Essays on Optimal Transport Theory and Causal Inference: A Theoretical and Empirical Approach

Wendao Xue

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2024

Reading Committee:

Yanqin Fan, Chair

Jing Tao

Alex Luedtke

Program Authorized to Offer Degree:

Economics

University of Washington

Abstract

Essays on Optimal Transport Theory and Causal Inference: A Theoretical and Empirical Approach

Wendao Xue

Chair of the Supervisory Committee:
Yanqin Fan
Department of Economics

This dissertation aims to study the causal inference from both theoretical and empirical perspectives. Specifically, the first two chapters seek to extend causal inference methods for problems that are underdeveloped in current methodologies, i.e., addressing non-overlap support and misreporting of outcome variables. The third chapter explores the use of COVID-19 as an instrumental variable from an empirical perspective to identify the causal impact of transportation on air pollution.

The first chapter explores identifying average treatment effects for the treated in the region where the covariate distributions across treatment and control groups have non-overlap support. We make a natural domain shift assumption for the non-overlap region based on the optimal transport theory. We study the identification of average treatment effects for the non-overlap region and propose three-step estimators of the average treatment effect and quantile treatment effect for the treated in the non-overlap region. We establish the consistency and asymptotic normality of the proposed estimators under high-level assumptions on the estimator of the optimal transport map. Three examples of the estimator of the optimal transport map are studied in detail and are shown to satisfy the high-level assumptions under primitive conditions. We investigate the finite sample performance of our estimator and Wald inference via simulation.

In the second chapter, we switch our focus to the misreporting of outcome variables in causal inference. In fact, self-reported outcomes are commonly used to identify average treatment effects. However, if reported outcomes are linked to misaligned incentives, individuals may strategically misreport their outcomes, thereby potentially biasing the estimation. We study the identification of the average treatment effect on the untreated (ATU) under two common scenarios – incentives linked to the value (Scenario 1) and the rank (Scenario 2) of the reported outcomes. An optimal transport map is leveraged to facilitate identification in Scenario 2. We introduce plug-in estimators for ATU and derive consistency and asymptotic normality of the estimators in both scenarios. As an extension to the plug-in estimators, we derive the Neyman orthogonal moments and introduce double machine learning (DML) estimators in both scenarios. We illustrate the performance of plug-in estimators through Monte Carlo simulations. Utilizing a self-reported criminal activity dataset with a validation subsample, we've shown the efficacy of the proposed estimators.

In the third chapter, we study the causal inference from an empirical perspective. We explore the causal effects of transportation on air pollution, using the unique context of the COVID-19 pandemic to address issues of reverse causality. By utilizing last-month COVID-19 infection rates and related online search queries as instruments for travel behavior and using a two-way fixed-effects model, we assess the impact of public and private transport on six air pollutants across 36 major Chinese cities. Our results reveal that the causal impact of transportation on air quality is significantly underestimated if endogeneity is not considered. Correcting for this, we find that a 1% increase in public transport usage and traffic congestion leads to increases of 0.039% and 0.368% in air pollution levels, respectively. The analysis highlights the heterogeneous effects of transportation modes and pollutants. These insights have significant implications for urban and environmental economics and policy evaluation.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I could not have completed this thesis without the support of many people, to whom I wish to express my sincere gratitude.

I would like to express my deepest gratitude to my advisor, Professor Yanqin Fan, for her invaluable contributions to my academic development. Professor Fan guided me through each step of this academic journey with remarkable patience and understanding, from the initial selection of a research topic to the nuances of academic writing. She is truly the best advisor I could ever imagine. She excels in providing rigorous academic training and inspires me with her work style and attitude. Moreover, Professor Fan has greatly impacted me with her caring and supportive approach to her students. She has always inspired me to become a great teacher like her.

I would like to express my most profound gratitude to Professor Xiaohong Chen for her critical and thoughtful approach to forming research ideas. Despite her renown, she has treated me as a co-author and provided extensive, valuable discussions on the technical details of sieve estimation. I am also deeply appreciative of my committee members, Professor Jing Tao, Professor Alex Luedtke, and Professor Yen-Chi Chen, for their constructive feedback and substantial support. Additionally, I am grateful to Heidi Hannah and Kim H. Lee for their indispensable administrative assistance.

Lastly, my deepest gratitude goes to my parents, Weichao Xue and Linglin Liu, for their unconditional love and support at all times. It is only upon becoming a parent myself that I have truly understood the depth of their love, which cherishes every single achievement of mine and supports my career path unconditionally. This thesis is as much theirs as it is mine. I would like to express my gratitude to my mother-in-law, Jingzhi Tu. Her unwavering

support means a lot to me. I would also like to thank my husband, Yifan Yu, who has consistently uplifted me, celebrated even my smallest accomplishments, offered constructive feedback, and contributed significantly to my life's journey.

DEDICATION

To my little boy, Dawson: "Success is not final, failure is not fatal, it is the courage to continue that counts."

Chapter 1

IDENTIFICATION AND ESTIMATION OF TREATMENT EFFECTS IN THE NON-OVERLAP REGION

1.1 Introduction

Let Y_1 and Y_0 denote the potential outcomes of a binary treatment. Define $Y := Y_1 D + Y_0 (1 - D)$ as the realized outcome, where D is the binary treatment indicator such that an individual with $D = 1$ receives the treatment and an individual with $D = 0$ does not receive the treatment. Let $X \in \mathcal{X} \subseteq \mathbb{R}^{d_X}$ denote individual's observable characteristics. Strong ignorability stated below is commonly adopted in the literature to identify various average treatment effect parameters, see e.g., Rosenbaum and Rubin (1983a,b), Hahn (1998), Heckman, Ichimura, Smith, and Todd (1998a); Heckman, Ichimura, and Todd (1998b), Dehejia and Wahba (1999), and Hirano, Imbens, and Ridder (2003), to name only a few.

Strong Ignorability. *(i) For all $x \in \mathcal{X}$, (Y_1, Y_0) is jointly independent of D conditional on $X = x$; (ii) For all $x \in \mathcal{X}$, $0 < p(x) < 1$, where $p(x) \equiv \Pr(D = 1 \mid X = x)$.*

Strong ignorability is composed of two parts: (i) is the unconfoundedness/selection-on-observables assumption, and (ii) is the overlap/common support assumption. Suppose a random sample $\{Y_i, X_i, D_i\}_{i=1}^n$ on (Y, X, D) is available. Then under Assumption SI (i), for all $x \in \mathcal{X}$, $F_1(y \mid x)$ and $F_0(y \mid x)$ are point identified: for $j = 1, 0$,

$$F_j(y \mid x) = \mathbb{P}(Y \leq y \mid X = x, D = j). \tag{1.1}$$

Moreover, since the distribution of X in each group is identified, the unconditional marginal cdfs $F_1(\cdot)$ and $F_0(\cdot)$ are also point identified. As a result, under SI (ii), any policy parameter depending on $F_1(y \mid x)$ and $F_0(y \mid x)$ or $F_1(\cdot)$ and $F_0(\cdot)$ is point identified and can be consis-

tently estimated nonparametrically. Although it is often argued that the unconfoundedness assumption may be made more plausible by conditioning on more observable covariates, the overlap assumption is difficult to satisfy with many covariates. Furthermore, even if the overlap assumption holds in population, there are typically non-overlap regions in the sample on covariates in the treated and control groups. This lack of overlap in the sample hampers nonparametric estimates of average treatment effects reflected in large bias and variance, see e.g., Crump, Hotz, Imbens, and Mitnik (2009).

Methods have been developed to address the limited overlap problem in the causal inference literature. Most works propose to estimate trimmed or weighted versions of average treatment effects. For example, Crump, Hotz, Imbens, and Mitnik (2009) identify and estimate ATE for the subpopulation corresponding to the most overlapped region, see Nethery, Mealli, and Dominici (2019) for a detailed discussion of other works and examples for which policy questions call for the estimation of treatment effect parameters for the subpopulation corresponding to the limited overlap region or of the untrimmed/unweighted average treatment effects for the whole population. They are different from the modified parameters studied in the current literature when treatment effect is heterogenous. This motivates Nethery, Mealli, and Dominici (2019) and the current chapter. Without any additional assumption, treatment effects in the limited or non-overlap region are not identified. Nethery, Mealli, and Dominici (2019) proposes to extrapolate the trend of the estimated treatment effect in the overlap region using parametric specification to estimate treatment effects in the limited overlap region. To be specific, they present a three-step approach in their study. Initially, they introduce a data-driven methodology for segregating data into overlap and non-overlap regions. Subsequently, they utilize the BART model to estimate the connection between potential outcomes, observable covariates, and treatment variables, with a normal error term assumption. They then impute the unobservable potential outcome. Finally, they extend the individual causal effect trends identified in the overlap region to the non-overlap region, while imposing the assumption that the trend can be extrapolated using spline smoothing.

In this chapter, we take a different approach and make a natural domain shift assumption for the limited overlap region. Our domain shift assumption is inspired by works on transfer learning or domain adaptation in the machine learning literature, especially Courty, Flamary, Tuia, and Rakotomamonjy (2016). Under the domain shift assumption, the counterfactual distribution of the potential outcome for the treated in the limited overlap region is identified by the distribution of the potential outcome for the untreated with the same "rank" as that of the treated, where the "rank" here refers to the optimal transport map pushing the distribution of the covariate for the treated to that for the control. Exploiting an estimator of the optimal transport map, we propose nonparametric estimators of treatment effects including the average treatment effect for the treated and quantile treated effect for the treated for the non-overlap region and establish asymptotic theory for our estimators under high level assumptions on the estimator of the optimal transport map. We then verify them for three specific estimators of the optimal transport map including affine map and a sieve estimator. We examine the finite sample performance of our estimator and inference via simulation.

Extension to the whole population is possible by first identifying the overlap and non-overlap regions based on existing approaches in Crump, Hotz, Imbens, and Mitnik (2009) or Nethery, Mealli, and Dominici (2019). In this chapter, we take the non-overlap region as given.

The rest of the chapter is organized as follows. Section 2 introduces the setup, the domain shift assumption, and identification results. Section 3 presents a three-step estimator of the counterfactual parameter and the three estimators of the optimal transport map. In Section 4, we establish asymptotic theory under high-level assumptions which are verified in Section 5. Section 6 develops the asymptotic theory for the treatment effects for the treated. Section 7 presents simulation results, and the last section concludes. Technical proofs are relegated to a series of appendices.

1.2 Treatment Effects in the Non-overlap Region

For notational simplicity, we let $X_1 \overset{d}{=} X|D=1$ and $X_0 \overset{d}{=} X|D=0$. Accordingly we sometimes use $\{(Y_{0i}, X_{0i})\}_{i=1}^{n_0}$ for the subsample with $D_i = 0$ (here the control group), and $\{(Y_{1i}, X_{1i})\}_{i=1}^{n_1}$ for the subsample with $D_i = 1$ (here the treated group), where $n_0 + n_1 = n$.

Denote the support of X_j as $\mathcal{X}_j \subseteq \mathbb{R}^{d_X}$ for $j = 0, 1$. Under Assumption SI (ii), $\mathcal{X}_1 = \mathcal{X}_0 = \mathcal{X}$ which is not required in this chapter. WLOG, we consider identification and estimation of treatment effects for the subpopulation $X_1 \in \mathcal{S}$, where $\mathcal{S}$ is a subset of $\mathcal{X}_1$. Although $\mathcal{S}$ can be any subset of $\mathcal{X}_1$, our methods are most important in the non-overlap region, where there is sufficient data in $\mathcal{S}$ in the treated group, but *no data* in $\mathcal{S}$ in the control group. In this chapter, We treat $\mathcal{S}$ as known.

Two commonly used parameters are average treated effect for the treated (ATT) and quantile treatment effect for the treated (QTT). Take ATT as an example, since

$$E[Y_1 - Y_0|X \in \mathcal{S}, D=1] = E[Y_1|X \in \mathcal{S}, D=1] - E[Y_0|X \in \mathcal{S}, D=1],$$

where $E[Y_1|X \in \mathcal{S}, D=1]$ is identified from the treated sample, it suffices to identify the counterfactual mean parameter $\beta_{o\mathcal{S}} = E[Y_0|X \in \mathcal{S}, D=1]$. Similarly for QTT with $q \in (0,1)$, since $F^{-1}_{Y_1|D}(q|X \in \mathcal{S}, D=1)$ is identified from the treated sample, it is sufficient to identify the counterfactual quantile parameter $\beta_{o\mathcal{S}} = F^{-1}_{Y_0|D}(q|X \in \mathcal{S}, D=1)$. To incorporate both parameters, we consider a *one-dimensional parameter* $\beta_{o\mathcal{S}} \in \mathcal{B} \subset \mathbb{R}$ *defined via the moment condition*

$$E[m(Y_0; \beta)|X \in \mathcal{S}, D=1] = 0 \text{ if and only if } \beta = \beta_{o\mathcal{S}}, \tag{1.2}$$

where $m(\cdot;\cdot)$ is a real-valued known function that is possibly non-linear and non-smooth in β and/or in Y_0. For example, $m(Y_0;\beta) = Y_0 - \beta$ for ATT, and $m(Y_0;\beta) = I\{Y_0 \leq \beta\} - q$ for QTT with $q \in (0,1)$.

In the rest of this section, we first introduce our identification assumption and then establish identification of $\beta_{o\mathcal{S}}$ in (1.2) and the corresponding ATT and QTT.

1.2.1 Measure Transportation and the Domain Shift Assumption

Our identification assumption relies critically on measure transportation. Measure transportation refers to the problem of pushing one distribution to another by a transformation or transport map. More precisely let ν_1 and ν_0 denote two probability measures on $\mathbb{R}^d$. We say that a transport map T pushes ν_1 to ν_0 written as $T\#\nu_1 = \nu_0$ if and only if $T(X) \sim \nu_0$ for any $X \sim \nu_1$. A remarkable result known as McCann's theorem ensures the existence of such a map under weak assumptions on ν_1 and ν_0. We restate McCann existence theorem below from Chernozhukov, Galichon, Hallin, Henry, et al. (2017), where $T = \nabla\psi$ is the transport map.

Lemma 1 (McCann's Existence Result)**.** Let ν_1 and ν_0 be two distributions on $\mathbb{R}^d$.

1. If ν_1 is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^d$, with support contained in a convex set $\mathcal{V}_1$, the following holds: there exists a convex function $\psi : \mathcal{V}_1 \to \mathbb{R} \cup \{+\infty\}$ such that $\nabla\psi\#\nu_1 = \nu_0$. The function $\nabla\psi$ exists and is unique, ν_1-almost everywhere.

2. If, in addition, ν_0 is absolutely continuous on $\mathbb{R}^d$ with support contained in a convex set $\mathcal{V}_0$, the following holds: there exists a convex function $\psi^* : \mathcal{V}_0 \to \mathbb{R} \cup \{+\infty\}$ such that $\nabla\psi^*\#\nu_0 = \nu_1$. The function $\nabla\psi^*$ exists, is unique and equal to $\nabla\psi^{-1}$, ν_0-almost everywhere.

Instead of Assumption SI, we adopt the following assumption throughout the chapter.

Assumption 1 (Domain Shift)**.** (i) For $j = 0, 1$ the distribution of X_j is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^{d_X}$, with density function f_{X_j} and convex support $\mathcal{X}_j$ with non-empty interior, and $E(|X_j|^2) < \infty$; (ii) The conditional distributions of Y_0 given X_1 and Y_0 given X_0 satisfy the *domain shift assumption*:

$$F(y|x, D = 1) = F(y|T_o(x), D = 0) \text{ for all } x \in \mathcal{S} \subset \mathcal{X}_1 \text{ and } y \in \mathcal{Y},$$

where $T_o \# F_{X_1} = F_{X_0}$ and $T_o = \nabla \psi_o$ for a convex function ψ_o, unique F_{X_1}-a.e.

Under Assumption 1 (i), McCann's theorem stated in Lemma 1 guarantees the existence of a unique Optimal Transport (OT) map T_o such that $T_o \# F_{X_1} = F_{X_0}$ and $T_o = \nabla \psi_o$ for a convex function ψ_o, unique a.e.- F_{X_1}. Assumption 1 (ii) playes the same role as the conditional independent assumption, i.e., SI (i).

To gain insight in the *domain shift assumption* in Assumption 1 (ii), consider the case $d_X = 1$. It is well known that for $d_X = 1$, $T_o(x) = F_{X_0}^{-1}(F_{X_1}(x))$, see for example Galichon (2016) or Santambrogio (2015). Thus Assumption 1 (ii) is equivalent to

$$F(y|x, D=1) = F\left(y|F_{X_0}^{-1}(F_{X_1}(x)), D=0\right) \text{ for all } x \in \mathcal{S} \subset \mathbb{R} \text{ and } y \in \mathcal{Y}. \tag{1.3}$$

(1.3) states that the conditional distribution of Y_0 given $X_1 = x$ is identical to the conditional distribution of Y_0 when X_0 takes the value in $\mathcal{X}_0$ that has the same rank as x in $\mathcal{S}$, i.e., $F_{X_0}^{-1}(F_{X_1}(x))$ for all $x \in \mathcal{S}$. The same interpretation carries over to the multivariate case, as under the conditions of Proposition 1, the transport map T_o is a F_{X_0}-quantile of F_{X_1} defined in Ekeland, Galichon, and Henry (2012) and Galichon and Henry (2012).

1.2.2 Identification in the Non-overlap Region

Assumption 1 allows us to tackle the non-overlap problem. Under Assumption 1, for $x \in \mathcal{S} \subset \mathcal{X}_1$, we have

$$\begin{aligned} E[m(Y_0;\beta)|X_1 = x, D=1] &= \int m(y;\beta) f(y|x, D=1)\, dy \\ &= \int m(y;\beta) f(y|T_o(x), D=0)\, dy \\ &= E[m(Y_0;\beta)|X_0 = T_o(x), D=0]. \end{aligned} \tag{1.4}$$

For $x' \in \mathcal{X}_0$, let

$$h(x', \beta) := \int m(y; \beta) f\left(y|x', D = 0\right) dy = E\left[m(Y_0; \beta) \mid X_0 = x', D = 0\right]. \tag{1.5}$$

Under Assumption 1 (i), the transport map T_o is identified from the samples $\{X_{1i}\}_{i=1}^{n_1}$ and $\{X_{0j}\}_{j=1}^{n_0}$, and $h(\cdot, \beta)$ is identified from the control sample $\{(Y_{0j}, X_{0j})\}_{j=1}^{n_0}$ for each $\beta \in \mathcal{B}$. As a result, $\beta \in \mathcal{B}$ is identified from the following equation (1.6), and Model (1.2) becomes

$$E[h\left(T_o(X_1), \beta\right) I\left\{X_1 \in \mathcal{S}\right\}] = 0 \text{ if and only if } \beta = \beta_{o\mathcal{S}} \in \mathcal{B}. \tag{1.6}$$

Remark 1. Our parameters of interest is the treatment effect for the treated parameter:

$$\tau_{o\mathcal{S}} := \kappa_{o\mathcal{S}} - \beta_{o\mathcal{S}}, \tag{1.7}$$

where $\kappa_{o\mathcal{S}} \in \mathcal{K}$ is identified by the moment condition (1.8) for the treated population:

$$E[m(Y_1; \kappa) I\left\{X_1 \in \mathcal{S}\right\} | D = 1] = 0 \text{ if and only if } \kappa = \kappa_{o\mathcal{S}} \in \mathcal{K}, \tag{1.8}$$

with $m(Y_1; \kappa) = Y_1 - \kappa$ for ATT, and $m(Y_1; \kappa) = I\left\{Y_1 \leq \kappa\right\} - q$ for QTT with a given quantile level $q \in (0, 1)$.

We conclude this subsection by restating our identification conditions as follows:

Assumption 2. (1) $E\left[h\left(T_o(X_1), \beta\right) I\left\{X_1 \in \mathcal{S}\right\} | D = 1\right] = 0$ has a unique solution $\beta_{o\mathcal{S}} \in \mathcal{B}$, where $\mathcal{B}$ is a closed bounded subset of $\mathbb{R}$;

(2) $E[m(Y_1; \kappa) I\left\{X_1 \in \mathcal{S}\right\} | D = 1] = 0$ has a unique solution $\kappa_{o\mathcal{S}} \in \mathcal{K}$, where $\mathcal{K}$ is a closed bounded subset of $\mathbb{R}$.

1.2.3 An Illustration of Assumption 1 (ii)

We constructed a toy model that shows the leverage of optimal transport. The horizontal axis of Figure 1.1a shows the distribution of one-dimensional covariate in the treatment

group (blue dots) and the control group (orange dots), respectively. In the treatment group, observations of covariate are drawn from a normal distribution, whereas in the control group, the observations are drawn from a uniform distribution. Besides, Figure 1.1a depicts the distribution of the potential outcomes along the vertical axis. According to Assumption 1 (ii), the potential outcome has the same distribution conditional on x in the treatment group and transported x in the control group, and therefore, the distributions on the vertical axis overlap.

Figure 1.1b depicts conditional expectation of potential outcome given covariate in the treatment and control groups, respectively. For the control group, this is the $h(x, \beta)$ function we are estimating when ATT is our parameter of interest. Figure 1.1c depicts a few optimal transport links between two sub-samples. Loosely speaking, optimal transport facilitates the transfer of subset $\mathcal{S}$ from the treatment group to the control group. Figure 1.1d depicts the observations from subset $\mathcal{S}$ in green and the transported observations in yellow. In the next section, we propose our three-step estimator.

1.3 Semiparametric Estimation of $\beta_{o\mathcal{S}}$ and $\tau_{o\mathcal{S}}$

We propose a three-step estimation of $\beta_{o\mathcal{S}}$ based on (1.5) and (1.6), and then estimate $\tau_{o\mathcal{S}}$ by a simple plug-in procedure in the final step.

Step 1. Estimate the OT map $T_o : \mathcal{X}_1 \mapsto \mathcal{X}_0$ using the two samples $\{X_{1i}\}_{i=1}^{n_1}$ and $\{X_{0j}\}_{j=1}^{n_0}$, and denote it as $\widehat{T} \in \mathcal{X}_0$ almost surely;

Step 2. For each $\beta \in \mathcal{B}$ and $x' \in \mathcal{X}_0$, estimate $h(x', \beta)$ defined in (1.5) using the control sample $\{Y_{0j}, X_{0j}\}_{j=1}^{n_0}$, and denote it as $\widehat{h}(x', \beta)$. One candidate for $\widehat{h}(x', \beta)$ is the sieve least squares (LS) estimator:

$$\widehat{h}(x', \beta) = \sum_{j=1}^{n_0} m(Y_{0j}, \beta)\, p^{J_{n_0}}(X_{0j})' \, (P_0' P_0)^{-1} \, p^{J_{n_0}}(x') \,,$$

where $p^{J_{n_0}}(x') = (p_1(x'), ..., \, p_{J_{n_0}}(x'))'$ and $P_0 = (p^{J_{n_0}}(X_{01}), ..., \, p^{J_{n_0}}(X_{0n_0}))'$ for some

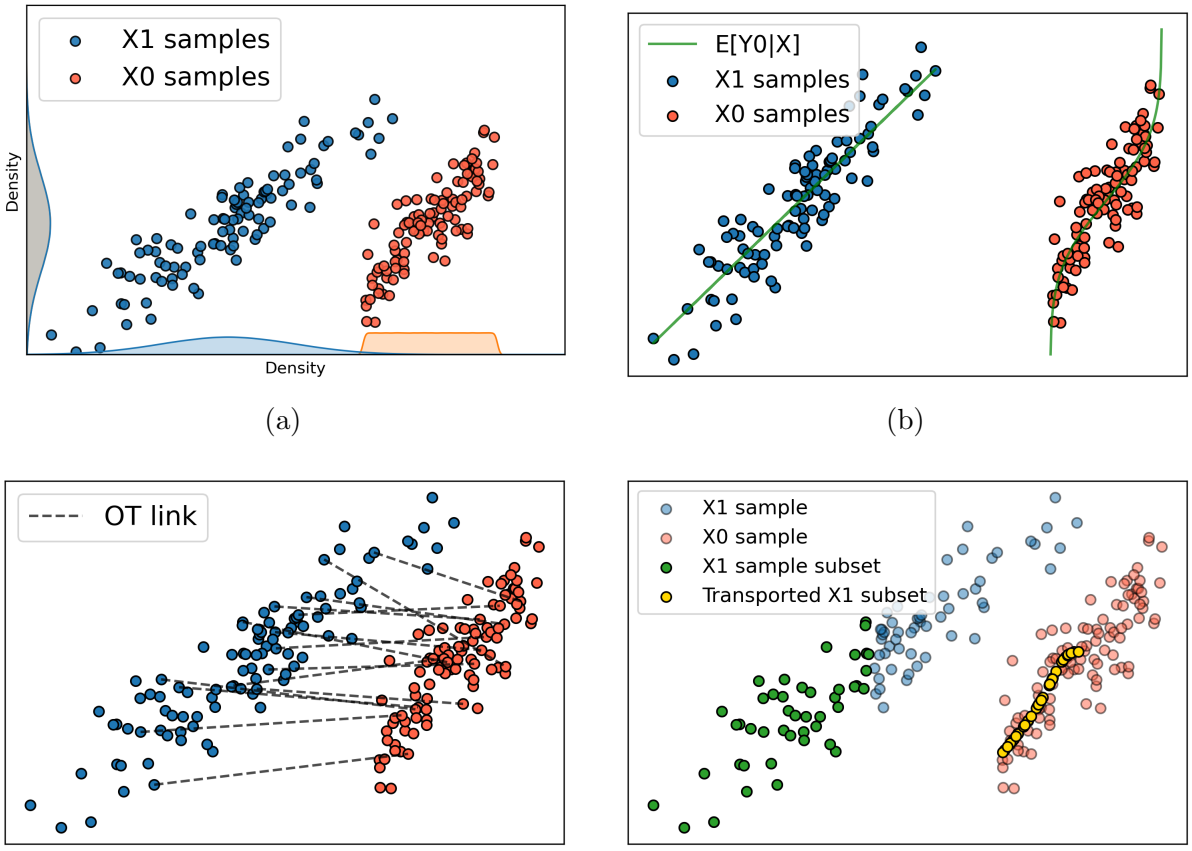


Figure 1.1: An Illustration of Assumption 1(ii)

In Figure (c), we randomly chose 15 optimal transportation links to illustrate the concept of an optimal transport (OT) map: a mapping between two distributions. Using this OT map, we can identify the subset of the population of interest based on the observations in the control group. After estimating $h(x, \beta)$ based on observations from the control group, we can estimate the parameter of interest using the transported subset, which appears as yellow dots in Figure (d).

integer J_{n_0}. Here $\left\{p_l(x'), l = 1, 2, ..., J_{n_0}\right\}$ denotes a sequence of known basis functions that can approximate any square-integrable function of $x' \in \mathcal{X}_0$ well as $J_{n_0} \to \infty$;

Step 3. Estimate $\beta_{o\mathcal{S}}$ by the Method of Moments:

$$\frac{1}{n_1}\sum_{i=1}^{n_1} \widehat{h}\left(\widehat{T}(X_{1i}), \widehat{\beta}\right) I\left\{X_{1i} \in \mathcal{S}\right\} = 0.$$

Step 4. Estimate $\tau_{o\mathcal{S}}$ by a simple plug-in estimator: $\widehat{\tau} = \widehat{\kappa} - \widehat{\beta}$, where $\widehat{\kappa} \in \mathcal{K}$ solves

$$\frac{1}{n_1}\sum_{i=1}^{n_1} [m(Y_{1i}; \widehat{\kappa}) I\left\{X_{1i} \in \mathcal{S}\right\}] = 0.$$

In Step 2, one can estimate the conditional mean function $h(x', \beta)$ using any nonparametric or machine learning procedures. We present the sieve LS estimator $\widehat{h}(x', \beta)$ for the sake of concreteness and simplicity. This will also allow us to establish asymptotic properties in the next section under low-level sufficient conditions regarding the nonparametric estimation of $h(x', \beta)$.

