A_k + A_1)^{-1}\right) \cdot U(\pi_k, \theta) \\ &= U(\pi_1, \theta) - \sum_{k=1}^{K} \exp\left(-2\Delta_{k1}^2 \cdot (A_k + A_1)^{-1}\right) \cdot \Delta_{1k} \end{aligned} \tag{3.21}$$

as required when $M^* = 1$.

□

3.5.2 Proof of Theorem 3.3.2

Without loss of generality, let $\max_{\boldsymbol{\pi} \in \Pi^L} \sum_{l=1}^{L} P(X = x_l) U_l(\boldsymbol{\pi}, \theta) = U(\boldsymbol{\pi}_1, \theta)$. The upper bound is straightforward; the highest attainable outcome for the CMES rule which conditions on all covariates is

$$\max_{\boldsymbol{\pi} \in \Pi^L} \sum_{l=1}^{L} P(X = x_l) U_l(\boldsymbol{\pi}, \theta) = U(\boldsymbol{\pi}_1, \theta)$$

Now, restate the expected outcome under the CMES rule as;

$$\begin{aligned} u(\boldsymbol{\delta}^{CMES}, \theta) &= \Pr\left(\hat{U}(\boldsymbol{\pi}_1, \theta) > \max_{\boldsymbol{\pi} \in \Pi_{-1}^L} \hat{U}(\boldsymbol{\pi}, \theta)\right) \cdot U(\boldsymbol{\pi}_1, \theta) + \sum_{k=2}^{K} \Pr\left(\hat{U}(\boldsymbol{\pi}_k, \theta) > \max_{\boldsymbol{\pi} \in \Pi_{-k}^L} \hat{U}(\boldsymbol{\pi}, \theta)\right) \cdot U(\boldsymbol{\pi}_k, \theta) \\ &= \left[1 - \sum_{k=2}^{K} \Pr\left(\hat{U}(\boldsymbol{\pi}_k, \theta) > \max_{\boldsymbol{\pi} \in \Pi_{-k}^L} \hat{U}(\boldsymbol{\pi}, \theta)\right)\right] \cdot U(\boldsymbol{\pi}_1, \theta) \\ &\quad + \sum_{k=2}^{K} \Pr\left(\hat{U}(\boldsymbol{\pi}_k, \theta) > \max_{\boldsymbol{\pi} \in \Pi_{-k}^L} \hat{U}(\boldsymbol{\pi}, \theta)\right) \cdot U(\boldsymbol{\pi}_k, \theta) \end{aligned} \tag{3.22}$$

Using the same arguments as in the proof of theorem 3.3.1, we can show that

$$\begin{aligned} u(\boldsymbol{\delta}^{CMES}, \theta) &\geq \left[1 - \sum_{k=2}^{K} \Pr(\hat{U}(\boldsymbol{\pi}_k, \theta) - \hat{U}(\boldsymbol{\pi}_1, \theta) > 0)\right] \cdot U(\boldsymbol{\pi}_1, \theta) \\ &\quad + \sum_{k=2}^{K} \left[\Pr(\hat{U}(\boldsymbol{\pi}_k, \theta) - \hat{U}(\boldsymbol{\pi}_1, \theta) > 0)\right] \cdot U(\boldsymbol{\pi}_k, \theta) \end{aligned} \tag{3.23}$$

Now, for $k = 2, \ldots, K$,

$$\hat{U}(\boldsymbol{\pi}_k, \theta) - \hat{U}(\boldsymbol{\pi}_1, \theta) = \sum_{l=1}^{L} \frac{P(X = x_l)}{N_k + N_1} \Bigg\{ \sum_{n \in N(0, \boldsymbol{\pi}_k, x_l)} (1 - \pi_{lk}) y_n \frac{N_k + N_1}{N_{k0l}} + \sum_{n \in N(1, \boldsymbol{\pi}_k, x_l)} \boldsymbol{\pi}_{lk} y_n \frac{N_k + N_1}{N_{k1l}} + \sum_{n \in N(0, \boldsymbol{\pi}_1, x_l)} (-(1 - \boldsymbol{\pi}_{l1})) y_n \frac{N_k + N_1}{N_{10l}} + \sum_{n \in N(1, \boldsymbol{\pi}_1, x_l)} (-\boldsymbol{\pi}_1) y_n \frac{N_k + N_1}{N_{11l}} \Bigg\}.$$

Thus, $\hat{U}(\boldsymbol{\pi}_k, \theta) - \hat{U}(\boldsymbol{\pi}_1, \theta)$ is the average of $N_k + N_1$ independent random variables whose ranges are $[0, P(X = x_l)(1 - \pi_{lk})(N_k + N_1)/N_{k0l}]$, $[0, P(X = x_l)\pi_{lk}(N_k + N_1)/N_{k1l}]$, $[-P(X = x_l)(1 - \pi_{l1})(N_k + N_1)/N_{10l}, 0]$, and $[-P(X = x_l)\pi_{l1}(N_k + N_1)/N_{11l}, 0]$.

For all $k \neq 1$, the Hoeffding inequality yields

$$\begin{aligned} Pr(\hat{U}(\boldsymbol{\pi}_k, \theta) - \hat{U}(\boldsymbol{\pi}_1, s) \geq 0) &= Pr(\hat{U}(\boldsymbol{\pi}_k, \theta) - \hat{U}(\boldsymbol{\pi}_1, \theta) + \Delta_{k1} \geq \Delta_{k1}) \\ &\leq \exp\left(-2\Delta_{k1}^2 \cdot \left\{ \sum_{l=1}^{L} P(X = x_l)^2 [(1 - \pi_{lk})^2 N_{k0l}^{-1} + \pi_{lk}^2 N_{k1l}^{-1} + (1 - \boldsymbol{\pi}_{l1})^2 N_{10l}^{-1} + \pi_{l1}^2 N_{11l}^{-1}] \right\}^{-1} \right) \\ &\equiv \exp\left(-2\Delta_{k1}^2 \cdot \left\{ \sum_{l=1}^{L} P(X = x_l)^2 (A_{kl} + A_{1l}) \right\}^{-1} \right) \end{aligned} \tag{3.24}$$

Plugging in (3.24) into (3.23), we obtain

$$\begin{aligned} u(\boldsymbol{\delta}^{CMES}, \theta) &\geq \left[1 - \sum_{k=2}^{K} \exp\left(-2\Delta_{k1}^2 \cdot \left\{ \sum_{l=1}^{L} P(X = x_l)^2 (A_{kl} + A_{1l}) \right\}^{-1} \right) \right] \cdot U(\boldsymbol{\pi}_1, \theta) \\ &\quad + \sum_{k=2}^{K} \left[\exp\left(-2\Delta_{k1}^2 \cdot \left\{ \sum_{l=1}^{L} P(X = x_l)^2 (A_{kl} + A_{1l}) \right\}^{-1} \right) \right] \cdot U(\boldsymbol{\pi}_k, \theta) \\ &= U(\boldsymbol{\pi}_1, \theta) - \sum_{k=1}^{K} \exp\left(-2\Delta_{k1}^2 \cdot \left\{ \sum_{l=1}^{L} P(X = x_l)^2 (A_{kl} + A_{1l}) \right\}^{-1} \right) \cdot \Delta_{1k} \end{aligned} \tag{3.25}$$

as required when $M^* = 1$. □