In the rest of this section, we present three examples of $\widehat{T}$ for the OT map T_o in Step 1. The first and the second are simple plug-in estimators of T_o that utilizes the closed-form expressions of T_o in special cases. The third is an optimization-based sieve estimator when T_o does not have a closed-form expression in general situations.

1.3.1 Example 1: $\widehat{T}$ in Univariate Nonparametric Case

When $d_X = 1$ it is known that $T_o(x) = F_{X_0}^{-1}(F_{X_1}(x))$. We can estimate $T_o(x)$ directly by

$$\widehat{T}(x) = \widehat{F}_{X_0}^{-1}\left(\widehat{F}_{X_1}(x)\right), \ x \in \mathcal{X}_1,$$

where $\widehat{F}_{X_0}, \widehat{F}_{X_1}$ are the empirical cdfs of $\{X_{0j}\}_{j=1}^{n_0}$ and $\{X_{1i}\}_{i=1}^{n_1}$ respectively, i.e.,

$$\begin{aligned}\widehat{F}_{X_0}(x) &= \frac{1}{n_0}\sum_{j=1}^{n_0} I\{X_{0j} \le x\} \text{ for } x \in \mathcal{X}_0 \text{ and} \\ \widehat{F}_{X_1}(x) &= \frac{1}{n_1}\sum_{i=1}^{n_1} I\{X_{1i} \le x\} \text{ for } x \in \mathcal{X}_1.\end{aligned}$$

Let $\{X_{0(1)} \le X_{0(2)} \le \cdots \le X_{0(n_0)}\}$ be the order statistics for $\{X_{0j}\}_{j=1}^{n_0}$ and $\{X_{1(1)} \le X_{1(2)} \le \cdots \le X_{1(n_1)}\}$ be the order statistics for $\{X_{1i}\}_{i=1}^{n_0}$. Note that

$$\begin{aligned}\widehat{F}_{X_0}^{-1}(s) &:= \inf\{t : \widehat{F}_{X_0}(t) \ge s\} \\ &= X_{0(j)} \text{ for } \frac{j-1}{n_0} < s \le \frac{j}{n_0} \text{ and } 1 \le j \le n_0.\end{aligned}$$

Further,

$$\widehat{F}_{X_1}(x) = \begin{cases} 0 \text{ for } x < X_{1(1)}, \\ \frac{i}{n_1} \text{ for } x \in [X_{1(i)}, X_{1(i+1)}) \text{ and } 1 \le i \le n_1 - 1, \\ 1 \text{ for } x \ge X_{1(n_1)}. \end{cases}$$

1.3.2 Example 2: $\widehat{T}$ in Multivariate Affine Map Case

When $d_X > 1$, there is no closed form expression for T_o unless more is known about the distributions F_{X_1} and F_{X_0}.

Condition 1. Assume F_{X_1}, F_{X_0} two distributions belongs to the class of elliptical distribution families with mean values μ_1 and μ_0 and covariance matrices Σ_1 and Σ_0, respectively.

Under Condition 1 for the class of elliptical distribution families with the same generator, Theorems 2.1 and 2.4 in Gelbrich (1990) or Theorem 2.1 in Cuesta-Albertos, Matrán-Bea, and Tuero-Diaz (1996) implies that the optimal transport map is affine, i.e.,

$$T_o(x) = \mu_0 + A(x - \mu_1) \quad \text{for any } x \in \mathcal{X}_1, \tag{1.9}$$

where $\mu_j = E(X_j)$ and $\Sigma_j = Var(X_j)$ (assumed to be positive-definite) for $j = 0, 1$, and

$$A = \Sigma_1^{-\frac{1}{2}} \left(\Sigma_1^{\frac{1}{2}} \Sigma_0 \Sigma_1^{\frac{1}{2}} \right)^{\frac{1}{2}} \Sigma_1^{-\frac{1}{2}}.$$

We note that Condition 1 implies Assumption 1 (i), and hence Brenier's Theorem is applicable, which implies that $T_o(x) = \nabla \psi_o(x)$ for the convex function

$$\psi_o(x) = \frac{1}{2} x' A x + (\mu_0 - A\mu_1)' x + const. \tag{1.10}$$

that is unique a.e. up to an additive constant.

Following Flamary, Lounici, and Ferrari (2019), we estimate T_o given in (1.9) by

$$\widehat{T}(x) = \widehat{\mu}_0 + \widehat{A}(x - \widehat{\mu}_1) \quad \text{for any } x \in \mathcal{X}_1, \tag{1.11}$$

where

$$\widehat{\mu}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} X_{ji} \quad \text{and} \quad \widehat{\Sigma}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} (X_{ji} - \widehat{\mu}_j)(X_{ji} - \widehat{\mu}_j)' \ ,$$

for $j = 0, 1$ and

$$\widehat{A} = \widehat{\Sigma}_1^{-1/2} (\widehat{\Sigma}_1^{1/2} \widehat{\Sigma}_0 \widehat{\Sigma}_1^{1/2})^{1/2} \widehat{\Sigma}_1^{-1/2}. \tag{1.12}$$

1.3.3 Example 3: Sieve $\widehat{T}$ in Multivariate Nonparametric Case

For $d_X > 1$, there is no closed form expression for the OT map T_o when F_{X_1} and F_{X_0} are fully nonparametric satisfying Assumption 1 (i) only. We will propose a spline sieve estimator of T_o in this subsection. Recall that under Assumption 1 (i), Brenier's Theorem implies that $T_o(x) = \nabla \psi_o(x)$ for a convex function ψ_o, unique a.e. F_{X_1} up to an additive constant, where ψ_o is a solution to the semi-dual problem:

$$\min_{\psi} \int_{\mathcal{X}_1} \psi(x) \, dF_{X_1}(x) + \int_{\mathcal{X}_0} \psi^*(y) \, dF_{X_0}(y) \ \text{ s.t. } \psi \in L^1(F_{X_1}), \tag{1.13}$$

where ψ^* is the convex conjugate or the Legendre-Fenchel conjugate of ψ :

$$\psi^*(y) = \sup_{x \in \mathcal{X}_1} (\langle x, y \rangle - \psi(x)). \tag{1.14}$$

Recently Hütter and Rigollet (2019) proposed a wavelet sieve estimator of ψ_o using the criterion (1.13). Since their wavelet estimator is difficult to compute when $d_X \geq 3$ we propose a spline sieve estimator instead. First we introduce the following condition, which is the same as Assumptions C1 and C2 of Hütter and Rigollet (2019). Recall that f_{X_j} is the density of F_{X_j} with respect to the Lebesgue measure on $\mathbb{R}^{d_X}$ for $j = 0, 1$.

Condition 2. For $j = 0, 1$ and for some $\alpha > 1$,

1. $\mathcal{X}_j$ is the closure of a uniformly convex, bounded, open subset of $\mathbb{R}^{d_X}$ with $C^{\lfloor \alpha - 1 \rfloor + 2}$ boundary;

2. $f_{X_j} \in C^{\alpha-1}(\mathcal{X}_j)$, and f_{X_j} is bounded from above and below on its support $\mathcal{X}_j$.

We note that Condition 2 also implies Assumption 1 (i) is satisfied. Under Condition 2, Caffarelli's global regularity theorem implies that $\psi_0 \in C^{\alpha+1}(\mathcal{X}_1)$ for $\alpha > 1$; see Villani (2009) Theorem 12.50. Let $\widetilde{\mathcal{X}_1} = \mathcal{X}_1 + \epsilon B(0, 1)$ for an $\epsilon > 0$. Denote by $\widetilde{\psi}_o$ the extension of ψ_0 to $\widetilde{\mathcal{X}_1}$, using the same notation as in Lemma 34 in Hütter and Rigollet (2019). Let $\widetilde{T}_o(x) = \nabla \widetilde{\psi}_o(x)$. Then it follows from Corollary 35 and Proposition-Definition 7 in Hütter and Rigollet (2019) that $\psi_0 \in \mathcal{X}(M)$ for a finite constant M, where $\mathcal{X}(M)$ is the set of all twice continously differentiable functions $\psi : \widetilde{\mathcal{X}_1} \to \mathbb{R}$ such that

(i) $|\psi(x)| \leq 2M^2$ and $|\nabla \psi(x)| \leq M$ for all $x \in \widetilde{\mathcal{X}_1}$,

(ii) $M^{-1} \preceq D^2 \psi(x) \preceq M$ for all $x \in \widetilde{\mathcal{X}_1}$.

Let $\Psi = \left\{ \psi \in X(2M) \cap C^{\alpha+1}\left(\widetilde{\mathcal{X}_1}\right) : \int \psi(x) dx = 0 \right\}$. Let $\{b_i\}_{i=1}^{\infty}$ be a complete basis for the infinite dimensional Hilbert space $(\Psi, \|\cdot\|_\Psi)$ and let $B_{k_{n_1}}(\cdot) = \left(b_1(\cdot), \ldots, b_{k_{n_1}}(\cdot)\right)'$. Denote

$$\Psi_n \equiv \left\{ \psi(\cdot) = B_{k_{n_1}}(\cdot)'\gamma \in \Psi : \gamma \in \mathbb{R}^{k_{n_1}} \right\}.$$

Let $k_{n_1} = \dim(\Psi_n)$.

A sieve estimator of T_o is defined as

$$\widehat{T}(x) = \nabla\widehat{\psi}(x), \text{ where } \widehat{Q}\left(\widehat{\psi}\right) \leq \inf_{\psi\in\Psi_n} \widehat{Q}(\psi) + o_p\left(n^{-1}\right), \tag{1.15}$$

and

$$\begin{aligned}
\widehat{Q}(\psi) &= \int_{x\in\mathcal{X}_1} \psi(x)\, d\widehat{F}_{X_1}(x) + \int_{y\in\mathcal{X}_0} \left[\sup_{z\in\widetilde{\mathcal{X}_1}} (\langle z, y\rangle - \psi(z))\right] d\widehat{F}_{X_0}(y) && (1.16)\\
&= \frac{1}{n_1}\sum_{i=1}^{n_1} \psi(X_{1i}) + \frac{1}{n_0}\sum_{i=1}^{n_0} \left[\sup_{z\in\widetilde{\mathcal{X}_1}} (\langle z, X_{0i}\rangle - \psi(z))\right]. && (1.17)
\end{aligned}$$

Hütter and Rigollet (2019) establishes minimax optimal convergence rate of $\widehat{\psi}$ and $\widehat{T}$ when Ψ_n is a wavelet sieve. Since their wavelet estimator is computationally very demanding as soon as $d_X \geq 3$, in our Monte Carlo studies we use spline sieve instead.

A Closed-form Spline Estimator

Let $x = (x_1, \ldots, x_{d_x})$. If the covariates $x_1, \ldots, x_d$ are independent and quadratic spline basis is used to estimate ψ, $\widehat{\psi}^*(y) = \sup_{x\in\mathcal{X}}\left(\langle x, y\rangle - \widehat{\psi}(x)\right)$ admits a closed form. Recall that when covariates are independent, $\widehat{\psi}$ using quadratic spline basis can be written as

$$\widehat{\psi}(x) = a_0 + \sum_{d=1}^{d_x}\left\{a_1^d x_d + a_2^d x_d^2 + \sum_{j=1}^{M_n} b_j^d \left(x_d - t_j^d\right)_+^2\right\},$$

where d_x is the dimension of x, and M_n is the number of knots associated with the spline basis for each dimension. Without loss of generality, we assume that $\mathcal{X}$ can be characterized

by $\underline{M}_d \leq x_d \leq \overline{M}_d$ for $d = 1, \ldots, d_x$.[1] Let $t_0^d = \underline{M}_d$, $t_{M_n+1}^d = \overline{M}_d$ and $b_0^d = 0$. Let

$$f_d(x_d) := \left(a_1^d - y_d\right) x_d - a_2^d x_d^2 - \sum_{j=0}^{M_n} b_j^d \left(x_d - t_j^d\right)_+^2, \text{ for } d = 1, \ldots, d_x.$$

For $J \in \{0, \ldots, M_n\}$, let

$$x_{d,J}^* := \begin{cases} \frac{(y_d - a_1^d)/2 + \sum_{j=0}^J b_j^d \cdot t_j^d}{a_2^d + \sum_{j=0}^J b_j^d} & \text{if } t_J^d \leq \frac{(y_d - a_1^d)/2 + \sum_{j=0}^J b_j^d \cdot t_j^d}{a_2^d + \sum_{j=0}^J b_j^d} < t_{J+1}^d; \\ t_J^d & \text{if } 2a_2^d t_J^d + a_1^d y_d + 2\sum_{j=0}^J b_j^d \left(t_J^d - t_j^d\right) > 0; \\ \varnothing & \text{else.} \end{cases}$$

Then

$$\widehat{\psi}^*(y) = -a_0 + \sum_{d=1}^{d_x} f_d^{max},$$

where

$$f_d^{max} = \max_{x_d \in \left\{x_{d,1}^*, \ldots, x_{d,M_n}^*, \overline{M}_d\right\}} f_d(x_d).$$

1.4 Asymptotic Properties of $\widehat{\beta}$ and $\widehat{\tau}$

In this section, we establish the asymptotic properties of $\widehat{\beta}$ and $\widehat{\tau}$ under low-level sufficient conditions on the sieve Least Squares (LS) estimator $\widehat{h}$ for h, but high-level assumptions on any OT estimator $\widehat{T}$ for T_o. We will then provide low-level sufficient conditions for these high-level assumptions on $\widehat{T}$ in Examples 1-3 in the next section.

1.4.1 Smoothness Class for h

To account for the effect of estimating $h : \mathbb{R}^{d_X} \to \mathbb{R}$ by the sieve Least Squares estimator $\widehat{h}$ given in Step 2, we follow Chen, Hong, and Tamer (2005) by imposing smoothness on $h(\cdot, \beta)$ for each $\beta \in \mathcal{B}$. A typical smoothness assumption is that a function belongs to a

[1] As long as $\mathcal{X}$ can be characterized by inequalities, the convex conjugate for the spline estimator described above admits closed-form.

Hölder space. For any $1 \times d_X$ vector $\mathbf{a} = (a_1, \ldots, a_{d_X})$ of non-negative integers, we write $|\mathbf{a}| = \sum_{k=1}^{d_X} a_k$, and for any $x = (x_1, \ldots, x_{d_X})' \in \mathcal{X} \subseteq \mathbb{R}^{d_X}$, we denote the $|\mathbf{a}|$-th derivative of a function $h : \mathcal{X} \to \mathbb{R}$ as

$$\nabla^{\mathbf{a}} h(x) = \frac{\partial^{|\mathbf{a}|}}{\partial x_1^{a_1} \ldots \partial x_d^{a_{d_X}}} h(x).$$

For some $\gamma > 0$, let $\underline{\gamma}$ be the largest integer smaller than γ, and let $\Lambda^\gamma(\mathcal{X})$ denote a Hölder space with smoothness γ, i.e., a space of functions $h : \mathcal{X} \to \mathbb{R}$ which have up to $\underline{\gamma}$-th continuous derivatives, and the highest $\underline{\gamma}$-th derivatives are Hölder continuous with the Hölder exponent $\gamma - \underline{\gamma} \in (0, 1]$. The Hölder space becomes a Banach space when endowed with the Hölder norm:

$$\|h\|_{\Lambda^\gamma} = \sup_x |h(x)| + \max_{|\mathbf{a}|=\underline{\gamma}} \sup_{x \neq \overline{x}} \frac{|\nabla^{\mathbf{a}} h(x) - \nabla^{\mathbf{a}} h(\overline{x})|}{\sqrt{(x-\overline{x})'(x-\overline{x})}^{\gamma-\underline{\gamma}}} < \infty.$$

Following Chen, Hong, and Tamer (2005), we let $\Lambda^\gamma(\mathcal{X}, \omega_1)$ denote a weighted Hölder space of functions $h : \mathcal{X} \to \mathbb{R}$ such that $h(\cdot)\left[1 + |\cdot|^2\right]^{-\omega_1/2}$ is in $\Lambda^\gamma(\mathcal{X})$. We call

$$\Lambda_c^\gamma(\mathcal{X}, \omega_1) \equiv \left\{ h \in \Lambda^\gamma(\mathcal{X}, \omega_1) : \left\| h(\cdot)\left[1 + |\cdot|^2\right]^{-\omega_1/2} \right\|_{\Lambda^\gamma} \leq c < \infty \right\}$$

a weighted Hölder ball (with radius c). We say a function $h : \mathcal{X} \to \mathbb{R}$ is $H(\gamma, \omega_1)$-smooth if it belongs to a weighted Hölder ball $\Lambda_c^\gamma(\mathcal{X}, \omega_1)$ for some $\gamma > 0$ and $\omega_1 \geq 0$. As discussed in Chen, Hong, and Tamer (2005), the weighted Holder ball with $\omega_1 = 0$ reduces to the standard Hölder ball $\Lambda_c^\gamma(\mathcal{X})$, which is a typical sufficient condition especially when the support $\mathcal{X}$ is a bounded subset of $\mathbb{R}^{d_X}$. However, when $\mathcal{X} = \mathbb{R}^{d_X}$, the standard Hölder ball $\Lambda_c^\gamma(\mathcal{X})$ may exclude some functions such as $h(x, \beta) = x'\iota - \beta$. It is clear that the weighted Hölder ball $\Lambda_c^\gamma(\mathcal{X}, \omega_1)$ with $\omega_1 > 0$ is a strictly larger space and $x'\iota - \beta \in \Lambda_c^\gamma\left(\mathbb{R}^{d_X}, \omega_1\right)$ with $\omega_1 = 1$.

For $j = 0, 1$, for any square measurable function $h : \mathcal{X}_j \to \mathbb{R}$, we define a Hilbert norm $\|h\|_{2,j} \equiv \sqrt{\int_{\mathcal{X}_j} h(x)^2 f_{X_0}(x) dx} < \infty$, and $\mathcal{L}_2(\mathcal{X}_j) = \{h : \mathcal{X}_j \to \mathbb{R} : \|h\|_{2,j} < \infty\}$ the

corresponding Hilbert space. For $h(x,\beta)$, $x \in \mathcal{X}_0$, $\beta \in \mathcal{B}$ defined in (1.5), we denote

$$\begin{aligned} \|h\|_{\infty,\omega} &= \sup_{x\in\mathcal{X}_0,\beta\in\mathcal{B}} \left| h(x,\beta)\left[1+|x|^2\right]^{-\omega/2}\right| \text{ and} \\ \|h\left(T_o\left(\cdot\right),\beta\right)\|_{\infty,\omega} &= \sup_{x\in\mathcal{X}_1} \left| h(T_o\left(x\right),\beta)\left[1+|T_o\left(x\right)|^2\right]^{-\omega/2}\right|. \end{aligned}$$

1.4.2 Consistency of $\widehat{\beta}$

Assumption 3. Let the following holds:

1. Let $n_0 := \sum_{i=1}^n I\left\{D_i = 0\right\} \to \infty$, $n_1 := \sum_{i=1}^n I\left\{D_i = 1\right\} \to \infty$, and $\lambda = \lim_{n_1\to\infty}\left(n_1/n_0\right) \in [0,\infty)$, where $n = n_0 + n_1$;

2. The two i.i.d samples $\{(Y_{0i}, X_{0i})\}_{i=1}^{n_0}$ and $\{(Y_{1i}, X_{1i})\}_{i=1}^{n_1}$ are independent.

Assumption 4. Let the following holds:

1. For all $\beta \in \mathcal{B}$, $h(\cdot,\beta)$ defined in (1.5) is $H\left(\gamma,\omega_1\right)$-smooth for some $\gamma > 0$, $\omega_1 \geq 0$;

2. $\int_{\mathcal{X}_1}\left(1+|x|^2\right)^{\omega} f_{X_1}(x)dx < \infty$ and $\int_{\mathcal{X}_0}\left(1+|x|^2\right)^{\omega} f_{X_0}(x)dx < \infty$ for some $\omega > \omega_1 \geq 0$;

3. For each fixed $x \in \mathcal{X}_0$, $h(x,\beta)$ is continuous at β for all $\beta \in \mathcal{B}$;

4. $\mathrm{Var}\left[\left\{m\left(Y_0,\beta\right) - h\left(X_0,\beta\right)\right\} \mid X_0 = x, D = 0\right]$ is bounded uniformly over $x \in T_o(\mathcal{S}) \subset \mathcal{X}_0$ and $\beta \in \mathcal{B}$;

5. For any $H\left(\gamma,\omega_1\right)$-smooth function $h(\cdot,\beta) : \mathcal{X}_0 \to \mathbb{R}$, there is a function $\Pi_{\infty n}h$ in the sieve space $\mathcal{H}_n = \left\{f(\cdot,\beta) = p^{J_{n_0}}(\cdot)'\pi(\beta)\right\}$ such that $\|h(\cdot,\beta) - \Pi_{\infty n}h(\cdot,\beta)\|_{\infty,\omega} = o(1)$. Also $E\left[p^{J_{n_0}}(X_0)p^{J_{n_0}}(X_0)'\right]$ is non-singular uniformly in J_{n_0}. Let $\frac{J_{n_0}}{n_0} \to 0$ and $J_{n_0} \to \infty$.

Assumption 5. Suppose $\widehat{T}\left(x\right) \in \mathcal{X}_0$ w.p.1 for each $x \in \mathcal{X}_1$, and

$$\|h\left(\widehat{T}\left(\cdot\right),\beta\right) - h\left(T_o\left(\cdot\right),\beta\right)\|_{\infty,\omega} = o_p\left(1\right) \text{ for all } \beta \in \mathcal{B}.$$

Theorem 1. Let Assumptions 1, 2.1, 3, 4 and 5 hold. Then: $\widehat{\beta} - \beta_{o\mathcal{S}} = o_p(1)$.

1.4.3 Asymptotic Linear Expansion and Normality of $\widehat{\beta}$

Additional assumptions are imposed to establish the asymptotic normality of $\widehat{\beta}$.

Assumption 6. Denote $\|\boldsymbol{\eta}(\cdot)\|_{2,1} = \left(\|\eta_1(\cdot)\|_{2,1}, \|\eta_2(\cdot)\|_{2,1}, \ldots \|\eta_d(\cdot)\|_{2,1}\right)'$, for $\boldsymbol{\eta}(\cdot) = (\eta_1(\cdot), \ldots, \eta_d(\cdot))'$. Assume following holds:

$$\left\| \frac{\partial \left\{ \widehat{h}\left(T_o(\cdot), \beta_{o\mathcal{S}}\right) - h\left(T_o(\cdot), \beta_{o\mathcal{S}}\right) \right\}}{\partial T} \right\|_{2,1} \cdot \left\| \widehat{T}(\cdot) - T_o(\cdot) \right\|_{2,1} = o_p\left(n_1^{-1/2}\right).$$

Assumption 7. There exist functions $\varphi_1 \in L^2\left(F_{X_1}\right)$ and $\varphi_0 \in L^2\left(F_{X_0}\right)$ such that

$$\begin{aligned} & \frac{1}{n_1} \sum_{i=1}^{n_1} \left(h(\widehat{T}(X_{1i}), \beta_{o\mathcal{S}}) I\left\{X_{1i} \in \mathcal{S}\right\} - h\left(T_o(X_{1i}), \beta_{o\mathcal{S}}\right) I\left\{X_{1i} \in \mathcal{S}\right\} \right) \\ = \; & \frac{1}{n_1} \sum_{i=1}^{n_1} \varphi_1\left(X_{1i}\right) + \frac{1}{n_0} \sum_{j=1}^{n_0} \varphi_0\left(X_{0j}\right) + o_p\left(n_1^{-1/2}\right), \end{aligned}$$

where $E\left[\varphi_1\left(X_1\right)\right] = 0$ and $E\left[\varphi_0\left(X_0\right)\right] = 0$.

Assumptions 6 and 7 are high-level assumption on $\widehat{T}$ and will be verified in the next section for Examples 1-3.

Assumption 8. Let $\beta_{o\mathcal{S}}$ be an interior point of $\mathcal{B}$. Suppose

1. G^2 is finite and positive, where $G = \frac{\partial}{\partial \beta} E\left[h\left(T_o(X_1), \beta_{o\mathcal{S}}\right) I\left\{X_1 \in \mathcal{S}\right\}\right]$;

2. Ω_1 and Ω_0 are finite and positive, where

$$\begin{aligned} \Omega_1 &= E\left[h\left(T_o(X_1), \beta_{o\mathcal{S}}\right) I\left\{X_1 \in \mathcal{S}\right\} + \varphi_1\left(X_1\right)\right]^2 \text{ and} \\ \Omega_0 &= E\left[\varphi_0\left(X_0\right) + \left(m\left(Y_0; \beta_{o\mathcal{S}}\right) - h\left(X_0, \beta_{o\mathcal{S}}\right)\right) I\left\{X_0 \in T_o(\mathcal{S})\right\}\right]^2; \end{aligned}$$

3. For each fixed $x \in \mathcal{X}_0$, and for some $\delta > 0$, $\frac{\partial h(x,\beta)}{\partial \beta}$ is continuous in $\beta \in \mathcal{B}$ with

$$E\left[\sup_{\beta:|\beta-\beta_{o\mathcal{S}}|\leq\delta}\left|\frac{\partial h\left(x,\beta\right)}{\partial\beta}\right|\right] < \infty;$$

4. There exist a constant $\epsilon \in (0,1]$, a $\delta > 0$, and a measurable function $b(\cdot)$ with $E\left[b\left(X_0\right)\right] < \infty$ such that $\left|\frac{\partial \widetilde{h}(x,\beta)}{\partial\beta} - \frac{\partial h(x,\beta)}{\partial\beta}\right| \leq b\left(x\right)\left[\|\widetilde{h}-h\|_{\infty,\omega}\right]^{\epsilon}$ for all $x \in \mathcal{X}_0$ and $\beta \in \mathcal{B}$ with $|\beta - \beta_{o\mathcal{S}}| \leq \delta$ and all $H\left(\gamma,\omega_1\right)$-smooth function $\widetilde{h}$ with $\|\widetilde{h}-h\|_{\infty,\omega} \leq \delta$.

Assumption 9. Let the following hold:

1. Assumption 4.1 is satisfied with $\gamma > d_X/2$, and Assumption 4.2 is satisfied with $\omega > \omega_1 + \gamma$;
2. $J_{n_0} = O\left(n_0^{\frac{d_X}{2\gamma+d_X}}\right)$.

Theorem 2. Let $\widehat{\beta} - \beta_{o\mathcal{S}} = o_p(1)$ and Assumptions 3, 6, 7, 8 and 9 hold. Then:

$$\begin{aligned}
&\sqrt{n_1}\left(\widehat{\beta}-\beta_{o\mathcal{S}}\right)\\
&\quad= -G^{-1}\left\{\frac{1}{\sqrt{n_1}}\sum_{i=1}^{n_1}\left[h\left(T_o(X_{1i}),\beta_{o\mathcal{S}}\right)I\left\{X_{1i}\in\mathcal{S}\right\}+\varphi_1\left(X_{1i}\right)\right]\right.\\
&\qquad\qquad\left.+\frac{\sqrt{n_1}}{n_0}\sum_{j=1}^{n_0}\left[\varphi_0\left(X_{0j}\right)+\left(m\left(Y_{0j},\beta_o\right)-h\left(X_{0j},\beta_{o\mathcal{S}}\right)\right)I\left\{X_{0j}\in T_o(\mathcal{S})\right\}\right]+o_p\left(1\right)\right\}\\
&\quad\xrightarrow{d}\mathcal{N}(0,V_\beta),
\end{aligned}$$

where $V_\beta = G^{-2}\left[\Omega_1+\lambda\Omega_0\right]$ and

$$\begin{aligned}
&\Omega_1+\lambda\Omega_0\\
&\quad=\mathrm{Avar}\left\{\frac{1}{\sqrt{n_1}}\sum_{i=1}^{n_1}\left[h\left(T_o(X_{1i}),\beta_{o\mathcal{S}}\right)I\left\{X_{1i}\in\mathcal{S}\right\}+\varphi_1\left(X_{1i}\right)\right]\right.
\end{aligned}$$

$$+\frac{\sqrt{n_1}}{n_0}\sum_{j=1}^{n_0}\left[\varphi_0\left(X_{0j}\right)+\left(m\left(Y_{0j},\beta_{o\mathcal{S}}\right)-h\left(X_{0j},\beta_{o\mathcal{S}}\right)\right)I\left\{X_{0j}\in T_o(\mathcal{S})\right\}\right]\Bigg\}. \quad (1.18)$$

The effect of estimating T_o is incorporated through $\varphi_1\left(X_1\right)$ and $\varphi_0\left(X_0\right)$ in (1.18). The effect of estimating h is characterized by the term $m\left(Y_{0j},\beta_{o\mathcal{S}}\right)-h\left(X_{0j},\beta_{o\mathcal{S}}\right)$.

1.4.4 Asymptotic properties of the plug-in estimator $\widehat{\tau}$ for $\tau_{o\mathcal{S}}$

Recall that $\widehat{\tau}=\widehat{\kappa}-\widehat{\beta}$ in Step 4 is our simple plug-in estimator for the general treatment effect on the treated parameter of interest $\tau_{o\mathcal{S}}=\kappa_{o\mathcal{S}}-\beta_{o\mathcal{S}}$ where $\kappa_{o\mathcal{S}}\in\mathcal{K}$ is identified by Assumption 2.2.

Corollary 1. Let $\widehat{\beta}-\beta_{o\mathcal{S}}=o_p(1)$, Assumptions 2.2 and 3 hold. Suppose that $m\left(Y_1;\kappa\right)$ is continuous in $\kappa\in\mathcal{K}$ for all y in the support of Y_1, and $E\left[\sup_{\kappa\in\mathcal{K}}\left|m(Y_1;\kappa)\right|\right]<\infty$. Then: $\widehat{\tau}-\tau_{o\mathcal{S}}=o_p\left(1\right)$.

Assumption 10. Let $\kappa_{o\mathcal{S}}$ be an interior point of $\mathcal{K}$,

1. M^2 is finite and positive, where $M=\frac{\partial}{\partial\kappa}E\left[m\left(Y_1;\kappa_{o\mathcal{S}}\right)I\left\{X_1\in\mathcal{S}\right\}\right]$;

2. $E\left(\left[m\left(Y_1;\kappa_{o\mathcal{S}}\right)I\left\{X_1\in\mathcal{S}\right\}\right]^2\right)$ is finite and positive;

3. For some $\delta>0$, $\frac{\partial}{\partial\kappa}E\left[m(Y_1;\kappa)I\left\{X_1\in\mathcal{S}\right\}\right]$ is continuous in $\kappa\in\mathcal{K}$ with

$$E\left[\sup_{\kappa:|\kappa-\kappa_{o\mathcal{S}}|\leq\delta}\left|\frac{\partial}{\partial\kappa}E\left[m(Y_1;\kappa)I\left\{X_1\in\mathcal{S}\right\}\right]\right|\right]<\infty.$$

Corollary 2. Let $\widehat{\tau}-\tau_{o\mathcal{S}}=o_p\left(1\right)$, Assumption 10 and Assumptions for Theorem 2 hold. Then:

$$\sqrt{n_1}\left(\widehat{\tau}-\tau_{o\mathcal{S}}\right)\xrightarrow{d}\mathcal{N}\left(0,V_\tau\right),$$

where $V_\tau = G^{-2}\left[\Omega_{\tau,1} + \lambda\Omega_0\right]$, and

$$\Omega_{\tau,1} = E\left[-\frac{G}{M} m\left(Y_1; \kappa_{o\mathcal{S}}\right) I\left\{X_1 \in \mathcal{S}\right\} + h\left(T_o(X_{1i}), \beta_{o\mathcal{S}}\right) I\left\{X_1 \in \mathcal{S}\right\} + \varphi_1\left(X_1\right)\right]^2.$$

1.5 *Verification of Assumptions 5, 6 and 7 for Different* $\widehat{T}$

Assumption 11. Suppose for ω in Assumptions 4.1 and 4.2, $h\left(T_o\left(\cdot\right), \beta\right)$ is continuous in T_o uniformly in $\beta \in \mathcal{B}$ under the norm $\left\|h\left(T_o\left(\cdot\right), \beta\right)\right\|_{\infty,\omega}$.

1.5.1 *Example 1: Univariate Nonparametric Case (Cont'd)*

Condition 3. The density function f_{X_j} is continuously differentiable, bounded from above by a finite constant $\overline{f}_{X_j} > 0$, bounded from below by a finite constant $\underline{f}_{X_j} > 0$ with support $\mathcal{X}_j = \left[\underline{x}_j, \overline{x}_j\right]$, where $\underline{x}_j, \overline{x}_j$ are finite constants and $\underline{x}_j < \overline{x}_j$ for $j = 0, 1$.

Denote

$$Q(x, X_1, \beta_{o\mathcal{S}}) = -\frac{h_1\left(T_o\left(X_1\right), \beta_{oS}\right)}{f_{X_0}\left(T_o\left(X_1\right)\right)} \times \left[I\left\{x \leq T_o\left(X_1\right)\right\} - F_{X_1}(X_1)\right] \times I\left\{X_1 \in \mathcal{S}\right\}, \quad x \in \mathcal{X}_0,$$

where $T_o\left(x\right) = F_{X_0}^{-1}\left(F_{X_1}(x)\right)$. Further let

$$q(x, \beta_{o\mathcal{S}}) = E\left[Q\left(x, X_1, \beta_{o\mathcal{S}}\right)\right], \quad x \in \mathcal{X}_0.$$

When the subset $\mathcal{S} = [\underline{\mathcal{S}}, \overline{\mathcal{S}}]$, where $\underline{\mathcal{S}}$, $\overline{\mathcal{S}}$ are finite constants, $q\left(x, \beta_{o\mathcal{S}}\right)$ can be simplified to

$$\begin{aligned} q\left(x, \beta_{o\mathcal{S}}\right) =& h\left(x, \beta_{oS}\right) \times I\left\{\underline{\mathcal{S}} \leq T_o^{-1}(x) \leq \overline{\mathcal{S}}\right\} - h\left(T_o(\overline{\mathcal{S}}), \beta_{o\mathcal{S}}\right) \times I\left\{\underline{\mathcal{S}} \leq T_o^{-1}(x) \leq \overline{\mathcal{S}}\right\} \\ & - h\left(T_o(\overline{\mathcal{S}}), \beta_{o\mathcal{S}}\right) \times I\left\{T_o^{-1}(x) \leq \underline{\mathcal{S}}\right\} + h\left(T_o(\underline{\mathcal{S}}), \beta_{oS}\right) \times I\left\{T_o^{-1}(x) \leq \underline{\mathcal{S}}\right\} \\ & + h\left(T_o(\overline{\mathcal{S}}), \beta_{oS}\right) \times F_{X_1}(\overline{\mathcal{S}}) - h\left(T_o(\underline{\mathcal{S}}), \beta_{o\mathcal{S}}\right) \times F_{X_1}(\underline{\mathcal{S}}). \end{aligned}$$

Proposition 1. Assume Condition 3, Assumption 4.1 and 11 hold, then Assumption 5 in section 1.4.2 holds. Further assume 3.2, 4 and 9 hold, then Assumption 6 in section 1.4.3 holds.

Proposition 2. Suppose Condition 3, Assumption 4.2 and Assumption 4.5 hold. Then Assumption 7 holds with

$$\varphi_1(X_1) = -q(T_o(X_1), \beta_{o\mathcal{S}}); \quad \varphi_0(X_0) = q(X_0, \beta_{o\mathcal{S}}).$$

1.5.2 Example 2: Multivariate Affine Map (Cont'd)

Recall $\mu_j = E(X_j), \Sigma_j = E\left[(x-\mu_j)(x-\mu_j)'\right]$ and for $j = 0, 1$.

Proposition 3. Assume Condition 1, Assumption 4.1 and 11 hold, then Assumption 5 in section 1.4.2 holds. Further assume Assumption 3.2, 4 and 9 hold, then Assumption 6 in section 1.4.3 holds.

Let $\Sigma_j(x) = (x-\mu_j)(x-\mu_j)'$. Denote the eigendecomposition $\Sigma_1\Sigma_0 = UDU^{-1}$, λ_d as the d-th eigenvalue of diagonal matrix D. Define matrices L and K, with elements $[L]_{d,s} := \frac{1}{\sqrt{\lambda_d}+\sqrt{\lambda_s}}$ and $[K]_{d,s} := \frac{\sqrt{\lambda_d\lambda_s}}{\sqrt{\lambda_d}+\sqrt{\lambda_s}}$ as the d-th and s-th element, constructed using the eigenvalue of diagonal matrix D. Denote $A \circ B$ the Hadamard product between matrices.

Proposition 4. Suppose Condition 1, Assumption 4.2 and Assumption 4.5 hold, then Assumption 7 holds with

$$\begin{aligned}\varphi_1(x_1) =& -E\left[I\{X_1 \in \mathcal{S}\} h_1(T_o(X_1), \beta_{o\mathcal{S}})' A(x_1-\mu_1)\right] \\ & - E\left[h_1(T_o(X_1), \beta_{o\mathcal{S}})' \left(\Sigma_0 U\left[K \circ \left(U^{-1}\Sigma_0^{-1}\Sigma_1^{-1}(\Sigma_1(x_1)-\Sigma_1)\Sigma_1^{-1}U\right)\right]U^{-1}\right)(X_1-\mu_1) I\{X_1 \in \mathcal{S}\}\right.\end{aligned}$$

$$\begin{aligned}\varphi_0(x_0) =& E\left[I\{X_1 \in \mathcal{S}\} h_1(T_o(X_1), \beta_{o\mathcal{S}})'(x_0-\mu_0)\right] \\ & + E\left[h_1(T_o(X_1), \beta_{o\mathcal{S}})' \left(\Sigma_1^{-1}U\left[L \circ \left(U^{-1}\Sigma_1(\Sigma_0(x_0)-\Sigma_0)U\right)\right]U^{-1}\right)\right)(X_1-\mu_1) I\{X_1 \in \mathcal{S}\}\Big].\end{aligned}$$

1.5.3 Example 3: Multivariate Nonparametric Case (Cont'd)

Note that $E\left[h\left(\nabla\psi(X_1), \beta_{o\mathcal{S}}\right) I\left\{X_1 \in \mathcal{S}\right\}\right]$ is a smooth functional of ψ. We follow Chen and Liao (2015) to verify Assumption 7 in three steps.

Step 1. Define a weak norm for the perturbation space. Let $\Psi = \mathcal{X}\left(2M\right) \cap C^{\alpha+1}\left(\widetilde{\mathcal{X}_1}\right)$ be endowed with a pseudo-metric $\|\cdot\|_\Psi = \|\cdot\|_{2,1}$. Let $\mathcal{V}$ be the closed linear span of $\Psi - \{\psi_o\}$ under norm $\|\cdot\|$, which will be described later. Since ψ_o is the unique minimizer of $Q(\psi)$ over Ψ up to an additive constant, within any shrinking $\|\cdot\|_\Psi$-neighborhood, Ψ_o, of ψ_o, we can define a local pseudo-metric for $v = \psi - \psi_o$ as

$$\|\psi - \psi_o\| = \left\{ \left[\frac{\partial^2}{\partial \tau^2} Q\left(\psi_o + \tau\left(\psi - \psi_o\right)\right) \right] \Big|_{\tau=0} \right\}^{1/2},$$

where $\|\psi - \psi_o\| \leq \text{const.} \times \|\psi - \psi_o\|_\Psi$ for any $\psi \in \Psi_o$.

Recall $\{b_i\}_{i=1}^{\infty}$ a complete basis for the infinite dimensional Hilbert space $(\mathcal{V}, \|\cdot\|)$ and let $B_{k_{n_1}}(\cdot) = \left(b_1(\cdot), \ldots, b_{k_{n_1}}(\cdot)\right)'$. Then $\mathcal{V}_n = \left\{v(\cdot) = B_{k_{n_1}}(\cdot)'\gamma : \gamma \in \mathbb{R}^{k_{n_1}}\right\}$ becomes dense in $(\mathcal{V}, \|\cdot\|)$ as $k_{n_1} \to \infty$.

Lemma 2. Assume Condition 2 holds, then

$$\|v\|^2 = E\left[\left\{\nabla v(X_1)\right\}^T \left[\nabla^2 \psi_o(X_1)\right]^{-1} \nabla v(X_1)\right].$$

Proof. Following Lemma 34 in Hütter and Rigollet (2019), there exists an $\varepsilon > 0$ and an extension $\widetilde{\psi}_o$ of ψ_o to $\widetilde{\mathcal{X}}_1 = \mathcal{X}_1 + \varepsilon B(0,1)$. Since $\widetilde{\psi}_o$ is strongly convex on an open set, we get the result directly from Proposition 2.2 in Delalande and Merigot (2021).

□

Step 2. Compute the Riesz representor of $E\left[h\left(\nabla\psi(X_1), \beta_{o\mathcal{S}}\right) I\left\{X_1 \in \mathcal{S}\right\}\right]$.

For any $v, \widetilde{v} \in \mathcal{V}$, the $\|\cdot\|$ induced inner product is given by

$$\langle v, \widetilde{v} \rangle = E\left[\{\nabla v(X_1)\}^T \left[\nabla^2 \psi_o(X_1)\right]^{-1} \nabla \widetilde{v}(X_1)\right] = E\left[\{\nabla v(X_1)\}^T \left[\nabla T_o(X_1)\right]^{-1} \nabla \widetilde{v}(X_1)\right].$$

We say that $E\left[h\left(\nabla \psi(X_1), \beta_{o\mathcal{S}}\right) I\left\{X_1 \in \mathcal{S}\right\}\right]$ is pathwise differentiable at $\psi_o \in \Psi$ in the direction v, if $\{\psi_o + \tau v : \tau \in [0,1]\} \subset \Psi$ and

$$\begin{aligned}\Gamma(h, \psi_o)[v] &:= \left.\frac{\partial E\left[h\left(\nabla\left(\psi_o + \tau v\right), \beta_{o\mathcal{S}}\right)\left(X_1\right)\right) I\left\{X_1 \in \mathcal{S}\right\}\right]}{\partial \tau}\right|_{\tau=0} \\ &= E\left[\{h_1(\nabla \psi_o(X_1), \beta_{o\mathcal{S}})\}^T \nabla v(X_1) I\left\{X_1 \in \mathcal{S}\right\}\right],\end{aligned}$$

where $h_1(z, \beta_{o\mathcal{S}}) = \frac{\partial h(z, \beta_{o\mathcal{S}})}{\partial z}$.

The linear functional $\Gamma(h, \psi_o)[\cdot]$ is bounded if and only if

$$\begin{aligned}\sup_{v \in \mathcal{V}, v \neq 0} \frac{\left|\Gamma(h, \psi_o)[v]\right|^2}{\|v\|^2} &= \sup_{v \in \mathcal{V}, v \neq 0} \frac{\left|E\left[\{h_1(T_o(X_1), \beta_{o\mathcal{S}})\}^T \nabla v(X_1) I\left\{X_1 \in \mathcal{S}\right\}\right]\right|^2}{E\left[\{\nabla v(X_1)\}^T \left[\nabla T_o(X_1)\right]^{-1} \nabla v(X_1)\right]} \\ &= E\left[\{h_1(T_o(X_1), \beta_{o\mathcal{S}})\}^T \nabla T_o(X_1) \{h_1(T_o(X_1), \beta_{o\mathcal{S}})\} I\left\{X_1 \in \mathcal{S}\right\}\right] < \infty.\end{aligned}$$

Condition 4. $E\left[\{h_1(T_o(X_1), \beta_{o\mathcal{S}})\}' \nabla T_o(X_1) \{h_1(T_o(X_1), \beta_{o\mathcal{S}})\} I\left\{X_1 \in \mathcal{S}\right\}\right] < \infty$.

By Riesz representation theorem, the linear functional $\Gamma(h, \psi_o)[\cdot]$ is bounded if and only if (iff) there is a Riesz representer $v^* \in \mathcal{V}$ such that

$$\Gamma(h, \psi_o)[v] = \langle v^*, v \rangle \text{ for all } v \in \mathcal{V} \tag{1.19}$$

and

$$\begin{aligned}\left\|v^*\right\|^2 &= \sup_{v \in \mathcal{V}, v \neq 0} \frac{\left|\Gamma(h, \psi_o)[v]\right|^2}{\|v\|^2} \\ &= E\left[\{h_1(T_o(X_1), \beta_{o\mathcal{S}})\}' \nabla T_o(X_1) \{h_1(T_o(X_1), \beta_{o\mathcal{S}})\} I\left\{X_1 \in \mathcal{S}\right\}\right] < \infty,\end{aligned}$$

where

$$\nabla v^*(X_1) = \nabla T_o(X_1)\{h_1(T_o(X_1), \beta_{oS})\} I\{X_1 \in \mathcal{S}\}. \tag{1.20}$$

Like ψ_0, $v^*(\cdot)$ is unique up to a constant $v^*(\cdot) + const$ for $x \in \mathcal{S}$. In the univariate case, we have $T_o(X_1) = F_{X_0}^{-1}(F_{X_1}(X_1))$ and

$$v^*(x) = q(T_o(x), \beta_{oS}) = -E_{X_1}\left[\frac{h_1(T_o(X_1), \beta_{oS})}{f_{X_0}(T_o(X_1))} \times [I\{T_o(x) \leq T_o(X_1)\} - F_{X_1}(X_1)] \times I\{X_1 \in \mathcal{S}\}\right].$$

Then we have

$$\begin{aligned}
\frac{d}{dx}v^*(x) &= -\frac{d}{dx}E_{X_1}\left[\frac{h_1(T_o(X_1), \beta_{oS})}{f_{X_0}(T_o(X_1))} \times I\{x \leq X_1\} \times I\{X_1 \in \mathcal{S}\}\right] \\
&= -\frac{d}{dx}\int_x^\infty \frac{h_1(T_o(t), \beta_{oS})}{f_{X_0}(T_o(t))} I\{t \in \mathcal{S}\} f_{X_1}(t)dt \\
&= \frac{f_{X_1}(x)}{f_{X_0}(T_o(x))} h_1(T_o(x), \beta_{oS}) I\{x \in \mathcal{S}\} \\
&= [\frac{d}{dx}T_o(x)]\{h_1(T_o(x), \beta_{oS})\} I\{x \in \mathcal{S}\},
\end{aligned}$$

which is consistent with the expression $\nabla v^*(x)$ in equation (1.20).

However, when $d_X > 1$, v^* may not have a closed form expression because it may depend on the choice of subset $\mathcal{S}$. We provide a sieve approximation of the Riesz representer v^* for which always has an explicit expression following Chen and Liao (2015). Recall $(\mathcal{V}_n, \|\cdot\|)$ as the sieve space, for each $k_{n_1} < \infty$, the restricted linear functional $\Gamma(h, \psi_o)[\cdot] : \mathcal{V}_n \to \mathbb{R}$ is always bounded and hence there always exists a sieve Riesz representer $\pi_n v^* \in \mathcal{V}_n$ such that

$$\begin{aligned}
&\Gamma(h, \psi_o)[v] = \langle v, \pi_n v^* \rangle \quad \text{for all } v \in \mathcal{V}_n \text{ and} \\
&\|\pi_n v^*\| = \sup_{v \in \mathcal{V}_n, v \neq 0} \frac{|\Gamma(h, \psi_o)[v]|}{\|v\|} < \infty.
\end{aligned}$$

By definition, the sieve Riesz representer $\pi_n v^*(\cdot) = B_{k_{n_1}}(\cdot)'\gamma^*_{k_{n_1}} \in \mathcal{V}_n$ solves the following

optimization problem:

$$\|\pi_n v^*\|^2 = \sup_{\gamma \in \mathbb{R}^{k_{n_1}}, \gamma \neq 0} \frac{\gamma' F_{k_{n_1}} F'_{k_{n_1}} \gamma}{\gamma' R_{k_{n_1}} \gamma},$$

where $F_{k_{n_1}} = \Gamma(h, \psi_o)\left[B_{k_{n_1}}(\cdot)\right] \equiv \left(\Gamma(h, \psi_o)\left[b_1(\cdot)\right], \ldots, \Gamma(h, \psi_o)\left[b_{k_{n_1}}(\cdot)\right]\right)'$ is a $k_{n_1} \times 1$ vector, and $R_{k_{n_1}}$ is a $k_{n_1} \times k_{n_1}$ positive definite matrix such that

$$\gamma' R_{k_{n_1}} \gamma \equiv -\left[\frac{\partial^2}{\partial \tau^2} Q\left(\psi_o(\cdot) + \tau B_{k_{n_1}}(\cdot)'\gamma\right)\right]\Big|_{\tau=0}$$

for all $\gamma \in \mathbb{R}^{k_{n_1}}$. Then $\pi_n v^* \in \mathcal{V}_n$ always has a closed form expression as

$$\pi_n v^*(\cdot) = B_{k_{n_1}}(\cdot)'\gamma^*_{k_{n_1}} = B_{k_{n_1}}(\cdot)'\left(R_{k_{n_1}}\right)^- F_{k_{n_1}} \in \mathcal{V}_n,$$

where $\left(R_{k_{n_1}}\right)^-$is a generalized inverse of $R_{k_{n_1}}$.

Step 3. Verify the assumptions

Proposition 5. Assume Condition 2, Assumption 3.1, 4.1 and 11 hold, then Assumption 5 in section 1.4.2 holds. Further Assume 3.2, 4 and 9 hold, then Assumption 6 in section 1.4.3 holds.

Condition 5. Let the following hold:

1. There is $\omega > 0$ such that

$$\left| \begin{array}{c} E\left[h\left(\nabla\psi\left(X_1\right), \beta_{o\mathcal{S}}\right) I\left\{X_1 \in \mathcal{S}\right\}\right] - E\left[h\left(\nabla\psi_o\left(X_1\right), \beta_{o\mathcal{S}}\right) I\left\{X_1 \in \mathcal{S}\right\}\right] \\ -\Gamma(h, \psi_o)[\psi - \psi_o] \end{array} \right| = O\left(\|\psi - \psi_o\|^{\omega}\right)$$

   uniformly in $\psi \in \Psi_n$ with $\|\psi - \psi_o\| = o(1)$;

2. $\|\pi_n v^* - v^*\| \times \left\|\widehat{\psi} - \psi_o\right\| = o_p\left(n^{-1/2}\right)$.

Proposition 6. Suppose Condition 2, 4, 5 hold, further assume Assumptions 3.1, 4.1 and 5.1 hold. Then Assumption 7 holds with

$$\varphi_1(X_1) = -[v^*(X_1) - E(v^*(X_1))]; \quad \varphi_0(X_0) = v^*\left(T_o^{-1}(X_0)\right) - E\left[v^*\left(T_o^{-1}(X_0)\right)\right],$$

where v^* is the Riesz representer defined in (1.19).

1.6 Consistent Variance Estimators for $\widehat{\tau}$

Here we make a smooth version of $m(\cdot,\cdot)$ in QTT case. To be specific, let $m(Y_1;\kappa_{o\mathcal{S}}) = 1 - \frac{1}{1+e^{-c(Y_1-\kappa_{o\mathcal{S}})}} - q$ instead of $m(Y_1;\kappa_{o\mathcal{S}}) = (I\{Y_1 \le \kappa_{o\mathcal{S}}\}) - q$ in QTT case. The choice of c is specified in simulation part.

Denote $\widehat{M}$ as a consistent estimator for M. Recall that $M = \frac{\partial}{\partial\kappa}E[m(Y_1;\kappa_{o\mathcal{S}})\, I\{X_1 \in \mathcal{S}\}]$. When parameter of interest is ATT, $m(Y_1;\kappa_{o\mathcal{S}}) = Y_1 - \kappa_{o\mathcal{S}}$. When parameter of interest is QTT, $m(Y_1;\kappa_{o\mathcal{S}}) = 1 - \frac{1}{1+e^{-c(Y_1-\kappa_{o\mathcal{S}})}} - q$, for a big $c > 0$. We can choose a consistent estimator for M as below: $\widehat{M} = -\frac{1}{n_1}\sum_{i=1}^{n_1} I\{X_{1i} \in \mathcal{S}\}$ in ATT case, and $\widehat{M} = \frac{1}{n_1}\sum_{i=1}^{n_1}\left\{\frac{c\cdot e^{-c(Y_{1i}-\widetilde{\kappa})}}{\left(1+e^{-c(Y_{1i}-\widetilde{\kappa})}\right)^2} I\{X_{1i} \in \mathcal{S}\}\right\}$ in QTT case where $\widetilde{\kappa}$ is an consistent estimator for $\kappa_{o\mathcal{S}}$. We choose $\widehat{G} = \frac{\partial}{\partial\beta}\left(\frac{1}{n_1}\sum_{i=1}^{n_1}\widehat{h}\left(\widehat{T}(X_{1i}),\widehat{\beta}\right) I\{X_{1i} \in \mathcal{S}\}\right)$ as a consistent estimator for G. Then, we estimate the asymptotic variance by the following estimator:

$$\widehat{V} = \widehat{G}^{-2}\left[\widehat{\Omega}_{\tau,1} + \widehat{\lambda}\widehat{\Omega}_0\right] \tag{1.21}$$

where

$$\widehat{\Omega}_{\tau,1} = \frac{1}{n_1}\sum_{i=1}^{n_1}\left[-\widehat{G}\widehat{M}^{-1}m(Y_{1i};\widehat{\kappa})\, I\{X_{1i} \in \mathcal{S}\} + \widehat{h}\left(\widehat{T}(X_{1i}),\widehat{\beta}\right) I\{X_{1i} \in \mathcal{S}\} + \widehat{\varphi}_1(X_{1i}) - \overline{m}_1\right]^2,$$

$$\text{for } \overline{m}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1}\left[-\widehat{G}\widehat{M}^{-1}m(Y_{1i};\widehat{\kappa})\, I\{X_{1i} \in \mathcal{S}\} + \widehat{h}\left(\widehat{T}(X_{1i}),\widehat{\beta}\right) I\{X_{1i} \in \mathcal{S}\} + \widehat{\varphi}_1(X_{1i})\right];$$

and

$$\widehat{\Omega}_0 = \frac{1}{n_0}\sum_{j=1}^{n_0}\left[\widehat{\varphi}_0\left(X_{0j}\right) + \left(m\left(Y_{0j};\widehat{\beta}\right) - \widehat{h}\left(X_{0j},\widehat{\beta}\right)\right) I\left\{\widehat{T}^{-1}\left(X_{0j}\right)\in\mathcal{S}\right\} - \overline{m}_0\right]^2,$$

$$\text{for } \overline{m}_0 = \frac{1}{n_0}\sum_{j=1}^{n_0}\left[\widehat{\varphi}_0\left(X_{0j}\right) + \left(m\left(Y_{0j};\widehat{\beta}\right) - \widehat{h}\left(X_{0j},\widehat{\beta}\right)\right) I\left\{\widehat{T}^{-1}\left(X_{0j}\right)\in\mathcal{S}\right\}\right].$$

Theorem 3. Let all conditions in Corollary 2 hold. Then:

$$\widehat{V} = V_\tau + o_p(1).$$

Consider a null hypothesis of the form $H_0 : \tau_{o\mathcal{S}} = 0$. One can consider different test statistics for this null hypothesis. Here we choose the Wald statistic:

$$\mathcal{W}_n = n_1\widehat{\tau}^2/\widehat{V} \xrightarrow{d} \chi^2(1) \text{ under } H_0,$$

where $\widehat{V}$ is given in (1.21).

1.6.1 Estimators of φ_1 and φ_0 in Examples 1-3 With Different $\widehat{T}$

Example 1: Univariate Nonparametric Case (Cont'd)

In univariate case, let

$$\widehat{q}(x,\widehat{\beta}) = -\frac{1}{n_1}\sum_{i=1}^{n_1}\frac{\widehat{h}_1\left(\widehat{T}\left(X_{1i}\right),\widehat{\beta}\right)}{\widehat{f}_{X_0}\left(\widehat{T}\left(X_{1i}\right)\right)} \times \left[I\left\{x\le\widehat{T}\left(X_{1i}\right)\right\} - \widehat{F}_{X_1}(X_{1i})\right] \times I\left\{X_{1i}\in\mathcal{S}\right\}.$$

Consistent estimators for φ_1 and φ_0 can be chosen as:

$$\widehat{\varphi}_1(x_1) = -q\left(\widehat{T}\left(x_1\right),\widehat{\beta}\right) \text{ and } \widehat{\varphi}_0(x_0) = q\left(x_0,\widehat{\beta}\right).$$

When the subset $\mathcal{S} = [\underline{\mathcal{S}},\overline{\mathcal{S}}]$, $\widehat{q}\left(x,\widehat{\beta}\right)$ is given by

$$\widehat{q}\left(x,\widehat{\beta}\right) = \widehat{h}\left(x,\widehat{\beta}\right) \times I\left\{\underline{\mathcal{S}}\le\widehat{T}^{-1}(x)\le\overline{\mathcal{S}}\right\} - \widehat{h}\left(\widehat{T}(\overline{\mathcal{S}}),\widehat{\beta}\right) \times I\left\{\underline{\mathcal{S}}\le\widehat{T}^{-1}(x)\le\overline{\mathcal{S}}\right\}$$

$$
\begin{aligned}
&- \widehat{h}\left(\widehat{T}(\overline{\mathcal{S}}), \widehat{\beta}\right) \times I\left\{\widehat{T}^{-1}(x) \leq \underline{\mathcal{S}}\right\} + \widehat{h}\left(\widehat{T}(\underline{\mathcal{S}}), \widehat{\beta}\right) \times I\left\{\widehat{T}^{-1}(x) \leq \underline{\mathcal{S}}\right\} \\
&+ \widehat{h}\left(\widehat{T}(\overline{\mathcal{S}}), \widehat{\beta}\right) \times \widehat{F}_{X_1}(\overline{\mathcal{S}}) - \widehat{h}\left(\widehat{T}(\underline{\mathcal{S}}), \widehat{\beta}\right) \times \widehat{F}_{X_1}(\underline{\mathcal{S}}),
\end{aligned}
$$

where $\widehat{T}^{-1}(x) = \widehat{F}_{X_1}^{-1}\left(\widehat{F}_{X_0}(x)\right)$.

Example 2: Multivariate Affine Map (Cont'd)

By Proposition 4 we can consider simple plug-in estimators of $\varphi_1(x_1)$ and $\varphi_0(x_0)$ as follows:

$$
\begin{aligned}
\widehat{\varphi}_1(x_1) = &-\frac{1}{n_1}\sum_{i=1}^{n_1}\left[I\left\{X_{1i} \in \mathcal{S}\right\}\widehat{h}_1(\widehat{T}(X_{1i}), \widehat{\beta})'\widehat{A}(x_1 - \widehat{\mu}_1)\right] \\
&- \frac{1}{n_1}\sum_{i=1}^{n_1}\left[\begin{array}{c}\widehat{h}_1(\widehat{T}(X_{1i}), \widehat{\beta})'\left(\widehat{\Sigma}_0\widehat{U}\left[\widehat{K} \circ \left(\widehat{U}^{-1}\widehat{\Sigma}_0^{-1}\widehat{\Sigma}_1^{-1}\left((x_1 - \widehat{\mu}_1)(x_1 - \widehat{\mu}_1)' - \widehat{\Sigma}_1\right)\widehat{\Sigma}_1^{-1}\widehat{U}\right)\right]\widehat{U}^{-1}\right) \\ \times (X_{1i} - \widehat{\mu}_1) I\left\{X_{1i} \in \mathcal{S}\right\}\end{array}\right];
\end{aligned}
$$

$$
\begin{aligned}
\widehat{\varphi}_0(x_0) = &\frac{1}{n_1}\sum_{i=1}^{n_1}\left[I\left\{X_{1i} \in \mathcal{S}\right\}\widehat{h}_1(\widehat{T}(X_{1i}), \widehat{\beta})'(x_0 - \widehat{\mu}_0)\right] \\
&+ \frac{1}{n_1}\sum_{i=1}^{n_1}\left[\begin{array}{c}\widehat{h}_1(\widehat{T}(X_{1i}), \widehat{\beta})'\left(\widehat{\Sigma}_1^{-1}\widehat{U}\left[\widehat{L} \circ \left(\widehat{U}^{-1}\widehat{\Sigma}_1\left((x_0 - \widehat{\mu}_0)(x_0 - \widehat{\mu}_0)' - \widehat{\Sigma}_0\right)\widehat{U}\right)\right]\widehat{U}^{-1}\right) \\ \times (X_{1i} - \widehat{\mu}_1) I\left\{X_{1i} \in \mathcal{S}\right\}\end{array}\right].
\end{aligned}
$$

where $\widehat{A} = \widehat{\Sigma}_1^{-1}\left(\widehat{\Sigma}_1\widehat{\Sigma}_0\right)^{1/2}$. Denote the eigendecomposition of $\widehat{\Sigma}_1\widehat{\Sigma}_0 = \widehat{U}\widehat{D}\widehat{U}^{-1}$ and we can construct $\widehat{L}$ and $\widehat{K}$ as:

$$
[\widehat{L}]_{d,s} = \frac{1}{\sqrt{\widehat{\lambda}_d} + \sqrt{\widehat{\lambda}_s}} \text{ and } [\widehat{K}]_{d,s} = \frac{\sqrt{\widehat{\lambda}_d\widehat{\lambda}_s}}{\sqrt{\widehat{\lambda}_d} + \sqrt{\widehat{\lambda}_s}}
$$

for $\widehat{\lambda}_d$ as the d-th eigenvalue of diagonal matrix $\widehat{D}$.

Example 3: Multivariate Nonparametric Case (Cont'd)

The estimate $\widehat{\pi_n v^*}$ of $\pi_n v^*$ is the Riesz representer of the estimated functional $\widehat{\Gamma}(\widehat{h},\widehat{\psi})[\cdot]$ on $\mathcal{V}_n$, i.e. $\widehat{\pi_n v^*}$ satisfies

$$\widehat{\Gamma}(\widehat{h},\widehat{\psi})[v] = \frac{1}{n_1}\sum_{i=1}^{n_1}\widehat{h}_1\left(\nabla\widehat{\psi}\left(X_{1i}\right),\widehat{\beta}\right)^T\nabla v\left(X_{1i}\right)I\left\{X_{1i}\in\mathcal{S}\right\} = \left\langle\widehat{\pi_n v^*},v\right\rangle_n \quad \text{for all } v\in\mathcal{V}_n \text{ and}$$

$$\left\|\widehat{\pi_n v^*}\right\|_n = \sup_{v\in\mathcal{V}_n, v\neq 0}\frac{\left|\Gamma_n(\widehat{h},\widehat{\psi})[v]\right|}{\|v\|_n} < \infty,$$

where $\|\cdot\|_n$ is the empirical semi-norm associated with the theoretical semi-norm $\|\cdot\|$ defined as

$$\|v\|_n = \left\{\left[\frac{\partial^2}{\partial\tau^2}\widehat{Q}_n\left(\widehat{\psi}+\tau v\right)\right]\Big|_{\tau=0}\right\}^{1/2} \text{ for any } v\in\mathcal{V},$$

and $\langle\cdot,\cdot\rangle_n$ is the empirical inner product induced by the empirical semi-norm $\|\cdot\|_n$. Again if ψ_o is a real valued function and $\mathcal{V}_n = \left\{v(\cdot) = B_{k_{n_1}}(\cdot)'\gamma : \gamma\in\mathbb{R}^{k_{n_1}}\right\}$, then $\widehat{\pi_n v^*}$ can be computed in a closed form:

$$\widehat{\pi_n v^*} = B_{k_{n_1}}(\cdot)'\widehat{\gamma}^*_{k_{n_1}} = B_{k_{n_1}}(\cdot)'\left(\widehat{R}_{k_{n_1}}\right)^-\widehat{F}_{k_{n_1}},$$

where $\widehat{F}_{k_{n_1}} = \widehat{\Gamma}(\widehat{h},\widehat{\psi})\left[B_{k_{n_1}}(\cdot)\right]$ and $\widehat{R}_{k_{n_1}}$ is such that

$$\begin{aligned}
\gamma'\widehat{R}_{k_{n_1}}\gamma &\equiv \left[\frac{\partial^2}{\partial\tau^2}\widehat{Q}_n\left(\widehat{\psi}+\tau B_{k_{n_1}}(\cdot)'\gamma\right)\right]\Big|_{\tau=0}\\
&= -\left[\frac{\partial}{\partial\tau}\int B_{k_{n_1}}\left(x^*(\tau)\right)'\gamma\right]d\widehat{F}_{X_0}(y)\Big|_{\tau=0}\\
&= -\int\left[\left\{\nabla B_{k_{n_1}}\left(x^*(\tau)\right)'\gamma\right\}^T\frac{\partial x^*(\tau)}{\partial\tau}\Big|_{\tau=0}\right]d\widehat{F}_{X_0}(y)\\
&= \int\gamma'\left(\nabla B_{k_{n_1}}\left(x^*(0)\right)\left[\nabla^2\widehat{\psi}\left(x^*(0)\right)\right]^{-1}\left\{\nabla B_{k_{n_1}}\left(x^*(0)\right)\right\}^T\right)\gamma\, d\widehat{F}_{X_0}(y)\\
&= \gamma'\frac{1}{n_0}\sum_{i=1}^{n_0}\left\{\nabla B_{k_{n_1}}\left(x^*(0)_i\right)\left[\nabla^2\widehat{\psi}\left(x^*(0)_i\right)\right]^{-1}\left\{\nabla B_{k_{n_1}}\left(x^*(0)_i\right)\right\}^T\right\}\gamma
\end{aligned}$$

for all $\gamma \in \mathbb{R}^{k_{n_1}}$, where $X_{0i} = \nabla\widehat{\psi}\left(x^*(0)_i\right)$ for $i = 1, \ldots, n_0$. Thus, we propose to estimate φ_1 and φ_0 by

$$\begin{aligned}
-\widehat{\pi_n v^*}\left(x\right) =& - B_{k_{n_1}}(x)' \left\{ \frac{1}{n_0} \sum_{i=1}^{n_0} \nabla B_{k_{n_1}}\left(x^*(0)_i\right) \left[\nabla^2 \widehat{\psi}\left(x^*(0)_i\right)\right]^{-1} \left\{\nabla B_{k_{n_1}}\left(x^*(0)_i\right)\right\}^T \right\}^- \\
& \times \left\{ \frac{1}{n_1} \sum_{i=1}^{n_1} \nabla B_{k_{n_1}}\left(X_{1i}\right) \left[\widehat{h}_1\left(\nabla\widehat{\psi}\left(X_{1i}\right), \widehat{\beta}\right)\right] I\left\{X_{1i} \in \mathcal{S}\right\} \right\} + const.; \\
\widehat{\varphi}_1(x_1) =& - \widehat{\pi_n v^*}\left(x_1\right) + \frac{1}{n_1} \sum_{i=1}^{n_1} \widehat{\pi_n v^*}\left(X_{1i}\right),
\end{aligned}$$

and

$$\begin{aligned}
\widehat{\pi_n v^*}\left(\widehat{T}^{-1}\left(x\right)\right) =& B_{k_{n_1}}\left(\widehat{T}^{-1}(x)\right)' \left\{ \frac{1}{n_0} \sum_{i=1}^{n_0} \nabla B_{k_{n_1}}\left(x^*(0)_i\right) \left[\nabla^2 \widehat{\psi}\left(x^*(0)_i\right)\right]^{-1} \left\{\nabla B_{k_{n_1}}\left(x^*(0)_i\right)\right\}^T \right\}^- \\
& \times \left\{ \frac{1}{n_1} \sum_{i=1}^{n_1} \nabla B_{k_{n_1}}\left(X_{1i}\right) \left[\widehat{h}_1\left(\nabla\widehat{\psi}\left(X_{1i}\right), \widehat{\beta}\right)\right] I\left\{X_{1i} \in \mathcal{S}\right\} \right\} + const.; \\
\widehat{\varphi}_0(x_0) =& \widehat{\pi_n v^*}\left(\widehat{T}^{-1}\left(x_0\right)\right) - \frac{1}{n_0} \sum_{j=1}^{n_0} \widehat{\pi_n v^*}\left(\widehat{T}^{-1}\left(X_{0j}\right)\right)
\end{aligned}$$

1.7 Numerical Results

To examine the accuracy of the inference procedure, we carry out two simulation studies to estimate the average treatment effect on the treated group when in univariate and bivariate case, respectively.

1.7.1 Performance of the Estimator

Univariate Case

When covariate has a dimension of one, the following method of data generation is used: observations of covariate in treatment group are i.i.d drawn from a normal distribution, and observations of covariate in control group are i.i.d. drawn from a uniform distribution: $X_1 \sim \mathcal{N}(5, 1)$ and $X_0 \sim U[4, 6]$. Conditioning on $X, D = 1$, observations of outcome of

treated in the treatment group Y_1 are i.i.d. drawn from a normal distribution: $Y_1 \mid x, D = 1 \sim \mathcal{N}(x, 1)$. Conditioning on $x, D = 0$, observations of outcome of not treated in the control group Y_0 are i.i.d. drawn from a normal distribution: $Y_0 \mid x, D = 0 \sim \mathcal{N}\left(T_o^{-1}(x), 1\right)$, and satisfies our Assumption 1 (ii). Here $T_o^{-1} : \mathcal{X}_0 \to \mathcal{X}_1$ pushes the distribution of X_0 to that of X_1. In this case, we choose the subset $\mathcal{S}$ to be all the $X_1 \leq 5$. The true function $h(x, \beta)$ is given by $h(x, \beta) = T_o^{-1}(x) - \beta$, and the true optimal transport map $T_o(x)$ is given by $T_o(x) = \frac{2}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-5)^2} + 4$. The true average treatment effect for the treated $\tau_o^{ATT} = 0$ and true quantile treatment effect for the treated $\tau_o^{QTT} = 0$. We consider 5,000 Monto Carlo (MC) repetitions, with sample sizes of $(n_1, n_0) = (500, 1000)$ for the treatment group and control group, respectively.

To estimate τ_o, we follow these steps:

1. We estimate T_o by plugging in the empirical CDF: $\widehat{T}(x) = \widehat{F}_{X_0}^{-1}\left(\widehat{F}_{X_1}(x)\right)$.

2. We estimate $h(x, \beta) = E\left[m(Y_0, \beta) \mid X_0 = x\right]$ using a sieve estimator using quadratic spline basis with $k_{n_0} = 4$.

3. We estimate $\beta_{o\mathcal{S}}$ and $\kappa_{o\mathcal{S}}$ respectively using method of moments. This gives us $\widehat{\beta}$ by letting $\frac{1}{n_1}\sum_{i=1}^{n_1} \widehat{h}\left(\widehat{T}(X_{1i}), \widehat{\beta}\right) I\{X_{1i} \in \mathcal{S}\} = 0$, and $\widehat{\kappa}$ by letting $\frac{1}{n_1}\sum_{i=1}^{n_1} m(Y_{1i}, \widehat{\kappa}) I\{X_{1i} \in \mathcal{S}\} = 0$. The difference between $\widehat{\kappa}$ and $\widehat{\beta}$ is then calculated as $\widehat{\tau}$. In the simulation, we let $c = 10$. Thus in QTT case, $m(Y_1; \widehat{\kappa}) = 1 - \frac{1}{1+e^{-10(Y_1 - \widehat{\kappa})}} - q$.

4. To construct inference for the estimator, we follow Section 1.6 to construct consistent variance estimator for $\widehat{\tau}$.

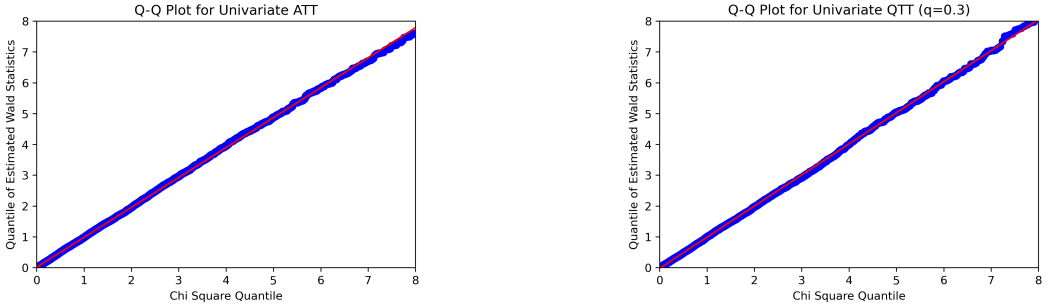
Table 1.1 shows the performance for different parameters of interest, i.e., ATT and QTT. Table 1.2 shows the performance for Wald test, and for different parameters of interest, Wald test generally performs well. Figure 1.2 shows the QQ-Plot for different parameters of interest. Based on the results, our estimator performs equally well for different parameters of interest, and so does the Wald test.

Table 1.1: Univariate Case Performance

Parameter of interest	bias	variance	MSE
ATT	-0.009	0.010	0.010
QTT, q=0.3	-0.008	0.013	0.014
QTT, q=0.5	-0.008	0.014	0.014
QTT, q=0.7	-0.007	0.014	0.014

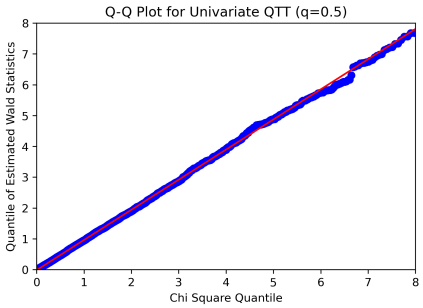
Table 1.2: Univariate Case Inference: Wald Test of $\tau = 0$

Parameter of interest	10%	5%	1%
ATT	9.9%	4.6%	0.8%
QTT, q=0.3	9.8%	5.0%	1.0%
QTT, q=0.5	9.3%	4.7%	0.9%
QTT, q=0.7	9.3%	4.7%	1.0%

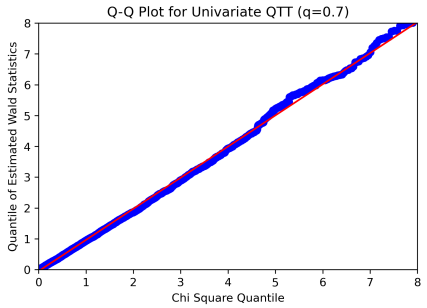


(a)

(b)



(c)



(d)

Figure 1.2: Q-Q plot for Wald tests $\mathcal{W}_n$ in univariate case

Multivariate Case

Bivariate Case. The data generating process for bivariate case can be described as follows: $\{X_{1i}\}_{i=1}^{n_1}$ are drawn from multivariate uniform distribution on $[0,1]^2$ and each dimension is drawn independently, and $\{X_{0j}\}_{j=1}^{n_0}$ are drawn from multivariate uniform distribution on $[0,2]^2$ and each dimension is drawn independently. The outcome variable of being treated in the treatment group $\{Y_{1i}\}_{i=1}^{n_1}$ are drawn from normal distribution conditioning on X_{1i}. That is, $Y_1 \mid X_1 = (x_1, x_2)' \sim N(x_1 + 5x_2, 1)$. $T_o(x)$ is an affine map and admits a closed form, e.g.,

$$T_o(x) = \begin{pmatrix} 1 \\ 1 \end{pmatrix} + \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix} \begin{pmatrix} x_1 - \frac{1}{2} \\ x_2 - \frac{1}{2} \end{pmatrix}, \quad T_o^{-1}(x) = \begin{pmatrix} \frac{1}{2} \\ \frac{1}{2} \end{pmatrix} + \begin{pmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{2} \end{pmatrix} \begin{pmatrix} x_1 - 1 \\ x_2 - 1 \end{pmatrix}.$$

We choose a subset of interest to be $\mathcal{S} = \{x \in \mathcal{X}_1 | x_1 \leq 0.7, x_2 \leq 0.7\}$ in bivariate case. We consider 5,000 Monto Carlo (MC) repetitions, with sample sizes of $(n_1, n_0) = (1000, 2000)$ for the treatment group and control group, respectively.

Multivariate Case with dimension of covariates = 5. The data generating process for bivariate case can be described as follows: $\{X_{1i}\}_{i=1}^{n_1}$ are drawn from multivariate uniform distribution on $[0,1]^5$ and each dimension is independent, and $\{X_{0j}\}_{j=1}^{n_0}$ are drawn from multivariate uniform distribution on $[0,2]^5$ and each dimension is independent. The outcome variable of treated in the treatment group Y_{1i} are drawn from normal distribution conditioning on X_{1i}. That is, $Y_1 \mid X_1 = (x_1, x_2, x_3, x_4, x_5)' \sim N(x_1 + 2x_2 + 3x_3 + 4x_4 + 0x_5, 1)$. $T_o(x)$ is an affine map and admits a closed form, e.g.,

$$T_o(x) = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} + \begin{pmatrix} 2 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & 2 \end{pmatrix} \begin{pmatrix} x_1 - \frac{1}{2} \\ x_2 - \frac{1}{2} \\ x_3 - \frac{1}{2} \\ x_4 - \frac{1}{2} \\ x_5 - \frac{1}{2} \end{pmatrix},$$

$$T_o^{-1}(x) = \begin{pmatrix} 1/2 \\ 1/2 \\ 1/2 \\ 1/2 \\ 1/2 \end{pmatrix} + \begin{pmatrix} 1/2 & 0 & 0 & 0 & 0 \\ 0 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 1/2 & 0 & 0 \\ 0 & 0 & 0 & 1/2 & 0 \\ 0 & 0 & 0 & 0 & 1/2 \end{pmatrix} \begin{pmatrix} x_1 - 1 \\ x_2 - 1 \\ x_3 - 1 \\ x_4 - 1 \\ x_5 - 1 \end{pmatrix}.$$

We choose a subset of interest to be $\mathcal{S} = \{x \in \mathcal{X}_1 | x_d \leq 0.9, \text{ for } d = 1, \ldots, d_X\}$ in multivariate case. We consider 2,000 Monto Carlo (MC) repetitions, with sample sizes of $(n_1, n_0) = (1000, 2000)$ for the treatment group and control group, respectively.

We let $Y_{0i} = Y_{1i}$, so both the individual treatment effect and the average treatment effect $\tau_{o\mathcal{S}}$ is zero. We follow a similar procedure to construct $\widehat{h}(x, \beta)$ in the multivariate case as in the univariate case, with the basis functions are tensor product of the spline basis for each dimension. To estimate $T_o(x) = \nabla\psi_o(x)$, we adopt two estimators: an affine map estimator and a spline estimator. For the affine map estimator, we follow the following procedure:

1. For $j = 0, 1$, calculate $\widehat{\mu}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} X_{ji}$ and $\quad \widehat{\Sigma}_j = \frac{1}{n_j}\sum_{i=1}^{n_j}(X_{ji} - \widehat{\mu}_j)(X_{ji} - \widehat{\mu}_j)'$.

2. Calculate $\widehat{A} = \widehat{\Sigma}_1^{-1/2}\left(\widehat{\Sigma}_1^{1/2}\widehat{\Sigma}_0\widehat{\Sigma}_1^{1/2}\right)^{1/2}\widehat{\Sigma}_1^{-1/2}$, then $\widehat{T}(x) = \widehat{\mu}_0 + \widehat{A}(x - \widehat{\mu}_1)$.

For the spline estimator, we adopt two different approaches to estimate optimal transport map: the FOC approach and the non-FOC approach.

FOC: We estimate ψ_o using quadratic spline basis functions, denoted as $B_{k_{n_1}}(\cdot) = \left(b_1(\cdot), \ldots, b_{k_{n_1}}(\cdot)\right)'$, and $\widehat{\psi}(x) = B_{k_{n_1}}(x)'\widehat{\gamma}$. $\widehat{\gamma}$ solves

$$\widehat{\gamma} = \underset{\substack{\gamma \\ \nabla^2 B_{k_{n_1}}(x)'\gamma \text{ is p.s.d for } x \in \mathcal{X}_1}}{\arg\min} \left\{ \frac{1}{n_1}\sum_{i=1}^{n_1} B_{k_{n_1}}(X_{1i})'\gamma + \frac{1}{n_0}\sum_{j=1}^{n_0}\left(\langle z_j^*, X_{0j}\rangle - B_{k_{n_1}}\left(z_j^*\right)'\gamma\right)\right\},$$

and $\left\{z_j^*\right\}_{j=1}^{n_0}$ is a sequence of vectors that for $j = 1, \ldots, n_0$, each vector z_j^* solves $X_{0j} = \nabla B_{k_{n_1}}\left(z_j^*\right)'\gamma$. In this approach, we solve conjugate using the first-order condition, and thus it's called the FOC approach.

Non-FOC: We estimate ψ_o using quadratic spline basis functions, denoted as $B_{k_{n_1}}(\cdot) = \left(b_1(\cdot), \ldots, b_{k_{n_1}}(\cdot)\right)'$, and $\widehat{\psi}(x) = B_{k_{n_1}}(x)'\widehat{\gamma}$. $\widehat{\gamma}$ solves

$$\widehat{\gamma} = \underset{\substack{\gamma \\ \nabla^2 B_{k_{n_1}}(x)'\gamma \text{ is p.s.d for } x \in \mathcal{X}_1}}{\arg\min} \left\{ \frac{1}{n_1} \sum_{i=1}^{n_1} B_{k_{n_1}}(X_{1i})'\gamma + \frac{1}{n_0} \sum_{j=1}^{n_0} \left(\max_{z, z \in \mathcal{X}_1} \left\{ \langle z, X_{0j} \rangle - B_{k_{n_1}}(z)'\gamma \right\} \right) \right\}.$$

For each $X_{0j}, j = 1, \ldots, n_0$, we used sequential quadratic programming to solve the inner maximization problem, then use another sequential quadratic programming to solve for the outer minimization.

We use both FOC and non-FOC approaches in bivariate case, and only non-FOC in multivariate case. The constraint of convexity of $\widehat{\psi}(x)$ is satisfied by adding a constraint to let the Hessian matrix to be positive semi-definite. The Hessian matrix can be easily computed by taking the second derivative of the quadratic spline basis functions. We then follow the same procedure as the univariate case to construct $\widehat{\beta}$, $\widehat{\kappa}$, and $\widehat{\tau}$.

The results for bivariate case are summarized in Table 1.3 and 1.5, and the results for multivariate case are summarized in Table 1.4 and 1.6. The sieve estimator performs equally well as the affine map estimator. The reported time is measured in seconds and shows the result for running one simulation run using one CPU. Figure 1.3 and Figure 1.4 shows the QQ-Plot for affine map estimator and sieve estimator in bivariate case. Figure 1.5 shows the QQ-Plot for sieve estimator in multivariate case.

Table 1.3: Bivariate Case Performance

Parameter of interest FOC	OT estimation	bias	variance	MSE
ATT	Affine Map	0.001	0.007	0.007
	Sieve	0.003	0.008	0.008
QTT, q=0.3	Affine Map	0.011	0.011	0.012
	Sieve	0.015	0.011	0.011
QTT, q=0.5	Affine Map	0.003	0.011	0.011
	Sieve	0.002	0.012	0.012
QTT, q=0.7	Affine Map	-0.029	0.009	0.010
	Sieve	-0.029	0.013	0.013
non-FOC				
ATT	Affine Map	0.001	0.007	0.007
	Sieve	-0.001	0.008	0.008
QTT, q=0.3	Affine Map	0.015	0.011	0.012
	Sieve	0.010	0.011	0.011
QTT, q=0.5	Affine Map	0.003	0.010	0.010
	Sieve	0.002	0.012	0.012
QTT, q=0.7	Affine Map	-0.027	0.012	0.013
	Sieve	0.032	0.012	0.013

Table 1.4: Multivariate Case Performance (dimension = 5)

Parameter of interest Dimension = 5	OT estimation	bias	variance	MSE
ATT	Sieve	-0.002	0.007	0.007
QTT, q=0.3	Sieve	-0.002	0.012	0.012
QTT, q=0.5	Sieve	-0.017	0.016	0.016

Table 1.5: Bivariate Inference

Parameter of interest	OT Estimation	10%	5%	1%	time (in seconds)
FOC					
ATT	Affine Map	10.15%	4.85%	1.00%	1,581
	Sieve	10.30%	5.30%	1.00%	2,184
QTT, q=0.3	Affine Map	9.30%	4.70%	0.95%	5,768
	Sieve	10.15%	4.78%	0.76%	6,142
QTT, q=0.5	Affine Map	9.35%	5.05%	0.80%	5,733
	Sieve	9.96%	4.85%	0.80%	6,110
QTT, q=0.7	Affine Map	9.37%	4.40%	0.92%	6,109
	Sieve	9.75%	5.20%	1.25%	6,470
non-FOC					
ATT	Affine Map	10.0%	5.3%	0.9%	1,580
	Sieve	10.3%	5.2%	0.9%	2,305
QTT, q=0.3	Affine Map	10.3%	5.3%	1.1%	5,741
	Sieve	10.3%	5.3%	1.2%	2,557
QTT, q=0.5	Affine Map	9.6%	4.6%	0.9%	5,724
	Sieve	10.4%	5.1%	1.0%	2,542
QTT, q=0.7	Affine Map	10.4%	5.4%	1.1%	6,107
	Sieve	10.0%	5.1%	1.3%	2,694

Table 1.6: Multivariate Inference (dimension = 5)

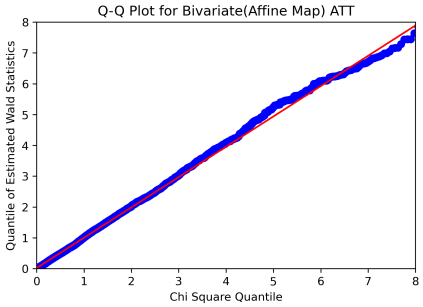
Parameter of interest	OT Estimation	10%	5%	1%	time (in seconds)
ATT	Sieve	10.2%	4.7%	0.6%	56,614
QTT, q=0.3	Sieve	10.3%	5.3%	1.5%	89,433
QTT, q=0.5	Sieve	10.2%	5.1%	1.1%	148,715

Table 1.7: Bivariate Performance When True OT Map Is Not Affine

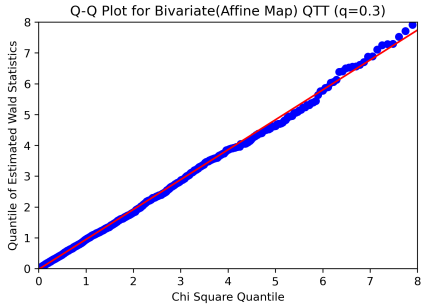
Parameter of interest	OT estimation	bias	variance	MSE
ATT	Affine Map	-0.015	0.006	0.006
	Sieve	0.005	0.006	0.006
QTT, q=0.3	Affine Map	0.103	0.011	0.022
	Sieve	0.005	0.010	0.010

Table 1.8: Bivariate Inference When True OT Map Is Not Affine

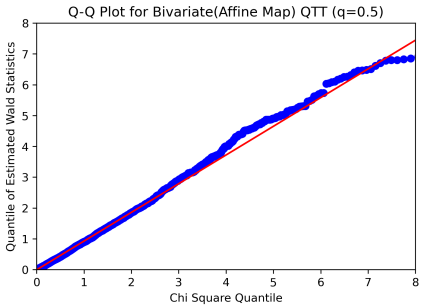
Parameter of interest	OT Estimation	10%	5%	1%
ATT	Affine Map	7.2%	3.3%	0.5%
	Sieve	6.5%	2.7%	0.5%
QTT, q=0.3	Affine Map	21.6%	12.7%	3.8%
	Sieve	8.6%	4.4%	1.1%



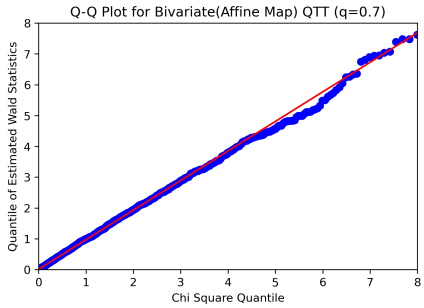
(a) FOC



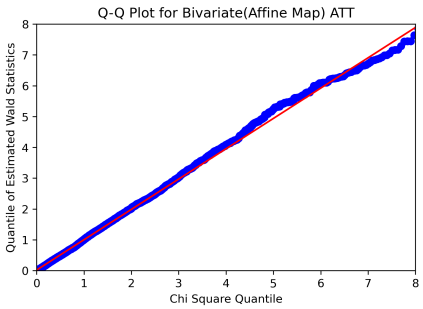
(b) FOC



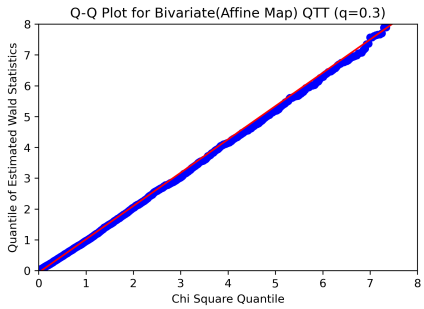
(c) FOC



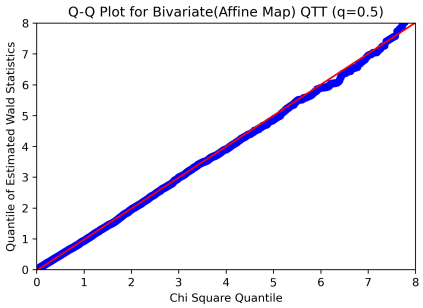
(d) FOC



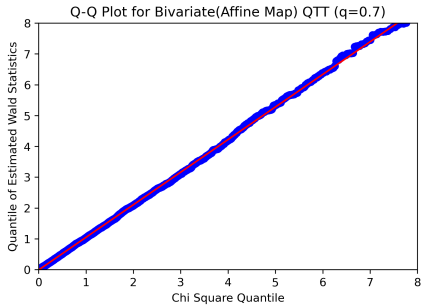
(e) non-FOC



(f) non-FOC

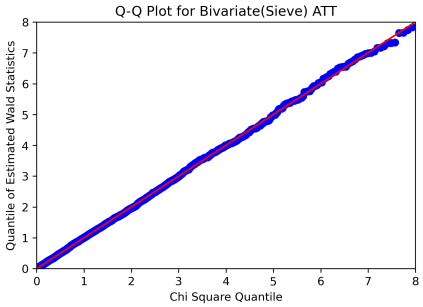


(g) non-FOC

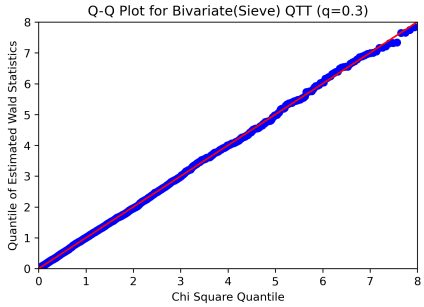


(h) non-FOC

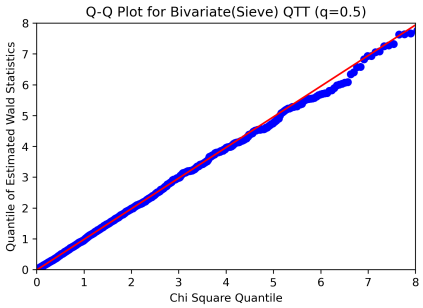
Figure 1.3: Q-Q plot for Wald tests $\mathcal{W}_n$ in bivariate (Affine) case



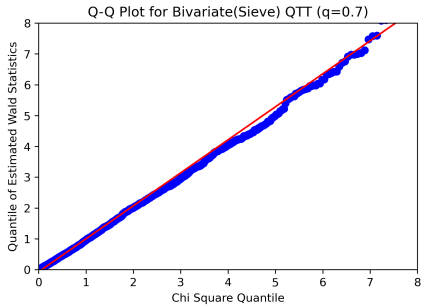
(a) FOC



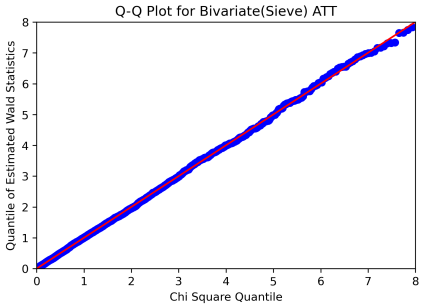
(b) FOC



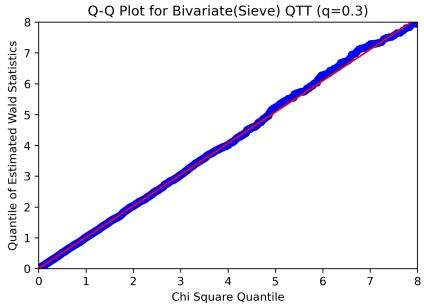
(c) FOC



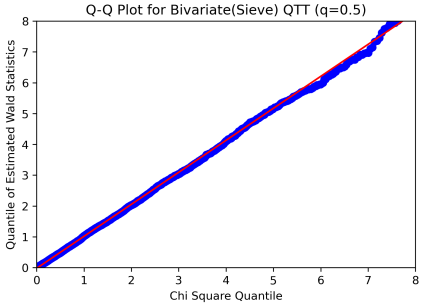
(d) FOC



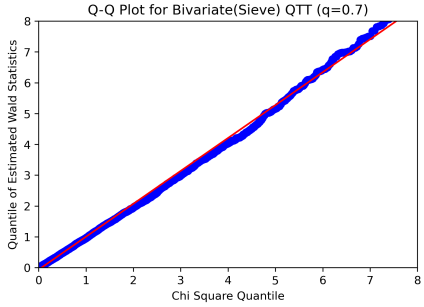
(e) non-FOC



(f) non-FOC

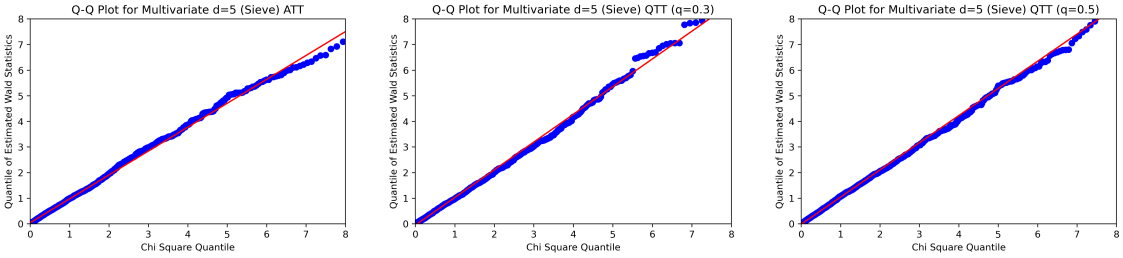


(g) non-FOC



(h) non-FOC

Figure 1.4: Q-Q plot for Wald tests $\mathcal{W}_n$ in bivariate (Sieve) case



(a) non-FOC (b) non-FOC (c) non-FOC

Figure 1.5: Q-Q plot for Wald tests $\mathcal{W}_n$ in Multivariate d=5 (Sieve) case

Chapter 2

CORRECTING STRATEGIC MISREPORTING BEHAVIOR ON OUTCOME IN ESTIMATING TREATMENT EFFECT

2.1 Introduction

Self-reported outcomes are commonly used in research across multiple disciplines to evaluate treatment effects (Blattman et al., 2016). However, evidence indicates that self-reported outcomes are not entirely reliable. Individuals may misreport health status (Angrist et al., 2010), wages (MaCurdy, 1982), and various outcomes in lab studies (Gillen et al., 2019). One crucial reason is that the incentive linked to the reported outcome may induce strategic misreporting behavior (Andreoni et al., 1998; Bénabou and Tirole, 2003; Fischbacher and Föllmi-Heusi, 2013). For example, a high reported income is linked to high tax compliance, and therefore, individuals are incentivized to strategically under-report their income (Andreoni et al., 1998). Workers may be incentivized to over-report their disability status for welfare receipt (Black et al., 2017).

Strategic misreporting behaviors on outcomes pose challenges for causal identification under the potential outcome framework. The presence of strategically misreported outcomes may violate the Strong Ignorability assumption as introduced in Rosenbaum and Rubin (1983b), thereby impeding the identification of the treatment effect (Millimet, 2011). The literature on strategic misreporting mainly focuses on empirically demonstrating the existence of such behavior (Andreoni et al., 1998; Benítez-Silva et al., 2004; Blattman et al., 2016) instead of proposing econometric solutions. Misreporting behavior is also related to the literature on causal inference and measurement error (Lewbel, 2007). However, limited research in this literature focuses on measurement errors in outcomes, possibly because it is often assumed that measurement errors in outcomes, if any, are the same between the treatment group and

the control group, and therefore, identification is still achieved with misreported outcomes (Shu and Yi, 2019). However, this assumption may not hold when the misreporting behavior on outcomes is *strategic*. After receiving treatment assignments, individuals in the treatment group will obtain different true outcomes as compared to individuals in the control group. Because the optimal strategic misreporting behavior is based on the true outcomes, such behavior would naturally differ across the treatment group and the control group, which violates the above assumption and breaks down the identification strategy.

Hence, this work aims to advance the literature by proposing estimators that account for strategic misreporting behaviors on outcomes. More specifically, we concentrate on the case where the treatment group represents a relatively smaller population compared to that of the control group. Thus, our parameter of interest is the treatment effect on the untreated (ATU), as the untreated group represents a larger population. For instance, if the treatment is a job training program, a researcher might be interested in understanding the impact of this program on the broader population beyond the treatment group. However, our framework can be easily extended to other parameters of interest, such as the average treatment effect (ATE) or the average treatment effect on the treated (ATT).

We first introduce a simple economic model to demonstrate an individual's misreporting decision when there are incentives and costs of misreporting in their utility function. Next, we study the identification of ATU in three different scenarios: (i) the baseline scenario, where no incentive is linked to the reported outcome; (ii) Scenario 1, where incentives are linked to the *value* of the reported outcome; and (iii) Scenario 2, incentives are linked to the *rank* of the reported outcomes. For example, a salesperson's wage may depend on their total monthly sales (Scenario 1) or the rank of their total monthly sales within the sales group (Scenario 2).

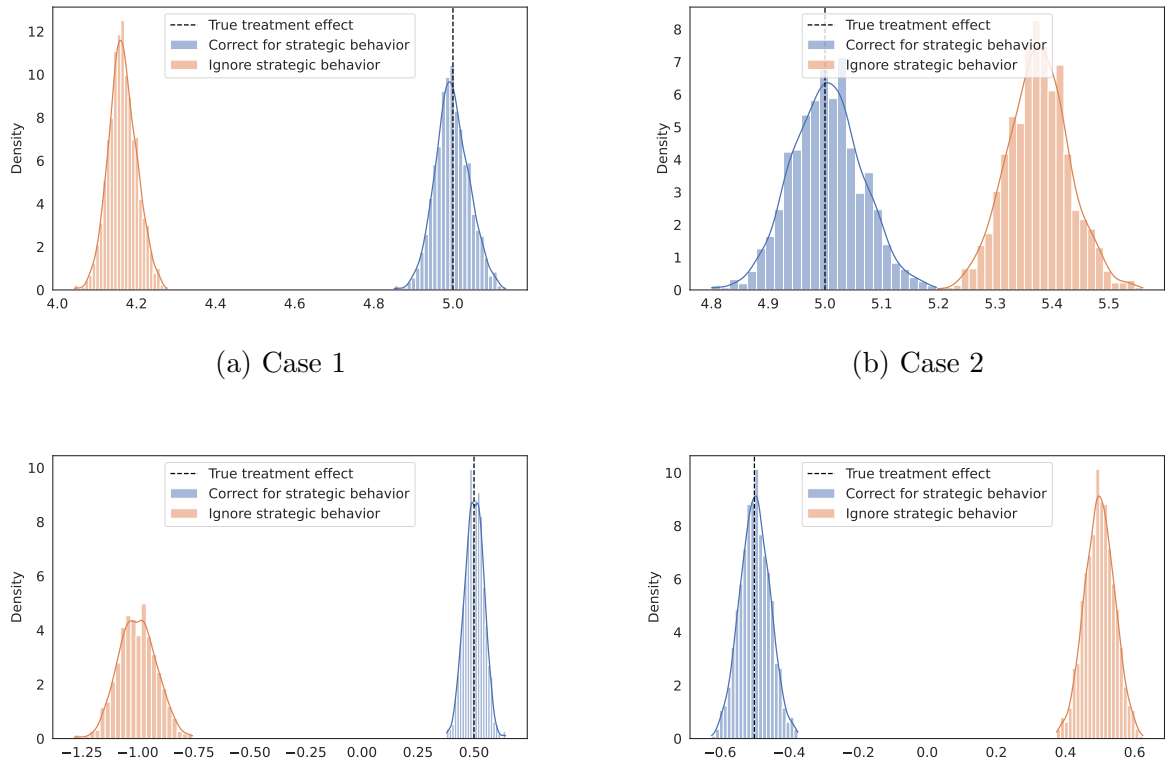
Our approach is based on the use of a validation dataset for the treatment group. Validation datasets are commonly used to correct bias induced by misreporting variables (Freedman, 1991; Chen et al., 2005; Martinelli and Parker, 2009; Blattman et al., 2016). In each scenario, we provide a plug-in estimator for ATU. Notably, in Scenario 2, we leverage the optimal transport map to facilitate identification. We demonstrate that our proposed plug-in estimators are

consistent and asymptotically normal. As an extension to the plug-in estimators, we derive the Neyman orthogonal moment and provide a double machine learning (DML) estimator (Chernozhukov et al., 2018) for each scenario. Lastly, we study the performance of our plug-in estimators through (1) Monte Carlo simulations and (2) empirical application. We apply our approach to a self-reported criminal activity dataset in Blattman et al. (2016) to estimate the treatment effect of cash and therapy on reducing the sensitive behavior of the respondents. Utilizing only a small proportion of the validation dataset, we can rectify the bias and attain an estimator with enhanced statistical power.

Our framework adds value to the literature on causal inference and measurement error in three ways. First, we show that using the reported outcome to estimate ATU can yield a biased estimator if individuals strategically misreport the outcome. Accordingly, we study the non-parametric identification condition of ATU for this problem. Second, we provide theoretical results for our plug-in estimators: we show that our estimators are consistent and asymptotic normal. Finally, we incorporate the multivariate ranks into the identification assumption based on the optimal transport theory. We add to the emerging literature on integrating the optimal transport theory into causal inference frameworks (Chen et al., 2023).

Simulated example. We present a simulated example in Figure 2.1 to show that when individuals engage in strategic misreporting, using the reported outcome to estimate the ATU can lead to a biased estimator for the true ATU. Figure 2.1 shows that using the reported outcome could lead to either underestimation or overestimation. It could even reverse the sign of the true ATU. Contrary to this, our proposed plug-in estimator demonstrates robust performance in this Monte Carlo simulation setting.

The remainder of this chapter is structured as follows. Section 2.2 discusses the related work. Section 2.3 use a simple economic model to illustrate individual strategic behavior and provides identification for the ATU under the baseline scenario and two alternative scenarios. We also introduce plug-in estimators for the ATUs in this section. Section 2.4 provides the asymptotic results for the proposed plug-in estimators. Section 2.5 extends the proposed estimators under the DML framework (Chernozhukov et al., 2018): we introduce



(a) Case 1

(b) Case 2

(c) Case 3

(d) Case 4

Figure 2.1: Comparison of different ATU estimators.

The true ATU is represented by the blue dashed line. We conducted 1,000 Monte Carlo simulation runs. Using the reported outcome, the estimators from 1,000 simulation runs are shown in the orange histogram. The results of our plug-in estimator are shown in the blue histogram. Figure 2.1a demonstrates that using the reported outcome to estimate ATU can lead to an underestimate of the true treatment effect. Figure 2.1b illustrates that it could overestimate the true treatment effect. In more extreme cases, the estimator using the reported outcome could even have a reverse sign of the true ATU, as shown in Figures 2.1c and 2.1d. Details of the estimator we used are described in Section 2.3.2.

DML estimators with Neyman orthogonal moments. Section 2.6 provides the numerical evidence for the plug-in estimators, and section 2.7 illustrates the estimator using an empirical dataset. The final section provides concluding remarks. Technical proofs are included in a series of appendices.

2.2 Related Work

2.2.1 Strategic Misreporting Behavior

The game theory literature predicts that rational individuals are incentivized to misreport their outcomes when their utility is based on these reported outcomes (Lazear and Rosen, 1981; Holmström, 1979; Frankel and Kartik, 2019; Yu et al., 2023). These studies aim to explore the potential consequences of strategic misreporting behavior from a theoretical perspective.

The empirical literature, on the other hand, focuses on demonstrating the existence of strategic misreporting behavior due to incentives linked to reported outcomes (Angrist et al., 2010; Black et al., 2017). For example, veterans may under-report their health status to qualify for the Veterans Disability Compensation Program (Angrist et al., 2010). To address the issue of strategic misreporting when estimating causal effects, these studies employ traditional econometric tools to deal with endogeneity, such as instrumental variables (Angrist et al., 2010) and using information provided by auxiliary datasets (Black et al., 2017).

Collectively, these studies establish the presence of strategic behavior in decision-making processes and highlight its potential impact on causal inference. However, there has been limited work on developing new econometric tools to correct strategic misreporting behavior in causal inference. One exception is Harris et al. (2022), which focuses on correcting strategic misreporting behavior in covariates. Our work, in contrast, focuses on the strategic misreporting of the outcome variables and studies two different scenarios of such behavior.

2.2.2 Causal Inference and Measurement Errors in Outcomes

Our work is closely related to the literature on differential measurement error in the outcome variable within the context of causal inference. Previous studies have examined the impact of non-differential measurement error on the outcome, which is defined as the measurement error that is independent of treatment assignment (Shu and Yi, 2019; Jiang and Ding, 2020). The effect of differential measurement error on the outcome in causal inference has not been extensively studied. However, empirical evidence suggests that differential measurement error exists and could hinder researchers from accurately estimating causal effects (Blattman et al., 2016; VanderWeele and Li, 2019).

While most research focuses on non-differential measurement error in the outcome variable, there are a few exceptions. Díaz and van der Laan (2013) offers a sensitivity analysis for making inferences about the treatment effect that cannot be identified from the observed data. Although their general framework could be applied to study differential measurement error in the outcome variable, it requires prior knowledge of the sensitivity parameter, i.e., the plausible range of the magnitude of the differential measurement error — which is not always available. VanderWeele and Li (2019) estimates the sensitivity parameter using validated outcomes in both the treatment and control groups. While this approach eliminates the need for prior knowledge of the sensitivity parameter, the availability of a validation dataset in both groups is not always guaranteed. In more recent work, Huang and Makar (2022) provides partial identification for the conditional treatment effect by making assumptions on the direction for the measurement error and the availability of the true outcomes in both groups.

Our work diverges from the existing literature in three perspectives. First, we focus on differential measurement errors arising from individuals' strategic misreporting behavior – individuals strategically misreport due to incentives linked to the reported outcomes. This narrower focus helps to understand the formulation of differential measurement error. Second, by gaining intuition of strategic misreporting behavior from a simple economic model, we

provide a natural identification assumption for the ATU. This eliminates the need for prior knowledge or estimation of the sensitivity parameter, as required in Díaz and van der Laan (2013); VanderWeele and Li (2019). Third, although a general framework like the one presented in Díaz and van der Laan (2013) could be applied to estimate the ATU in our context, their approach may yield conservative results. In contrast, our estimator is point-identified and is easier to make inferences.

2.2.3 Optimal Transport in Economics

The concept of Wasserstein distance and its associated optimal transport problem were first introduced by Monge (1781) and have been extensively studied in the field of mathematics. Recently, tools from optimal transport have been applied across various fields, including statistics (Hütter and Rigollet, 2019; Pooladian and Niles-Weed, 2022), computer science (Flamary et al., 2019; Peyré et al., 2019), and economics (Fan et al., 2022, 2023; Chen et al., 2023; Øystein Daljord et al., 2021; Carlier et al., 2016; Galichon, 2018).

It is relatively new for optimal transport to be applied to economics. In the field of microeconomics, Chiappori et al. (2022) explored the marriage matching market, which involves sorting based on multiple continuous attributes. To determine the equilibrium matching, they utilized the duality inherent in the transportation problem. In empirical research, Øystein Daljord et al. (2021) employed optimal transport methods to assess the size of the black market for license plates in Beijing.

In the econometrics literature, optimal transport has been applied in three main research streams. The first of these is vector quantile regression. Early work by Carlier et al. (2016) introduced vector quantile regression through the use of transport maps. Specifically, they defined the conditional vector quantile function of a random vector Y given a covariate X as a mapping between the quantile and the conditional distribution $Y|X$. Their main theorem demonstrates that under certain conditions, this map is the conditional Brenier map and can be solved using optimal transport theory. Building on this, Fan, Henry, Pass, and Rivero (2022) proposed a multivariate extension of the Lorenz curve based on the vector quantile

map. They contribute to visualize and compare inequality across multiple dimensions.

The second research stream focuses on distributionally robust estimation, which aims to make estimations robust to distributional perturbations in the data. This is achieved using Distributionally Robust Optimization (DRO) under the Wasserstein metric, which arises from optimal transport (Chen and Paschalidis, 2021). Fan, Park, and Xu (2023) studied the scenarios where sample information comes from multiple data sources and only marginal measures are identified. They incorporated DRO into their framework. Their general framework can be applied to various contexts, including the estimation of treatment effects, policy learning, and robust estimation under data combination.

The third research stream explores the application of optimal transport in causal inference. Chen, Fan, and Xue (2023) addressed the issue of limited overlap in the potential outcome framework. They used optimal transport to map between covariates in the treatment group and the control group when there was no overlap between the supports of covariates in the treatment group and the control group. Their identification assumption relies on the transport map, and they offer partial identification for the average treatment effect on the treated. While our work builds on some of the theoretical results presented in Chen, Fan, and Xue (2023), it diverges by focusing on a different problem: how to make valid causal inferences when individuals strategically misreport their outcomes. In our work, we use optimal transport to help with the identification of one specific scenario: when the incentives are linked to the rank of the reported outcome.

2.3 Model

We begin by reviewing the definition of the treatment effect within the potential outcome framework, considering individuals sampled from a population. Let D be the binary treatment indicator, where an individual with $D = 1$ receives the treatment and an individual with $D = 0$ does not. Let $Y_1^* \in \mathcal{Y} \subset \mathbb{R}^{d_Y}$ and $Y_0^* \in \mathcal{Y} \subset \mathbb{R}^{d_Y}$ denote the d_Y-dimensional true potential outcome. Define $Y^* := DY_1^* + (1 - D)Y_0^*$. Similarly, let $Y_1 \in \mathcal{Y} \subset \mathbb{R}^{d_Y}$ and $Y_0 \in \mathcal{Y} \subset \mathbb{R}^{d_Y}$ denote the d_Y-dimensional reported potential outcome. Define $Y := DY_1 + (1 - D)Y_0$.

Individual observable characteristics are denoted by $X \in \mathcal{X} \subset \mathbb{R}^{d_X}$. To simplify the notation, let $X_1 := X|D = 1$ and $X_0 := X|D = 0$.

We assume observations of $\{Y_{1i}^*, Y_{1i}, X_{1i}\}_{i=1}^{n_1}$ in the treatment group and $\{Y_{0i}, X_{0i}\}_{i=1}^{n_0}$ in the control group. Denote the sample size of the treatment group as n_1, and the sample size of the control group as n_0. We abuse the notation Y_{1i} to denote the observation of the reported outcome in the treatment group for individual i, and Y_{0j} to denote the observation of the reported outcome in the control group for individual j. Similarly, we abuse the notation of Y_{1i}^* to denote the observation for the true outcome in the treatment group for individual i. We are interested in estimating the ATU, given by:

$$\tau_o = E\left(Y_1^*|D = 0\right) - E\left(Y_0^*|D = 0\right). \tag{2.1}$$

Assumption 12 (Strong Ignorobility in Rosenbaum and Rubin (1983b))**.** For each dimension $d = 1, \ldots, d_Y$, we have (i) For all $x \in \mathcal{X}$, Y_1^{*d} is independent of D conditional on $X = x$; (ii) For all $x \in \mathcal{X}$, $0 < p(x) < 1$, where $p(x) \equiv Pr(D = 1|X = x)$.

The first part on the right-hand side of (2.1) is identified under Assumption 12. The second part remains unidentified without additional assumptions. Under Assumption 12, $E\left(Y_1^*|D = 0\right) = E\left(E\left(Y_1^*|X, D = 1\right)|D = 0\right)$. Let $W_1 := Y_1 - Y_1^*$, $W_0 := Y_0 - Y_0^*$ as the difference between the true potential outcome and the reported potential outcome. We can write the ATU as:

$$\begin{aligned}\tau_o =& E\left(E(Y_1|X, D = 1)|D = 0\right) - E\left(Y_0|D = 0\right) && (2.2)\\ &+ E\left(E(W_1|X, D = 1)|D = 0\right) - E\left[E(W_0|X, D = 0)|D = 0\right] && (2.3)\end{aligned}$$

It is noted that (2.2) can be identified using the reported outcome and the covariates in the data. However, it remains uncertain whether (2.3) is equal to zero. In the following, we use a simple economic model to demonstrate that (2.3) is not always equal to zero when strategic behavior exists.

2.3.1 A Simple Economic Model for Strategic Behavior

Baseline Scenario

Following the notation in the previous section, each individual has observed exogenous characteristics $X \in \mathcal{X}$. For $k \in \{0, 1\}$, each individual receives a potential true outcome $Y_k^* \in \mathcal{Y}^*$, and declares an potential reported outcome $Y_k \in \mathcal{Y}$. The misreporting behavior of each individual is measured by $W_k = Y_k - Y_k^* \in \mathcal{W}$. Each individual obtains a payoff by $V(\cdot) : \mathcal{X} \to \mathbb{R}$. By choosing the amount of misreporting, each individual incurs a cost $C(\cdot, \cdot) : \mathcal{W} \times \mathcal{X} \to \mathbb{R}$. The individual obtains a utility:

$$U\left(W_k; X, \xi\right) = V\left(X\right) - C\left(W_k, X\right) + \xi.$$

where ξ is an error term that represents the exogenous unobserved attribute of the individual. The utility has a standard cost-payoff structure. We do not assume a specific functional form for the payoff function $V(\cdot)$ or the cost function $C(\cdot, \cdot)$, allowing the model to be applicable in various contexts. Instead, we only require the cost function to be strictly convex with respect to W_k for every $X \in \mathcal{X}$, that is $\frac{\partial^2}{\partial W_k^2} C(W_k, X) > 0$. The convexity of the cost function results in an increasing marginal cost for deviating from reporting a true outcome.

The individual maximizes the utility by choosing how much they want to deviate from their true outcome W_k, that is

$$W_{k,optimal} = \arg\max_{W_k} U\left(W_k; X, \xi\right).$$

The first order condition gives

$$\underbrace{0}_{\text{marginal benefit}} = \underbrace{\frac{\partial}{\partial W_k} C\left(W_{k,optimal}, X\right)}_{\text{marginal cost}}. \tag{2.4}$$

Since the payoff is independent of the individual's reporting behavior and only depends on

the individual's exogenous characteristics, the marginal benefit of misreporting is zero. By the convexity of the cost function and the implicit function theorem, solving equation (2.4) gives the expression of $W_{k,optimal}$ as follows:

$$W_{k,optimal} = h(X).$$

This implies that an individual's optimal misreporting behavior depends solely on their observed characteristics X, regardless of the treatment assignment.

Figure 2.2a illustrates an individual's decision to misreport. The cost function for receiving the treatment is shown on the right, while the cost function for not receiving the treatment is shown on the left. Both functions share the same shape. However, when the true treatment effect is positive, the cost function shifts to the right. Conditional on X, an individual aims to report an outcome where the marginal cost of misreporting equals the marginal benefit of misreporting, which is zero. This results in the same amount of misreporting, regardless of treatment status.

Recalling the notation for the amount of misreporting: $W_1 = Y_1 - Y_1^*$ and $W_0 = Y_0 - Y_0^*$, we can deduce that

$$E(W_1|X = x, D = 1) = E(W_0|X = x, D = 0). \tag{2.5}$$

This equation again suggests that the misreporting behavior of the individual with the same covariates remains consistent, irrespective of their assignment to the treatment or the control group. In other words, if an individual is observed to misreport in the treatment group, they would misreport the same amount counterfactually if they were in the control group, conditional on the same observed characteristics X.

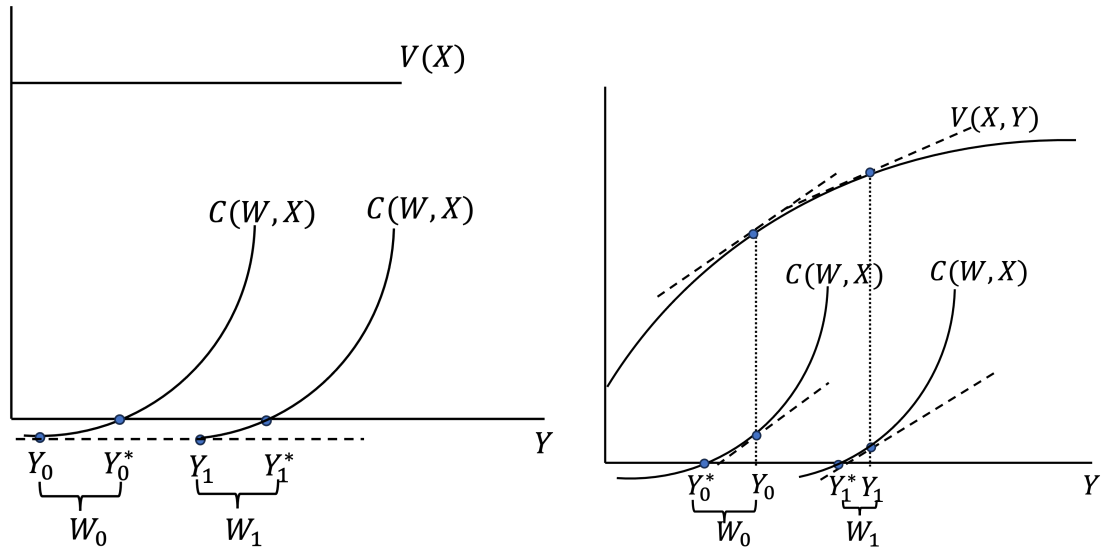


Figure 2.2: Illustration of Misreporting

Each individual faces the same optimization problem. The optimal potential reported outcome is determined by the point where the slope of the payoff function $V(X, Y)$ equals the slope of the cost function $C(W, X)$, conditional on X. Assuming that the treatment has a positive effect, Y_1^* will be higher than Y_0^*, leading to a shift in the cost function. In the left panel, the slope of the payoff function conditional on X is zero. This results in an equivalent amount of misreporting in both the treatment and control groups, conditional on X. In the right panel, the individual's optimal potential reported outcome is determined by the point where the slope of $V(X, Y)$ and $C(W, X)$ are equal, conditional on X. Due to a steeper slope of $V(X, Y)$ at Y_0 than at Y_1 conditional on X, this results in a larger amount of misreporting in the control group than in the treatment group.

Scenario 1: Strategic Behavior toward the Value of the Reported Outcome

In this section, we model the scenario where there is an incentive linked to the individual's reported outcome. Each individual's payoff depends on the observed characteristics and the reported outcome. Specifically, the individual receives a payoff of $V(X, Y_k)$, where $V(\cdot,\cdot) : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is a function determined by their observed characteristics X and the potential reported outcome Y_k for $k \in \{0, 1\}$. Assume $C(W_k, X)$ is strictly convex in W_k conditional on all $X \in \mathcal{X}$. Each individual receives a true potential outcome Y_k^* and decides how much they want to deviate from the true potential outcome. In this case, plug in $Y_k = Y_k^* + W_k$, each individual's utility is given by:

$$U(W_k; Y_k^*, X) = V(X, Y_k) - C(W_k, X) + \xi, \text{ for } k \in \{0, 1\}. \tag{2.6}$$

This framework aligns with several real-world scenarios. To understand this payoff function, we can consider university admissions as an example. Assume that admission is determined by a combination of the student's exam score and other characteristics such as gender, race, etc. In this context, the exam score is a self-reported measure of the student's understanding of the subject matter. This score can be improved—or "misreported"—by seeking assistance from other sources, such as ChatGPT. Other examples include company promotion decisions, which often depend on employee performance. In such cases, the employees have incentives to exaggerate their achievements by inflating key performance indicators (KPIs), sales records, or customer satisfaction ratings. Similarly, eligibility for specific subsidy programs often relies on self-reported income. Therefore, people have incentives to underreport their incomes to be eligible for the subsidy program.

The individual maximizes the utility by choosing how much they want to deviate from the potential true outcome, that is

$$W_k^* = \arg\max_{W_k} U(W_k; Y_k^*, X).$$

By the first order condition and the chain rule, we have

$$\underbrace{\frac{\partial}{\partial Y_k} V\left(X, Y_k\right)}_{\text{marginal benefit}} = \underbrace{\frac{\partial}{\partial W_k} C\left(W_{k,optimal}, X\right)}_{\text{marginal cost}}. \tag{2.7}$$

Figure 2.2b illustrates an individual's decision to misreport in this scenario. Although the marginal cost remains the same regardless of the treatment assignment when conditional on X, the marginal benefit varies. Due to the diminishing marginal benefit of the reported outcome, a higher potential true outcome results in a smaller marginal benefit for misreporting. For example, when a student understands the subject matter well, the marginal benefit of misreporting (e.g., cheating) diminishes. This leads to different patterns of misreporting between the treatment and control groups when the treatment effect is significant.

By implicit function theorem, we can write $W_{k,optimal}$ in (2.7) as a function of X and Y_k. This result implies that an individual's optimal misreporting behavior depends on their observable characteristics and the potential reported outcome. Hence, we assume that the following equation is valid when there is strategic behavior toward the value of the reported outcome:

$$E\left(W_1 | Y_1 = y, X = x, D = 1\right) = E\left(W_0 | Y_0 = y, X = x, D = 0\right). \tag{2.8}$$

To gain some intuition behind equation (2.8), let's assume that the cost function takes the special form $C(W_k) = \frac{W_k^2}{2}$. In this case, equation (2.7) can be rewritten as

$$W_{k,\text{optimal}} = \frac{\partial}{\partial Y_k} V(X, Y_k).$$

With this simplification, equation (2.8) reveals that individuals in both the treatment and the control groups are facing the same marginal benefit function.

Scenario 2: Strategic Behavior toward the Rank of the Reported Outcome

In some cases, incentives are linked to the rank of the reported outcome, leading individuals to engage in strategic misreporting behavior. Hence, we modify equation (2.8) to account for a scenario in which the individual values the rank of their reported outcome, rather than solely the absolute value. For example, if promotion opportunities within a company are limited to only a few top-performing employees, the rank of an employee's performance becomes a critical factor.

Similarly, various real-world situations highlight the importance of the rank of the outcome. For example, consumers in fitness services compete for rewards based on their relative number of workouts within a group. Customers exaggerate negative emotions to have their inquiries prioritized according to complaint severity. Students hire tutors to help with the rank of their exam scores for university admissions. These examples illustrate the broader applicability of the model and its relevance to diverse contexts where rank plays a pivotal role in decision-making.

To address the scenario where strategic misreporting is due to incentives linked to the rank of the reported outcome, we employ the concept of optimal transport to match the distribution of the potential reported outcome in the treatment group $Y_1|D = 1$ to the potential reported outcome in the control group $Y_0|D = 0$. We extend the case in equation (2.8) to the following:

$$E\left(W_1|Y_1 = y, X = x, D = 1\right) = E\left(W_0|Y_0 = T_o\left(y\right), X = x, D = 0\right), \tag{2.9}$$

where $T_o\left(\cdot\right)$ is a transport map that pushes the distribution of $Y_1|D = 1$ to the distribution of $Y_0|D = 0$, and is assumed to exist.

This identification assumption is grounded in the concept of measure transportation, which is the problem of transforming one distribution into another using a transport map. Specifically, let ν_1 and ν_0 represent two probability measures on $\mathbb{R}^d$. A transport map T is said to push ν_1 to ν_0, denoted as $T\#\nu_1 = \nu_0$, if and only if $T\left(X\right) \sim \nu_0$ for any $X \sim \nu_1$. The

existence and uniqueness of such a map under mild assumptions on ν_1 and ν_0 is proven by McCann (1995). This theorem has been restated in the following lemma, where $T = \nabla\psi$ defines the transport map.

Lemma 3 (McCann's Main Theorem in McCann (1995))**.** Let ν_1 and ν_0 be two probability distributions on $\mathbb{R}^d$.

1. If ν_1 is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^d$, with support contained in a convex set $\mathcal{V}_1$, the following holds: there exists a convex function $\psi : \mathcal{V}_1 \to \mathbb{R} \cup \{+\infty\}$ such that $\nabla\psi \# \nu_1 = \nu_0$. The function $\nabla\psi$ exists and is unique, ν_1-almost everywhere.

2. If, in addition, ν_0 is absolutely continuous on $\mathbb{R}^d$ with support contained in a convex set $\mathcal{V}_0$, the following holds: there exists a convex function $\psi^* : \mathcal{V}_0 \to \mathbb{R} \cup \{+\infty\}$ such that $\nabla\psi^* \# \nu_0 = \nu_1$. The function $\nabla\psi^*$ exists, is unique and equal to $\nabla\psi^{-1}$, ν_0-almost everywhere.

To gain insight on (2.9), consider the case $d_Y = 1$. It is well known that for $d_Y = 1$, $T_o(y) = F^{-1}_{Y_0|D=0}\left(F_{Y_1|D=1}(x)\right)$, see Remark 2.6 in Santambrogio (2015) or Galichon (2016). Thus (2.9) is equivalent to

$$E(W_1|Y_1 = y, X = x, D = 1) = E\left(W_0|Y_0 = F^{-1}_{Y_0|D=0}\left(F_{Y_1|D=1}(y)\right), X = x, D = 0\right). \quad (2.10)$$

Equation (2.9) states that the misreporting behavior, while considering the covariates and the rank of the misreported outcome, remains the same between the treatment group and the control group. This interpretation also extends to the multivariate scenario, where the transport map T_o serves as the $F_{Y_1|D=1}$-quantile of $F_{Y_0|D=0}$, as defined in works by Ekeland, Galichon, and Henry (2012) and Galichon and Henry (2012).

2.3.2 Identification and Estimation

Baseline Scenario

Given Assumption 12 and (2.5), our identification strategy is outlined as follows:

$$\begin{aligned}\tau_{o,baseline} &= E\left(Y_1^*|D=0\right) - E\left(Y_0^*|D=0\right) \\ &= E\left(E\left(Y_1^*|X,D=1\right)|D=0\right) - E\left(Y_0|D=0\right) + E\left(E\left(W_0|X,D=0\right)|D=0\right) \\ &= E\left(E\left(Y_1|X,D=1\right)|D=0\right) - E\left(Y_0|D=0\right).\end{aligned}$$

Let $\boldsymbol{\mu_o}(x) := E\left(Y_1|X=x,D=1\right)$, we have

$$\tau_{o,baseline} = E\left[\boldsymbol{\mu_o}(X_0)\right] - E\left(Y_0|D=0\right). \tag{2.11}$$

The estimation strategy can be described as below:

Step 1. Estimate $\boldsymbol{\mu_o}(x) = E\left(Y_1|X=x,D=1\right)$ using $\{Y_{1i}, X_{1i}\}_{i=1}^{n_1}$. Denote the estimator as $\widehat{\boldsymbol{\mu}}(x) = (\widehat{\mu}^1(x), \ldots, \widehat{\mu}^{d_Y}(x))$, where $\widehat{\mu}^d(x) = \widehat{E}\left(Y_1^d|X=x,D=1\right)$ for $d = 1, \ldots, d_Y$.

Generally, one can use any machine learning algorithm or non-parametric methods to estimate the conditional expectation function. To be self-contained, we provide the nonparametric sieve least square (LS) estimator for $\widehat{\mu}^d(x)$ here and derive the asymptotic results for this estimator in the next section. The sieve least squares (LS) estimator is described as below:

$$\widehat{\mu}^d(x) = \sum_{i=1}^{n_1} Y_{1i}^d \, p^{J_{n_1}}(X_{1i})' \left(P_1'P_1\right)^{-1} p^{J_{n_1}}(x),$$

where $p^{J_{n_1}}(x) = (p_1(x), ..., \ p_{J_{n_1}}(x))'$ and $P_1 = (p^{J_{n_1}}(X_{11}), ..., \ p^{J_{n_1}}(X_{1n_1}))'$ for some integer J_{n_1}. Here $\{p_l(x), l = 1, 2, ..., J_{n_1}\}$ denotes a sequence of known basis functions that can approximate any square-integrable function of $x \in \mathcal{X}_1$ well as $J_{n_1} \to \infty$.

Step 2.

$$\widehat{\tau}_{baseline} = \frac{1}{n_0}\sum_{j=1}^{n_0}\widehat{\boldsymbol{\mu}}(X_{0j}) - \frac{1}{n_0}\sum_{j=1}^{n_0} Y_{0j}. \tag{2.12}$$

Scenario 1: Strategic Behavior toward the Value of the reported Outcome Cont'd

Let $\boldsymbol{h_o}(x,y) := E(W_1|X=x, Y_1=y, D=1)$. Under (2.8), we can write the ATU as follows:

$$\begin{aligned}
\tau_{o,S1} &= E\left(E\left(Y_1^*|X, D=1\right)|D=0\right) - E\left(Y_0|D=0\right) + E\left(W_0|D=0\right) \\
&= E\left(E\left(Y_1^*|X, D=1\right)|D=0\right) - E\left(Y_0|D=0\right) + E\left(E\left(W_0|X, Y_0, D=0\right)|D=0\right) \\
&= E\left(E\left(Y_1|X, D=1\right)|D=0\right) - E\left(Y_0|D=0\right) + E\left(\boldsymbol{h_o}(X, Y_0)|D=0\right) - E\left(E\left(W_1|X, D=1\right)|D=0\right).
\end{aligned}$$

Recall $\tau_{o,baseline} = E\left(E\left(Y_1|X, D=1\right)|D=0\right) - E\left(Y_0|D=0\right)$ is the ATU under the baseline scenario. We can write

$$\tau_{o,baseline} = \tau_{o,S1} - \underbrace{\left[E\left(\boldsymbol{h_o}(X, Y_0)|D=0\right) - E\left(E\left(W_1|X, D=1\right)|D=0\right)\right]}_{\text{bias term}}.$$

The first component in the bias term, $E\left(\boldsymbol{h_o}(X, Y_0)|D=0\right)$, quantifies the expected level of misreporting in the control group when individuals are assumed to engage in strategic behavior toward the value of the reported outcome. This captures how much individuals would deviate from reporting the true outcome if they optimize their utility in (2.6). The second component, $E\left(E\left(W_1|X, D=1\right)|D=0\right)$, represents the expected level of misreporting in the control group under the baseline scenario: individuals' misreporting pattern is the same in the treatment group and the control group. The difference between these two terms gives us the "bias term", which helps correct the estimate of the treatment effect to account for the differential misreporting behavior.

We can also write

$$\tau_{o,S1} = E\left(E\left(Y_1^*|X, D=1\right)|D=0\right) - E\left(Y_0|D=0\right) + E\left(\boldsymbol{h_o}(X, Y_0)|D=0\right). \tag{2.13}$$

Let $\boldsymbol{\mu}_{\boldsymbol{o}}^{*}(x) := E\left(Y_1^{*}|X=x, D=1\right)$, we have

$$\tau_{o,S1} = E\left(\boldsymbol{\mu}_{\boldsymbol{o}}^{*}(X_0)|D=0\right) - E\left(Y_0|D=0\right) + E\left(\boldsymbol{h_o}\left(X, Y_0\right)|D=0\right).$$

It is noted that in this alternative scenario, validating the true outcome for the treatment group is needed[1]. Instead of imposing additional parametric assumptions about the misreporting behavior (which might not be reasonable to assume), we rely on obtaining information from validation to comprehend the nature of strategic misreporting. When the treatment group is large, validating a small proportion of the dataset is also manageable.

The estimation strategy can be described below:

Step 1. Estimate $\boldsymbol{\mu}_{\boldsymbol{o}}^{*}(x) = E\left(Y_1^{*}|X=x, D=1\right)$ using $\{Y_{1i}^{*}, X_{1i}\}_{i=1}^{n_1}$. Denote the estimator as $\widehat{\boldsymbol{\mu}}^{*}(x) = \left(\widehat{\mu}^{*1}(x), \ldots, \widehat{\mu}^{*d_Y}(x)\right)$, where $\widehat{\mu}^{*d}(x) = \widehat{E}\left(Y_1^{*d}|X=x, D=1\right)$ for $d = 1, \ldots, d_Y$. Similar to the estimation strategy in the baseline scenario, one can use any machine learning algorithm or non-parametric estimator to estimate the function.

Step 2. Estimate $\boldsymbol{h_o}(x,y) = E\left(W_1|X=x, Y_1=y, D=1\right)$ using $\{W_{1i}, X_{1i}, Y_{1i}\}_{i=1}^{n_1}$. Denote the estimator as $\widehat{\boldsymbol{h}}(x,y) = \left(\widehat{h}^{1}(x,y), \ldots, \widehat{h}^{d_Y}(x,y)\right)$, where $\widehat{h}^{d}(x,y) = \widehat{E}\left(W_1^{d}|X=x, Y_1=y, D=1\right)$ for $d = 1, \ldots, d_Y$;

Step 3.

$$\widehat{\tau}_{S1} = \frac{1}{n_0}\sum_{j=1}^{n_0}\widehat{\boldsymbol{\mu}}^{*}(X_{0j}) - \frac{1}{n_0}\sum_{j=1}^{n_0}Y_{0j} + \frac{1}{n_0}\sum_{j=1}^{n_0}\widehat{\boldsymbol{h}}\left(X_{0j}, Y_{0j}\right). \tag{2.14}$$

Scenario 2: Strategic Behavior toward the Rank of the Reported Outcome Cont'd

Under (2.9), we can write the ATU as follows:

$$\begin{aligned}\tau_{o,S2} &= E\left(E\left(Y_1^{*}|X, D=1\right)|D=0\right) - E\left(Y_0|D=0\right) + E\left(E\left(W_0|Y_0=y, X=x, D=0\right)\right)\\ &= E\left(E\left(Y_1|X, D=1\right)|D=0\right) - E\left(Y_0|D=0\right) + E\left(\boldsymbol{h_o}(X, T_o(Y_0))|D=0\right) - E\left(E\left(W_1|X, D=1\right)|D=\right.\end{aligned}$$

[1]This can be easily extended to partially validate the true outcome in one group.

$\tau_{o,baseline}$ can then be expressed as

$$\tau_{o,baseline} = \tau_{o,S2} - \underbrace{\left[E\left(\boldsymbol{h_o}(X, T_o(Y_0))|D=0\right) - E\left(E\left(W_1|X, D=1\right)|D=0\right)\right]}_{\text{bias term}}.$$

In this equation, $\tau_{o,S2}$ denotes the true ATU, which takes into account the individual's optimal misreporting behavior toward the tank of the outcome. The term $E\left(E\left(W_1|X, D=1\right)|D=0\right)$ estimates the extent of misreporting in the control group that would occur where there is no incentive linked to the reported outcome. The term $E\left(\boldsymbol{h_o}(X, T_o(Y_0))|D=0\right)$ quantifies the misreporting behavior in the control group when there are incentives linked to the rank of the reported outcome. While $\boldsymbol{h_o}(x, y) = E\left(W_1|Y_1 = y, X = x, D=1\right)$ captures the individual's optimal reporting behavior in the treatment group, the transport map $T_o(\cdot)$ maps the reported outcome in the control group to the reported outcome in the treatment group, preserving their relative ranks.

The benefit of using the optimal transport map becomes clear when the outcome variable is multidimensional. In such cases, straightforward rank-based mapping could lead to non-unique or ambiguous mappings. For instance, when a university admits students, it may consider multiple dimensions, e.g., GPA and participation in social activities. Although two students with offsetting high and low values on these dimensions (i.e., high GPA and low social activities engagement vs. low GPA and high social activities engagement) have the same rank, it would be problematic to claim that their misreporting behaviors are comparable. Optimal transport map solves this issue by providing a unique mapping that also minimizes a cost function. Recall that a transport map T push ν_1 to ν_0 is denoted as $T\#\nu_1 = \nu_0$, where ν_1 and ν_0 are two probability measures in $\mathbb{R}^d$. An optimal transport map T_o is a transport map that minimizes the object

$$\min_T \int_{\mathbb{R}^d} \|T(x) - x\|_2^2 \, \mathrm{d}\nu_1(x), \quad \text{s.t. } T_{\#}\nu_1 = \nu_0.$$

This is known as the Monge problem. Back to the previous example, it ensures that a high

GPA and low social engagement student in the treatment group is compared with a similarly high GPA and low social engagement student in the control group, making the comparison more reasonable.

We can write the treatment effect on the untreated as

$$\tau_{o,S2} = E\left(\boldsymbol{\mu}_{\boldsymbol{o}}^{*}(X)|D=0\right) - E\left(Y_0|D=0\right) + E\left(\boldsymbol{h_o}\left(X, T_o\left(Y_0\right)\right)|D=0\right).$$

The estimation strategy can be described below:

Step 1. Estimate $\boldsymbol{\mu}_{\boldsymbol{o}}^{*}(x) = E\left(Y_1^*|X=x, D=1\right)$ using $\{Y_{1i}^*, X_{1i}\}_{i=1}^{n_1}$. Denote the estimator as $\widehat{\boldsymbol{\mu}}^{*}(x) = \left(\widehat{\mu}^{*1}(x), \ldots, \widehat{\mu}^{*d_Y}(x)\right)$, where $\widehat{\mu}^{*d}(x) = \widehat{E}\left(Y_1^{*d}|X=x, D=1\right)$ for $d = 1, \ldots, d_Y$;

Step 2. Estimate $\boldsymbol{h_o}(x,y) = E\left(W_1|X=x, Y_1=y, D=1\right)$ using $\{W_{1i}, X_{1i}, Y_{1i}\}_{i=1}^{n_1}$. Denote the estimator as $\widehat{\boldsymbol{h}}(x,y) = \left(\widehat{h}^1(x,y), \ldots, \widehat{h}^{d_Y}(x,y)\right)$, where $\widehat{h}^d(x) = \widehat{E}\left(W_1^d|X=x, Y_1=y, D=1\right)$ for $d = 1, \ldots, d_Y$;

Step 3. Estimate $T_o\left(Y_0\right)$ using $\{Y_{1i}\}_{i=1}^{n_1}$ and $\{Y_{0j}\}_{j=1}^{n_0}$, denote as $\widehat{T}$. Compute $\left\{\widehat{T}\left(Y_{0j}\right)\right\}_{j=1}^{n_0}$;

Step 4.

$$\widehat{\tau}_{S2} = \frac{1}{n_0}\sum_{j=1}^{n_0}\widehat{\boldsymbol{\mu}}^{*}\left(X_{0j}\right) - \frac{1}{n_0}\sum_{j=1}^{n_0}Y_{0j} + \frac{1}{n_0}\sum_{j=1}^{n_0}\widehat{\boldsymbol{h}}\left(X_{0j}, \widehat{T}\left(Y_{0j}\right)\right). \tag{2.15}$$

The details of estimating the optimal transport map T_o are described in the next subsection.

Estimating Transport Map $T_o(\cdot)$

Following Chen, Fan, and Xue (2023), we provide the estimator for the optimal transport map in three different cases.

Case 1: $\widehat{T}$ in Univariate Nonparametric Case

When $d_Y = 1$ it is known that $T_o(y) = F^{-1}_{Y_1|D=1}\left(F_{Y_0|D=0}(y)\right)$ for $y \in \mathcal{Y}$. We can estimate $T_o(y)$ directly by

$$\widehat{T}(y) = \widehat{F}^{-1}_{Y_1|D=1}\left(\widehat{F}_{Y_0|D=0}(y)\right), \ y \in \mathcal{Y},$$

where $\widehat{F}_{Y_1|D=1}, \widehat{F}_{Y_0|D=0}$ are the empirical cdfs using $\{Y_{1i}\}_{i=1}^{n_1}$ and $\{Y_{0i}\}_{i=1}^{n_0}$. That is, for $k = 0, 1$,

$$\widehat{F}_{Y_k|D=k}(y) = \frac{1}{n_k}\sum_{i=1}^{n_k} I\{Y_{ki} \leq y\} \text{ for } y \in \mathcal{Y}.$$

Let $\{Y_{0(1)} \leq Y_{0(2)} \leq \cdots \leq Y_{0(n_0)}\}$ be the order statistics for $\{Y_{0j}\}_{j=1}^{n_0}$ and $\{Y_{1(1)} \leq Y_{1(2)} \leq \cdots \leq Y_{1(n_1)}\}$ be the order statistics for $\{Y_{1i}\}_{i=1}^{n_1}$. We can write

$$\begin{aligned}\widehat{F}^{-1}_{Y_1|D=1}(s) &:= \inf\{t : \widehat{F}_{Y_1|D=1}(t) \geq s\} \\ &= Y_{1(j)} \text{ for } \frac{j-1}{n_1} < s \leq \frac{j}{n_1} \text{ and } 1 \leq j \leq n_1.\end{aligned}$$

Case 2: $\widehat{T}$ in Multivariate Affine Map Case

When $d_Y > 1$, there is no closed form expression for T_o unless more is known about the distributions $F_{Y_1|D=1}$ and $F_{Y_0|D=0}$.

Condition 6. $F_{Y_1|D=1}$ and $F_{Y_0|D=0}$ belong to the same location-scale family with finite second moments such that

$$F_{Y_k|D=k}(y) = F_o\left(\Sigma_k^{-1/2}(y - \mu_k)\right) \text{ for } k = 0, 1,$$

for some distribution F_o with zero mean and identity variance covariance matrix, where $\mu_k = E(Y_k|D = k)$ and $\Sigma_k = Var(Y_k|D = k)$ (assumed to be positive-definite) for $k = 0, 1$.

Under Condition 6 or, more generally, for the class of elliptical distribution families with the same generator, Theorems 2.1 and 2.4 in Gelbrich (1990) or Theorem 2.1 in Cuesta-Albertos, Matrán-Bea, and Tuero-Diaz (1996) implies that the optimal transport map is

affine, i.e.,

$$T_o(y) = \mu_1 + A\,(y - \mu_0) \quad \text{for any } y \in \mathcal{Y}, \tag{2.16}$$

and

$$A = \Sigma_0^{-\frac{1}{2}} \left(\Sigma_0^{\frac{1}{2}} \Sigma_1 \Sigma_0^{\frac{1}{2}} \right)^{\frac{1}{2}} \Sigma_0^{-\frac{1}{2}} = A'$$

We note that Condition 6 implies conditions in Lemma 3, which implies that $T_o(y) = \nabla \psi_o\,(y)$ for the convex function

$$\psi_o(x) = \frac{1}{2} x' A x + (\mu_0 - A\mu_1)' x + const. \tag{2.17}$$

that is unique a.e. $F_{Y_1|D=1}$ up to an additive constant.

Following Flamary, Lounici, and Ferrari (2019), we estimate T_o given in (2.16) by

$$\widehat{T}\,(y) = \widehat{\mu}_1 + \widehat{A}\,(y - \widehat{\mu}_0) \quad \text{for any } y \in \mathcal{Y}, \tag{2.18}$$

where for $k = 0, 1$

$$\widehat{\mu}_k = \frac{1}{n_k} \sum_{i=1}^{n_k} Y_{ki}, \qquad \widehat{\Sigma}_k = \frac{1}{n_k} \sum_{i=1}^{n_k} (Y_{ki} - \widehat{\mu}_k)\,(Y_{ki} - \widehat{\mu}_k)',$$

and

$$\widehat{A} = \widehat{\Sigma}_0^{-1/2} (\widehat{\Sigma}_0^{1/2} \widehat{\Sigma}_1 \widehat{\Sigma}_0^{1/2})^{1/2} \widehat{\Sigma}_0^{-1/2}. \tag{2.19}$$

Case 3: Sieve $\widehat{T}$ in Multivariate Nonparametric Case

For $d_Y > 1$, there is no closed form expression for the optimal transport map T_o when $F_{Y_1|D=1}$ and $F_{Y_0|D=0}$ are fully nonparametric. We propose a spline sieve estimator of T_o in this subsection.

First we introduce the following condition, which is the same as Assumptions C1 and C2 of Hütter and Rigollet (2019). Denote $f_{Y_k|D=k}$ as the density of $F_{Y_k|D=k}$ with respect to the

Lebesgue measure on $\mathbb{R}^{d_Y}$ for $k = 0, 1$.

Condition 7. For $k = 0, 1$ and for some $\alpha > 1$,

1. The support of Y_k, $\mathcal{Y}$, is the closure of a uniformly convex, bounded, open subset of $\mathbb{R}^{d_Y}$ with $C^{\lfloor \alpha - 1 \rfloor + 2}$ boundary;

2. $f_{Y_k|D=k} \in C^{\alpha-1}(\mathcal{Y})$, and $f_{Y_k|D=k}$ is bounded from above and below on its support $\mathcal{Y}$.

We note that Condition 7 also implies that Lemma 3 is satisfied. Under Condition 7, Caffarelli's global regularity theorem implies that $\psi_0 \in C^{\alpha+1}(\mathcal{Y})$ for $\alpha > 1$; see Villani (2009) Theorem 12.50. Denote by $\widetilde{\psi}_o$ the extension of ψ_0 to $\widetilde{\mathcal{Y}}$, where $\widetilde{\mathcal{Y}} = \mathcal{Y} + \epsilon B(0,1)$ for an $\epsilon > 0$, the same as notation in Lemma 34 in Hütter and Rigollet (2019). Let $\widetilde{T}_o(y) = \nabla \widetilde{\psi}_o(y)$. Then it follows from Proposition-Definition 7 in Hütter and Rigollet (2019) that $\psi_0 \in \mathcal{Y}(M)$ for a finite constant M, where $\mathcal{Y}(M)$ is the set of all twice continously differentiable functions $\psi : \widetilde{\mathcal{Y}} \to \mathbb{R}$ such that

(i) $|\psi(y)| \leq 2M^2$ and $|\nabla \psi(y)| \leq M$ for all $y \in \widetilde{\mathcal{Y}}$,

(ii) $M^{-1} \preceq D^2\psi(y) \preceq M$ for all $y \in \widetilde{\mathcal{Y}}$.

Let $\Psi = \left\{ \psi \in \mathcal{Y}(2M) \cap C^{\alpha+1}\left(\widetilde{\mathcal{Y}}\right) : \int \psi(x)dx = 0 \right\}$. Let $\{b_i\}_{i=1}^{\infty}$ be a complete basis for the infinite dimensional Hilbert space $(\Psi, \|\cdot\|_\Psi)$ and let $b^k(\cdot) = (b_1(\cdot), \ldots, b_k(\cdot))'$. Denote

$$\Psi_n \equiv \left\{ \psi(\cdot) = b^{k(n)}(\cdot)'\gamma \in \Psi : \gamma \in \mathbb{R}^{k(n)} \right\}.$$

Let $k(n) = \dim(\Psi_n)$ where $n = n_1 + n_0$.

A sieve estimator of T_o is defined as

$$\widehat{T}(y) = \nabla \widehat{\psi}(y), \text{ where } \widehat{Q}\left(\widehat{\psi}\right) \leq \inf_{\psi \in \Psi_n} \widehat{Q}(\psi) + o_p\left(n^{-1}\right), \tag{2.20}$$

and

$$\widehat{Q}(\psi) = \int_{y \in \mathcal{Y}} \psi(y) \, d\widehat{F}_{Y_1|D=1}(y) + \int_{y \in \mathcal{Y}} \left[\sup_{z \in \widetilde{\mathcal{Y}}} \left(\langle z, y \rangle - \psi(z) \right) \right] d\widehat{F}_{Y_0|D=0}(y) \tag{2.21}$$

$$= \frac{1}{n_0}\sum_{i=1}^{n_0}\psi\left(Y_{0i}\right) + \frac{1}{n_1}\sum_{i=1}^{n_1}\left[\sup_{z\in\widetilde{\mathcal{Y}}}\left(\langle z, Y_{1i}\rangle - \psi\left(z\right)\right)\right]. \tag{2.22}$$

2.4 Asymptotic Results

Definition 1. For all β, a function $g(\cdot)$ is $H\left(\gamma, \omega_1\right)$-smooth[2] if it belongs to a weighted Hölder ball $\Lambda_c^{\gamma}\left(\mathcal{X}, \omega_1\right)$ for some $\gamma > 0$ and $\omega_1 \geq 0$.

2.4.1 Baseline Scenario Cont'd

Assumption 13. Let the following hold:

1. The two i.i.d samples $\{(Y_{0i}, X_{0i})\}_{i=1}^{n_0}$ and $\{(Y_{1i}, X_{1i})\}_{i=1}^{n_1}$ are independent;

2. Let $n_0 := \sum_{i=1}^{n} I\left\{D_i = 0\right\} \to \infty, n_1 := \sum_{i=1}^{n} I\left\{D_i = 1\right\} \to \infty$, and $\lambda = \lim_{n_1\to\infty}\left(n_1/n_0\right), \lambda \in (0, 1]$, where $n = n_0 + n_1$;

3. $\boldsymbol{\mu_o}(\cdot)$ is $H\left(\gamma, \omega_1'\right)$-smooth for some $\gamma > 0, \omega_1' \geq 0$;

4. $\int\left(1 + |x|^2\right)^{\omega} f_{X_1}(x)dx < \infty, \int\left(1 + |x|^2\right)^{\omega} f_{X_0}(x)dx < \infty$ for some $\omega > \omega_1' \geq 0$;

5. $\mathrm{Var}\left(Y_1 - \boldsymbol{\mu_o}\left(X\right) \mid X = x, D = 1\right)$ is bounded uniformly over x;

6. For any $H\left(\gamma, \omega_1'\right)$-smooth function $\mu^d(\cdot)$, $d = 1, \ldots, d_Y$, there is a function $\Pi_{\infty n}\mu^d$ in the sieve space $\mathcal{M}_n = \left\{\mu^d(\cdot) = p^{k_{nv}}(\cdot)'\pi\right\}$ such that $\left\|\mu^d(\cdot) - \Pi_{\infty n}\mu^d(\cdot)\right\|_{\infty,\omega} = o(1)$. Also $E\left[p^{k_{n_1}}(X)p^{k_{n_1}}(X)'|D = 1\right]$ is non-singular uniformly in k_{n_1}.

Assumption 13.1 and 13.2 impose mild constraints on the relationship between the treatment and control groups. Assumption 13.3 applies a conventional weighted smoothness condition to the function $\boldsymbol{\mu_o}(\cdot)$. This is done to facilitate accurate estimation of the unknown function $\boldsymbol{\mu_o}(\cdot)$ using the sieve estimator, by ensuring that $\boldsymbol{\mu_o}(\cdot)$ exhibits some form of

[2]See Appendix A for details on definition of smoothness of function.

smoothness with respect to x. Assumption 13.4 is a typical condition for the tail behavior of marginal densities. Assumption 13.5 provides the necessary conditions to achieve consistency. Finally, Assumption 13.6 is crucial for demonstrating that $\left\|\widehat{\mu}^d(\cdot) - \mu^d(\cdot)\right\|_{\infty,\omega} = o(1)$, as shown in the appendix.

Theorem 4. Under (2.5), Assumptions 12, 13, if $\frac{k_{n_1}}{n_1} \to 0, k_{n_1} \to \infty$, then

$$\widehat{\tau}_{baseline} - \tau_{o,baseline} = o_p(1).$$

Assumption 14. Let the following hold:

1. $E\left[\boldsymbol{\mu_o}(X)\boldsymbol{\mu_o}(X)'|D=0\right]$ is finite and positive definite;
2. $E\left[\left(\frac{f_{X_0}(X)}{f_{X_1}(X)}\right)^2 \middle| D=1\right] < \infty$;
3. $\gamma > d_X/2$ and $\omega > \omega_1 + \gamma$;
4. $k_{n_1} = O\left((n_1)^{\frac{d_X}{2\gamma+d_X}}\right)$;
5. $(n_1)^{-\frac{\gamma}{2\gamma+d_X}} \times \left\|\frac{f_{X_0}(\cdot)}{f_{X_1}(\cdot)} - \Pi_{2n}\frac{f_{X_0}(\cdot)}{f_{X_1}(\cdot)}\right\|_{2,P(X_1)} = o\left(n_1^{-1/2}\right)$.

Assumption 14.1 is a standard regularity and dominance condition required to achieve root-n consistency. Assumption 14.2 holds true when f_{X_0} is absolutely continuous relative to f_{X_1}, and $\sup_x \frac{f_{X_0}(x)}{f_{X_1}(x)} < \infty$. Assumption 14.3 imposes a higher degree of smoothness for the nuisance function. When combined, Assumptions 14.3 and 14.4 lead to the conclusion that $\|\widehat{\boldsymbol{\mu}}(\cdot) - \boldsymbol{\mu_o}(\cdot)\|_{2,P(X_1)} = O_p\left(n_1^{-\frac{\gamma}{2\gamma+d_X}}\right)$. Finally, Assumption 14.5 is met if the ratio $\frac{f_{X_0}(\cdot)}{f_{X_1}(\cdot)}$ exhibits a certain degree of smoothness, such that $\left\|\frac{f_{X_0}(\cdot)}{f_{X_1}(\cdot)} - \Pi_{2n}\frac{f_{X_0}(\cdot)}{f_{X_1}(\cdot)}\right\|_{2,P(X_1)} = o\left(n_1^{-\frac{d_X}{2(2\gamma+d_X)}}\right)$.

Theorem 5. Under (2.5), Assumptions 12, 13, 14, we have

$$\sqrt{n_1}\left(\widehat{\tau}_{baseline}-\tau_{o,baseline}\right)=\frac{1}{\sqrt{n_1}}\sum_{i=1}^{n_1}\left(Y_{1i}-\boldsymbol{\mu_o}\left(X_{1i}\right)\right)\frac{f_{X_0}\left(X_{1i}\right)}{f_{X_1}\left(X_{1i}\right)}+\sqrt{\frac{n_1}{n_0}}\frac{1}{\sqrt{n_0}}\sum_{j=1}^{n_0}\left\{\boldsymbol{\mu_o}\left(X_{0j}\right)-E\left[\boldsymbol{\mu_o}\left(X\right)|D=\right.\right. \tag{2.23}$$

$$-\sqrt{\frac{n_1}{n_0}}\frac{1}{\sqrt{n_0}}\sum_{j=1}^{n_0}\left(Y_{0j}-E\left[Y_0|D=0\right]\right)+o_p\left(1\right)$$

$$\xrightarrow{d}\mathcal{N}\left(0,V_{baseline}\right)$$

where $V_{baseline}=\lambda\Omega_{0,baseline}+\Omega_{1,baseline}$ for

$$\Omega_{0,baseline}:=E\left[\begin{matrix}\left(\boldsymbol{\mu_o}\left(X\right)-E\left(\boldsymbol{\mu_o}\left(X\right)|D=0\right)-Y_0+E\left(Y_0|D=0\right)\right)\\ \left(\boldsymbol{\mu_o}\left(X\right)-E\left(\boldsymbol{\mu_o}\left(X\right)|D=0\right)-Y_0+E\left(Y_0|D=0\right)\right)'\end{matrix}\middle|D=0\right],$$

$$\Omega_{1,baseline}:=E\left[\left(Y_1-\boldsymbol{\mu_o}\left(X\right)\right)\left(Y_1-\boldsymbol{\mu_o}\left(X\right)\right)'\frac{f_{X_0}\left(X\right)^2}{f_{X_1}\left(X\right)^2}\middle|D=1\right].$$

Theorem 5 establishes the asymptotic linear representation of the ATU estimator. The impact of estimating the unknown function $\boldsymbol{\mu_o}$ enters in the first term on the right-hand side of equation (2.23). This term captures the approximation error associated with estimating $\boldsymbol{\mu_o}$ from the data.

2.4.2 Scenario 1: Strategic Behavior toward the Value of the reported Outcome Cont'd

Let $z:=(x,y)$, $Z_1:=(X,Y_1)|D=1$, $Z_0:=(X,Y_0)|D=0$.

Assumption 15. Let the following hold:

1. The two i.i.d samples $\{(Y_{0i}^*,Y_{0i},X_{0i})\}_{i=1}^{n_0}$ and $\{(Y_{1i},X_{1i})\}_{i=1}^{n_1}$ are independent;

2. Let $n_0:=\sum_{i=1}^{n}I\left\{D_i=0\right\}\to\infty, n_1:=\sum_{i=1}^{n}I\left\{D_i=1\right\}\to\infty$, and $\lambda=\lim_{n_1\to\infty}\left(n_1/n_0\right),\lambda\in(0,1]$, where $n=n_0+n_1$;

3. $\boldsymbol{\mu_o}(\cdot)$ is $H(\gamma_1, \omega_1)$-smooth for some $\gamma_1 > 0, \omega_1 \geq 0$, $\boldsymbol{h_o}(\cdot)$ is $H(\gamma_2, \omega_2)$-smooth for some $\gamma_2 > 0, \omega_2 \geq 0$;

4. $\int (1+|x|^2)^{\omega} f_{X_1}(x)dx < \infty, \int (1+|x|^2)^{\omega} f_{X_0}(x)dx < \infty$ for some $\omega > \omega_1 \geq 0$;

5. $\int (1+|z|^2)^{\omega} f_{Z_1}(z)dz < \infty, \int (1+|z|^2)^{\omega} f_{Z_0}(z)dz < \infty$ for some $\omega > \omega_2 \geq 0$

6. $\mathrm{Var}\left(Y_1^* - \boldsymbol{\mu_o^*}(X) \mid X = x, D = 1\right)$ is bounded uniformly over x, $\mathrm{Var}\left(W_1 - \boldsymbol{h_o}(Z_1) \mid Z_1 = z, D = 1\right)$ is bounded uniformly over z;

7. For any $H(\gamma_1, \omega_1)$-smooth function $\mu^{*d}(\cdot)$, $d = 1, \ldots, d_Y$, there is a function $\Pi_{\infty n}\mu^{*d}$ in the sieve space $\mathcal{M}_n = \left\{\mu^{*d}(\cdot) = p^{k_{n_1}}(\cdot)'\pi\right\}$ such that $\left\|\mu^{*d}(\cdot) - \Pi_{\infty n}\mu^{*d}(\cdot)\right\|_{\infty,\omega} = o(1)$. Also $E\left[p^{k_{n_1}}(X)p^{k_{n_1}}(X)'|D=1\right]$ is non-singular uniformly in k_{n_1};

8. For any $H(\gamma_2, \omega_2)$-smooth function $h^d(\cdot)$, $d = 1, \ldots, d_Y$, there is a function $\Pi_{\infty n}h^d$ in the sieve space $\mathcal{H}_n = \left\{h^d(\cdot) = p^{q_{n_1}}(\cdot)'\pi\right\}$ such that $\left\|h^d(\cdot) - \Pi_{\infty n}h^d(\cdot)\right\|_{\infty,\omega} = o(1)$. Also $E\left[p^{q_{n_1}}(Z_1)p^{q_{n_1}}(Z_1)'|D=1\right]$ is non-singular uniformly in q_{n_1}.

Assumption 15 specifies that both the validated outcome and the reported outcome are jointly observable in the treatment group. This assumption can be relaxed by only observing a subset of the validated outcome in the treatment group, i.e., we have a validation set with $n_v < n_1$. The validation dataset provides information on $E[Y_1^*|X, D=1]$ and $E[W_1|X, Y_1, D=1]$. While a validation set with a smaller sample size may increase the estimation variance, the theoretical framework can be easily extended to accommodate this change, based on the existing results.

Theorem 6. Under (2.8), Assumptions 12, 15, if $\frac{k_{n_1}}{n_1} \to 0, k_{n_1} \to \infty$ and $\frac{q_{n_1}}{n_1} \to 0, q_{n_1} \to \infty$, then

$$\widehat{\tau}_{S1} - \tau_{o,S1} = o_p(1).$$

Assumption 16. Let the following hold:

1. $E\left[\boldsymbol{\mu}_{\boldsymbol{o}}^{*}(X)\boldsymbol{\mu}_{\boldsymbol{o}}^{*}(X)'|D=0\right]$ is finite and positive definite, $E\left[\boldsymbol{h}_{\boldsymbol{o}}(X,Y_0)\boldsymbol{h}_{\boldsymbol{o}}(X,Y_0)'|D=0\right]$ is finite and positive definite;
2. $E\left[\left(\frac{f_{X_0}(X)}{f_{X_1}(X)}\right)^2\middle|D=1\right]<\infty$, $E\left[\left(\frac{f_{Z_0}(z)}{f_{Z_1}(z)}\right)^2\middle|D=1\right]<\infty$;
3. $\gamma_1>d_X/2$, $\gamma_2>(d_X+d_Y)/2$ and $\omega>\max\left\{\omega_1'+\gamma_1,\omega_2+\gamma_2\right\}$;
4. $k_{n_1}=O\left((n_1)^{\frac{d_X}{2\gamma_1+d_X}}\right)$;
5. $q_{n_1}=O\left((n_1)^{\frac{d_X+d_Y}{2\gamma_2+d_X+d_Y}}\right)$;
6. $(n_1)^{-\frac{\gamma_1}{2\gamma_1+d_X}}\times\left\|\frac{f_{X_0}(\cdot)}{f_{X_1}(\cdot)}-\Pi_{2n_1}\frac{f_{X_0}(\cdot)}{f_{X_1}(\cdot)}\right\|_{2,P(X_1)}=o\left(n_1^{-1/2}\right)$;
7. $(n_1)^{-\frac{\gamma_2}{2\gamma+d_X+d_Y}}\times\left\|\frac{f_{Z_0}(\cdot)}{f_{Z_1}(\cdot)}-\Pi_{2n_1}\frac{f_{Z_0}(\cdot)}{f_{Z_1}(\cdot)}\right\|_{2,P(Z_1)}=o\left(n_1^{-1/2}\right)$.

Theorem 7. Under (2.8), Assumptions 12, 15, 16, we have

$$\begin{aligned}\sqrt{n_1}\left(\widehat{\tau}_{S1}-\tau_{o,S1}\right)=&\frac{1}{\sqrt{n_0}}\sum_{j=1}^{n_0}\sqrt{\frac{n_1}{n_0}}\left\{\begin{array}{c}\boldsymbol{\mu}_{\boldsymbol{o}}^{*}\left(X_{0j}\right)-E\left(\boldsymbol{\mu}_{\boldsymbol{o}}^{*}\left(X\right)|D=0\right)+\boldsymbol{h}_{\boldsymbol{o}}\left(X_{0j},Y_{0j}\right)-E\left(\boldsymbol{h}_{\boldsymbol{o}}\left(X,Y_0\right)|D=0\right)\\-Y_{0j}+E\left(Y_0|D=0\right)\end{array}\right\}\\&+\frac{1}{\sqrt{n_1}}\sum_{i=1}^{n_1}\left\{\frac{f_{X_0}\left(X_{1i}\right)}{f_{X_1}\left(X_{1i}\right)}\left(Y_{1i}^{*}-\boldsymbol{\mu}_{\boldsymbol{o}}^{*}\left(X_{1i}\right)\right)+\frac{f_{Z_0}\left(X_{1i},Y_{1i}\right)}{f_{Z_1}\left(X_{1i},Y_{1i}\right)}\left(W_{1i}-\boldsymbol{h}_{\boldsymbol{o}}\left(X_{1i},Y_{1i}\right)\right)\right\}\\&+o_p\left(1\right)\\\xrightarrow{d}&\mathcal{N}\left(0,V_{S1}\right),\end{aligned}\tag{2.24}$$

(2.25)

where $V_{S1}=\lambda\Omega_{0,S1}+\Omega_{1,S1}$ for

$$\Omega_{0,S1}:=E\left[\begin{array}{c}\left(\boldsymbol{\mu}_{\boldsymbol{o}}^{*}\left(X\right)-E\left(\boldsymbol{\mu}_{\boldsymbol{o}}^{*}\left(X\right)|D=0\right)+\boldsymbol{h}_{\boldsymbol{o}}\left(X,Y_0\right)-E\left(\boldsymbol{h}_{\boldsymbol{o}}\left(X,Y_0\right)|D=0\right)-Y_0+E\left(Y_0|D=0\right)\right)\\\left(\boldsymbol{\mu}_{\boldsymbol{o}}^{*}\left(X\right)-E\left(\boldsymbol{\mu}_{\boldsymbol{o}}^{*}\left(X\right)|D=0\right)+\boldsymbol{h}_{\boldsymbol{o}}\left(X,Y_0\right)-E\left(\boldsymbol{h}_{\boldsymbol{o}}\left(X,Y_0\right)|D=0\right)-Y_0+E\left(Y_0|D=0\right)\right)'\end{array}\middle|D=\right.$$

$$\Omega_{1,S1} := E\left[\begin{array}{c}\left(\frac{f_{X_0}(X)}{f_{X_1}(X)}\left(Y_1^* - \boldsymbol{\mu_o^*}(X)\right) + \frac{f_{Z_0}(X,Y_1)}{f_{Z_1}(X,Y_1)}\left(W_1 - \boldsymbol{h_o}(X,Y_1)\right)\right)' \\ \left(\frac{f_{X_0}(X)}{f_{X_1}(X)}\left(Y_1^* - \boldsymbol{\mu_o^*}(X)\right) + \frac{f_{Z_0}(X,Y_1)}{f_{Z_1}(X,Y_1)}\left(W_1 - \boldsymbol{h_o}(X,Y_1)\right)\right)\end{array}\middle| D=1\right].$$

In Theorem 7, the asymptotic linear representation (2.25) comprises two main components. The first part is related to estimating the nuisance function $\boldsymbol{\mu_o^*}$ from the data. The second part is related to estimating another nuisance function $\boldsymbol{h_o}$ from the data.

2.4.3 Scenario 2: Strategic Behavior toward the Rank of the Reported Outcome Cont'd

Assumption 17. Suppose $\widehat{T}(y) \in \mathcal{Y}$ w.p.1 for each $y \in \mathcal{Y}$,

$$\left\|\boldsymbol{h_o}\left(\cdot, \widehat{T}(\cdot)\right) - \boldsymbol{h_o}(\cdot, T_o(\cdot))\right\|_{\infty,\omega} = o_p(1).$$

Theorem 8. Under (2.9), Assumption 12, 15, 17, if $\frac{k_{n_1}}{n_1} \to 0, k_{n_1} \to \infty$ and $\frac{q_{n_1}}{n_1} \to 0, q_{n_1} \to \infty$, then

$$\widehat{\tau}_{S2} - \tau_{o,S2} = o_p(1).$$

Assumption 18. Assume following holds:

1. $$\left\|\frac{\partial\left\{\widehat{\boldsymbol{h}}(\cdot, T_o(\cdot)) - \boldsymbol{h_o}(\cdot, T_o(\cdot))\right\}}{\partial T}\right\|_{2,P(Z_1)} \times \left\|\widehat{T}(\cdot) - T_o(\cdot)\right\|_{2,P(Z_1)} = o_p\left(n_1^{-1/2}\right).$$

2. There exist functions φ_1 and φ_0 such that

$$\begin{aligned}&\frac{1}{n_1}\sum_{i=1}^{n_0}\left[\boldsymbol{h_o}\left(X_{0i}, \widehat{T}(Y_{0i})\right) - \boldsymbol{h_o}(X_{0i}, T_o(Y_{0i}))\right] \\ = \;& \frac{1}{n_0}\sum_{i=1}^{n_0}\varphi_0(Y_{0i}) + \frac{1}{n_1}\sum_{j=1}^{n_1}\varphi_1(Y_{1j}) + o_p\left(n_1^{-1/2}\right),\end{aligned}$$

where $E[\varphi_1(Y_1)] = 0$ and $E[\varphi_0(Y_0)] = 0$. The expressions for φ_1 and φ_0 are mentioned in the next subsection.

3. $E\left[\boldsymbol{\mu}_{\boldsymbol{o}}^{*}(X)\boldsymbol{\mu}_{\boldsymbol{o}}^{*}(X)'|D=0\right]$ is finite and positive definite, $E\left[\boldsymbol{h}_{\boldsymbol{o}}(X,Y_0)\boldsymbol{h}_{\boldsymbol{o}}(X,Y_0)'|D=0\right]$ is finite and positive definite;

4. $E\left[\left(\frac{f_{X_0}(X)}{f_{X_1}(X)}\right)^2\middle|D=1\right]<\infty$, $E\left[\left(\frac{f_{Z_0}\left(X,T_o^{-1}(Y_1)\right)}{f_{Z_1}(X,Y_1)}\frac{f_{Y_1|D=1}(Y_1)}{f_{Y_0|D=0}\left(T_o^{-1}(Y_1)\right)}\right)^2\middle|D=1\right]<\infty$;

5. $\gamma_1>d_X/2$, $\gamma_2>(d_X+d_Y)/2$ and $\omega>\max\left\{\omega_1'+\gamma_1,\omega_2+\gamma_2\right\}$;

6. $k_{n_1}=O\left((n_1)^{\frac{d_X}{2\gamma_1+d_X}}\right)$;

7. $q_{n_1}=O\left((n_1)^{\frac{d_X+d_Y}{2\gamma_2+d_X+d_Y}}\right)$;

8. $(n_1)^{-\frac{\gamma_1}{2\gamma_1+d_X}}\times\left\|\frac{f_{X_0}(\cdot)}{f_{X_1}(\cdot)}-\Pi_{2n_1}\frac{f_{X_0}(\cdot)}{f_{X_1}(\cdot)}\right\|_{2,P(Z_1)}=o\left(n_1^{-1/2}\right)$;

9. $(n_1)^{-\frac{\gamma_2}{2\gamma+d_X+d_Y}}\times\left\|\frac{f_{Z_0}\left(x,T_o^{-1}(y)\right)}{f_{Z_1}(x,y)}\frac{f_{Y_1|D=1}(y)}{f_{Y_0|D=0}\left(T_o^{-1}(y)\right)}-\Pi_{2n_1}\frac{f_{Z_0}\left(x,T_o^{-1}(y)\right)}{f_{Z_1}(x,y)}\frac{f_{Y_1|D=1}(y)}{f_{Y_0|D=0}\left(T_o^{-1}(y)\right)}\right\|_{2,P(Z_1)}=o\left(n_1^{-1/2}\right)$.

Assumption 17, 18.1, 18.2 are high-level assumption on $\widehat{T}$ and can be verified for optimal transport map estimator introduced in Section 2.3.2 cases 1-3. Chen, Fan, and Xue (2023) verified these assumptions in their paper in Section 5.

Theorem 9. Under (2.9). Assumptions 12, 15, 17, 18, we have

$$\sqrt{n_1}\left(\widehat{\tau}_{S2}-\tau_{o,S2}\right)=\frac{1}{\sqrt{n_0}}\sum_{j=1}^{n_0}\sqrt{\frac{n_1}{n_0}}\left\{\begin{array}{c}\boldsymbol{\mu}_{\boldsymbol{o}}^{*}\left(X_{0j}\right)-E\left(\boldsymbol{\mu}_{\boldsymbol{o}}^{*}\left(X\right)|D=0\right)\\+\boldsymbol{h}_{\boldsymbol{o}}\left(X_{0j},T_o(Y_{0j})\right)-E\left[\boldsymbol{h}_{\boldsymbol{o}}\left(X,T_o(Y_0)\right)|D=0\right]\\-Y_{0j}+E\left(Y_0|D=0\right)+\varphi_0\left(Y_{0j}\right)-E\left(\varphi_0\left(Y_0\right)|D=0\right)\end{array}\right\} \tag{2.26}$$

$$+\frac{1}{\sqrt{n_1}}\sum_{i=1}^{n_1}\left\{\begin{array}{c}\frac{f_{X_0}(X_{1i})}{f_{X_1}(X_{1i})}\left(Y_{1i}^{*}-\boldsymbol{\mu}_{\boldsymbol{o}}^{*}\left(X_{1i}\right)\right)\\+\frac{f_{Z_0}\left(X_1,T_o^{-1}(Y_{1i})\right)}{f_{Y_0|D=0}\left(T_o^{-1}(Y_{1i})\right)}\frac{f_{Y_1|D=1}(Y_{1i})}{f_{Z_1}(X_{1i},Y_{1i})}\left(W_{1i}-\boldsymbol{h}_{\boldsymbol{o}}\left(X_{1i},Y_{1i}\right)\right)\\+\varphi_1\left(Y_{1i}\right)-E\left(\varphi_1\left(Y_1\right)|D=1\right)\end{array}\right\} \tag{2.27}$$

$$
\begin{aligned}
&+ o_p(1) \\
&\xrightarrow{d} \mathcal{N}(0, V_{S2}),
\end{aligned}
$$

where $V_{S2} = \lambda\Omega_{0,S2} + \Omega_{1,S2}$ for

$$
\Omega_{0,S2} := E\left[\begin{Bmatrix} (\boldsymbol{\mu_o^*}(X) - E(\boldsymbol{\mu_o^*}(X)|D=0) + \boldsymbol{h_o}(X, T_o(Y_0)) - E(\boldsymbol{h_o}(X, T_o(Y_0))|D=0) \\ -Y_0 + E(Y_0|D=0) + \varphi_0(Y_0) - E(\varphi_0(Y_0)|D=0)) \end{Bmatrix}^{\otimes 2} \middle| D=0\right],
$$

$$
\Omega_{1,S2} := E\left[\begin{pmatrix} \frac{f_{X_0}(X)}{f_{X_1}(X)}(Y_1^* - \boldsymbol{\mu_o^*}(X)) \\ +\frac{f_{Z_0}\left(X, T_o^{-1}(Y_1)\right)}{f_{Y_0}\left(T_o^{-1}(Y_1)\right)} \frac{f_{Y_1}(Y_1)}{f_{Z_1}(X,Y_1)}(W_1 - \boldsymbol{h_o}(Z_1)) \\ +\varphi_1(Y_1) - E(\varphi_1(Y_1)|D=1) \end{pmatrix}^{\otimes 2} \middle| D=1\right],
$$

In Theorem 9, equations (2.26) and (2.27) provide the asymptotic linear representations for $\widehat{\tau}_{S2}$. Similar to Theorem 7, the first part of (2.27) pertains to the estimation of the nuisance function $\boldsymbol{\mu_o^*}$. However, unlike Theorem 7, the second part of the equation accounts for the estimation of $\boldsymbol{h_o}$ while also incorporating the transport map T_o. Specifically, this differentiation arises from the assumption of (2.9): $E[W_1|Y_1 = T_o(y), X = x, D = 1] = E[W_0|Y_0 = y, X = x, D = 0]$. This term introduces $h_o(\cdot)$ into the estimator in a way that accounts for $T_o(\cdot)$. The effect of estimating this transport map is brought into the estimator through the terms φ_1 and φ_0.

2.4.4 Expression for φ_1 and φ_0 in Assumption 8

Case 1: $\widehat{T}$ in Univariate Nonparametric Case

Denote

$$
Q(y, X, Y_0) = -\frac{h_{o,2}(X, T_o(Y_0))}{f_{Y_1|D=1}(T_o(Y_0))} \times \left[I\{y \leq T_o(Y_0)\} - F_{Y_0|D=0}(Y_0)\right], \quad y \in \mathcal{Y},
$$

Further let

$$q(y) = E\left[Q(y, X, Y_0)|D = 0\right], \quad y \in \mathcal{Y}.$$

Now $\varphi_1(\cdot)$ and $\varphi_0(\cdot)$ mentioned in the previous section take the form of

$$\varphi_0\left(Y_0\right) = -q\left(T_o\left(Y_0\right)\right); \quad \varphi_1\left(Y_1\right) = q\left(Y_1\right).$$

Case 2: $\widehat{T}$ in Multivariate Affine Map Case

Denote the eigendecomposition $\Sigma_0\Sigma_1 = UDU^{-1}$, λ_d as the d-th eigenvalue of diagonal matrix D, $[L]_{d,s} = \frac{1}{\sqrt{\lambda_d}+\sqrt{\lambda_s}}$ and $[K]_{d,s} = \frac{\sqrt{\lambda_d\lambda_s}}{\sqrt{\lambda_d}+\sqrt{\lambda_s}}$ as the d-th and s-th element of matrix L and K. Denote $A \circ B$ the Hadamard product between matrices. $\varphi_1(\cdot)$ and $\varphi_0(\cdot)$ mentioned in the previous section take the following form:

$$\begin{aligned}
\varphi_0(y) =& - E\left[h_{o,2}\left(X, T_o\left(Y_0\right)\right)' A\left(y - \mu_0\right)|D = 0\right] \\
& - E\left[h_{o,2}\left(X, T_o\left(Y_0\right)\right)' \left(\Sigma_1 U\left[K \circ \left(U^{-1}\Sigma_1^{-1}\Sigma_0^{-1}\left(\Sigma_0(y) - \Sigma_0\right)\Sigma_0^{-1}U\right)\right]U^{-1}\right)\left(Y_0 - \mu_0\right)|D = 0\right], \\
\varphi_1(y) =& E\left[h_{o,2}\left(X, T_o\left(Y_0\right)\right)'\left(y - \mu_1\right)|D = 0\right] \\
& + E\left[h_{o,2}\left(X, T_o\left(Y_0\right)\right)' \left(\Sigma_0^{-1} U\left[L \circ \left(U^{-1}\Sigma_0\left(\Sigma_1(y) - \Sigma_1\right)U\right)\right]U^{-1}\right)\right)\left(Y_0 - \mu_0\right)|D = 0\right].
\end{aligned}$$

Case 3: Sieve $\widehat{T}$ in Multivariate Nonparametric Case

Under this case, $\varphi_1(\cdot)$ and $\varphi_0(\cdot)$ mentioned in the previous section take the following form:

$$\varphi_0\left(y\right) = -[v^*\left(y\right) - E(v^*\left(Y_0\right)|D = 0)]; \quad \varphi_1\left(y\right) = v^*\left(T_o^{-1}\left(y\right)\right) - E\left[v^*\left(T_o^{-1}\left(Y_1\right)\right)|D = 1\right], \tag{2.28}$$

where

$$\nabla v^*\left(y\right) = E\left[\frac{h_{o,2}\left(X, T_o(y)\right) f_{X,Y_0|D=0}(X, y)}{f_{Y_0|D=0}(y) f_{X_0}(X)} \nabla T_o(y) \middle| D = 0\right]. \tag{2.29}$$

2.5 Extension to Double Machine Learning Estimator

2.5.1 Double Machine Learning Estimator for Baseline Scenario

We also derive the Neyman orthogonal moment described in Chernozhukov et al. (2018) to incorporate high-dimensional cases and use machine learning algorithms in estimating the nuisance function $\boldsymbol{\mu_o}$. The Neyman orthogonal moment is described as the following proposition:

Proposition 7. The Neyman orthogonal moment for $\tau_{o,baseline}$ is described as below:

$$\tau^{DML}_{o,baseline} = E\left[\left(Y_1 - \boldsymbol{\mu}\left(X_1\right)\right)\frac{p}{1-p}\frac{1-m(X_1)}{m(X_1)}\middle| D=1\right] + E\left[\boldsymbol{\mu}\left(X_0\right) - Y_0 | D=0\right].$$

where the nuisance parameter $\eta = (\boldsymbol{\mu}, m, p)$ consists of function $\boldsymbol{\mu}$ mapping $\mathcal{X}$ to $\mathbb{R}^{d_Y}$, m mapping $\mathcal{X}$ to $\mathbb{R}$, and a constant $p \in (\varepsilon, 1-\varepsilon)$, for some $\varepsilon \in (0, 1/2)$. The true value of η is $\eta_0 = (\boldsymbol{\mu_o}, m_o, p_o)$, where $m_o(x) = E[D|x]$ and $p_o = E[D]$.

With Proposition 7, the DML estimator is described below:

Step 1. Take a K-fold random partition $(I_{k,1})_{k=1}^K$ of observation indices $[n_1] = \{1, \ldots, n_1\}$ such that the size of each fold $I_{k,1}$ is $N_1 = n_1/K$. Also, for each $k \in [K] = \{1, \ldots, K\}$, define $I^c_{k,1} := \{1, \ldots, n_1\} \backslash I_{k,1}$. Similarly, Take a K-fold random partition $(I_{k,0})_{k=1}^K$ of observation indices $[n_0] = \{1, \ldots, n_0\}$ such that the size of each fold $I_{k,0}$ is $N_0 = n_0/K$, define $I^c_{k,0} := \{1, \ldots, n_0\} \backslash I_{k,0}$;

Step 2. For each $k \in [K]$, estimate $\boldsymbol{\mu_o}(x) = E\left(Y_1 | X = x, D = 1\right)$ by any machine learning algorithm, denote as $\widehat{\boldsymbol{\mu}}_k(x)$, using observations in $I_{k,1}$. Estimate $\nu\left(x\right) := \frac{p}{1-p}\frac{1-m(X_1)}{m(X_1)}$ using any machine learning algorithm, denote as $\widehat{\nu_k}\left(x\right)$, using observations in $I_{k,1}$ and $I_{k,0}$;

Step 3.

$$\widehat{\tau}_{baseline}^{DML} = \frac{1}{K}\sum_{k=1}^{K}\left\{\frac{1}{N}\sum_{i=1}^{N}\left(Y_{1i} - \widehat{\boldsymbol{\mu}}_k\left(X_{1i}\right)\right)\widehat{\nu}_k\left(X_{1i}\right) + \frac{1}{n_0}\sum_{j=1}^{n_0}\left(\widehat{\boldsymbol{\mu}}_k\left(X_{0j}\right) - Y_{0j}\right)\right\}.$$

2.5.2 Double Machine Learning Estimator for Scenario 1

To integrate a DML estimator into this framework, we propose the following Neyman orthogonal moment.

Proposition 8. The Neyman orthogonal moment for $\tau_{o,S1}$ is described as below:

$$\begin{aligned}\tau_{o,S1}^{DML} =& E\left[\boldsymbol{\mu}^{*}(X_0)|D=0\right] - E\left[Y_0|D=0\right] + E\left[\boldsymbol{h}\left(Z_0\right)|D=0\right] \\ &+ E\left[\nu_1\left(X_1\right)\left(Y_1^{*} - \boldsymbol{\mu}^{*}\left(X_1\right)\right) + \nu_2\left(Z_1\right)\left(W_1 - \boldsymbol{h}\left(Z_1\right)\right)|D=1\right],\end{aligned}$$

where $Z_1 = (X_1, Y_1)$, $\nu_1(x) = \frac{f_{X_0}(x)}{f_{X_1}(x)}$ and $\nu_2(z) = \frac{f_{Z_0}(z)}{f_{Z_1}(z)}$.

With Proposition 8, the DML estimator is described below:

Step 1. Take a K-fold random partition $(I_k)_{k=1}^{K}$ of observation indices $[n_1] = \{1, \ldots, n_1\}$ such that the size of each fold I_k is $N = n_1/K$. Also, for each $k \in [K] = \{1, \ldots, K\}$, define $I_k^c := \{1, \ldots, n_1\} \backslash I_k$;

Step 2 For each $k \in [K]$, estimate $\boldsymbol{\mu}_{\boldsymbol{o}}^{*}(x) = E\left(Y_1^{*}|X=x, D=1\right)$ by any machine learning algorithm, denote as $\widehat{\boldsymbol{\mu}}_k^{*}(x)$, using observations in I_k. For $d = 1, \ldots, d_Y$, $\widehat{\mu}_k^{*d}(x) = \widehat{E}\left(Y_1^{*d}|X=x, D=1\right)$. Estimate $\boldsymbol{h}_{\boldsymbol{o}}(z) = E\left(W_1|X=x, Y_1=y, D=1\right)$, denote the estimator as $\widehat{\boldsymbol{h}}_k(z)$, using observations in I_k;

Step 3. For each $k \in [K]$, estimate $\nu_1(x) = \frac{f_{X_0}(x)}{f_{X_1}(x)}$, denote as $\widehat{\nu}_{1,k}(x)$. Specifically, estimate $f_{X_0}\left(x\right)$ using all observations in control group and estimate $f_{X_1}\left(x\right)$ using observations in I_k. Similarly, for each $k \in [K]$, estimate $\nu_2(Z) = \frac{f_{Z_0}(Z)}{f_{Z_1}(Z)}$, denote as $\widehat{\nu}_{2,k}(z)$;

Step 4. The estimator is then given by

$$\widehat{\tau}_{S1}^{DML} = \frac{1}{K}\sum_{k=1}^{K}\left\{\frac{1}{n_0}\sum_{j=1}^{n_0}\widehat{\boldsymbol{\mu}}_k^*(X_{0j}) - \frac{1}{n_0}\sum_{j=1}^{n_0}Y_{0j} + \frac{1}{n_0}\sum_{j=1}^{n_0}\widehat{\boldsymbol{h}}_k\left(Z_{0j}\right) + \frac{1}{N}\sum_{i=1}^{N}\left\{\widehat{\nu}_{1,k}\left(X_{1i}\right)\left(Y_{1i}^* - \widehat{\boldsymbol{\mu}}_k^*\left(X_{1i}\right)\right) + \widehat{\nu}_{2,k}\left(Z_{1i}\right)\left(W_{1i} - \widehat{\boldsymbol{h}}_k\left(Z_{1i}\right)\right)\right\}\right\}.$$

2.5.3 Double Machine Learning Estimator for Scenario 2

Proposition 9. The Neyman orthogonal moment for $\tau_{o,S2}$ is described as below:

$$\begin{aligned}\tau_{o,S2}^{DML} =& E\left[\boldsymbol{\mu}^*\left(X_0\right) + \boldsymbol{h}\left(X_0, T_o(Y_0)\right) - Y_0 + \varphi_0\left(Y_0\right) - E\left(\varphi_0\left(Y_0\right)|D=0\right)|D=0\right] \\ &+ E\left[\nu_1(X_1)\left(Y_1^* - \boldsymbol{\mu}^*\left(X_1\right)\right) + \nu_2'(X_1, Y_1)\left(W_1 - \boldsymbol{h}\left(Z_1\right)\right) + \varphi_1\left(Y_1\right) - E\left(\varphi_1\left(Y_1\right)|D=1\right)\middle| D=1\right],\end{aligned}$$

where $\nu_1(x) = \frac{f_{X_0}(x)}{f_{X_1}(x)}$, $\nu_2'(x,y) = \frac{f_{Z_0}\left(x, T_o^{-1}(y)\right)}{f_{Z_1}(x,y)}\frac{f_{Y_1|D=1}(y)}{f_{Y_0|D=0}\left(T_o^{-1}(y)\right)}$ and the form of $\varphi_1(\cdot)$ and $\varphi_0(\cdot)$ is given in section 2.4.4.

Given the orthogonal moment in Proposition 9, the estimator is described as below:

Step 1. Take a K-fold random partition $(I_{1,k})_{k=1}^K$ of observation indices $[n_1] = \{1, \ldots, n_1\}$ such that the size of each fold $I_{1,k}$ is $N_1 = n_1/K$. Also, for each $k \in [K] = \{1, \ldots, K\}$, define $I_{1,k}^c := \{1, \ldots, n_1\} \backslash I_{1,k}$;

Step 2. Similarly, take a K-fold random partition $(I_{0,k})_{k=1}^K$ of observation indices $[n_0] = \{1, \ldots, n_0\}$ such that the size of each fold $I_{0,k}$ is $N_0 = n_0/K$. Also, for each $k \in [K] = \{1, \ldots, K\}$, define $I_{0,k}^c := \{1, \ldots, n_0\} \backslash I_{0,k}$;

Step 3. For each $k \in [K]$,

3.1 using observations in $I_{1,k}$, estimate $\boldsymbol{\mu}_{\boldsymbol{o}}^*(x) = E\left(Y_1^*|X=x, D=1\right)$ by any machine learning algorithm, denote as $\widehat{\boldsymbol{\mu}}_k^*(x)$. Estimate $\boldsymbol{h}_{\boldsymbol{o}}(z) = E\left(W_1|X=x, Y_1=y, D=1\right)$,

denote the estimator as $\widehat{\boldsymbol{h}}_k(z)$. Estimate $\varphi_1(y)$, denote as $\widehat{\varphi}_{1,k}(y)$. Estimate $f_{Z_1}(\cdot)$, denote as $\widehat{f}_{Z_1}(\cdot)$. Estimate $f_{Y_1|D=1}(\cdot)$, denote as $\widehat{f}_{Y_1|D=1}(\cdot)$;

3.2 using observations in $I_{0,k}$, estimate $\varphi_0(y)$, denote as $\widehat{\varphi}_{0,k}(y)$. Estimate $f_{Z_0}(\cdot)$, denote as $\widehat{f}_{Z_0}(\cdot)$. Estimate $f_{Y_0|D=0}(\cdot)$, denote as $\widehat{f}_{Y_0|D=0}(\cdot)$;

3.3 get $\widehat{\nu}_{1,k}(x) = \frac{\widehat{f}_{X_0}(x)}{\widehat{f}_{X_1}(x)}$, $\widehat{\nu}'_{2,k}(x,y) = \frac{\widehat{f}_{Z_0}\left(x,\widehat{T}^{-1}(y)\right)}{\widehat{f}_{Z_1}(x,y)} \frac{\widehat{f}_{Y_1|D=1}(y)}{\widehat{f}_{Y_0|D=0}\left(\widehat{T}^{-1}(y)\right)}$;

3.4 using observations in $I_{1,k}$ and $I_{0,k}$, estimate optimal transport map $\widehat{T}$;

Step 4. The estimator is then given by

$$\begin{aligned}\widehat{\tau}_{\mathrm{S2}}^{DML} = \frac{1}{K}\sum_{k=1}^{K}\Bigg\{ & \frac{1}{N_0}\sum_{j=1}^{N_0}\left[\widehat{\boldsymbol{\mu}}_k^*\left(X_{0j}\right) + \widehat{\boldsymbol{h}}_k\left(X_{0j}, \widehat{T}(Y_{0j})\right) - Y_{0j} + \widehat{\varphi}_0\left(Y_{0j}\right) - \frac{1}{N_0}\sum_{j=1}^{N_0}\widehat{\varphi}_0\left(Y_{0j}\right)\right] \\ & + \frac{1}{N_1}\sum_{i=1}^{N_1}\left[\widehat{\nu}_{1,k}\left(X_{1i}\right)\left(Y_{1i}^* - \widehat{\boldsymbol{\mu}}_k^*\left(X_{1i}\right)\right) + \widehat{\varphi}_1\left(Y_{1i}\right) - \frac{1}{N_1}\sum_{i=1}^{N_1}\widehat{\varphi}_1\left(Y_{1i}\right)\right] \\ & + \frac{1}{N_1}\sum_{i=1}^{N_1}\left[\widehat{\nu}'_{2,k}\left(X_{1i}, Y_{1i}\right)\left(W_{1i} - \widehat{\boldsymbol{h}}\left(Z_{1i}\right)\right)\right]\Bigg\}.\end{aligned}$$

2.6 Numerical Results

We carry out two simulation studies to estimate the ATU when there are two different incentives for strategic misreporting behavior: incentives linked to the value of the reported outcome and incentives linked to the rank of the reported outcome.

2.6.1 Scenario 1: Strategic Behavior toward the Value of the Reported Outcome

We focus on the outcome with dimension $d_Y = 1$. The data-generating process (DGP) is described as follows: in the treatment group, a univariate covariate is sampled from the standard normal distribution $\mathcal{N}(0,1)$. In the control group, the univariate covariate is sampled from a normal distribution with mean 0.5: $\mathcal{N}(0.5,1)$. The true outcome in the treatment group is generated as $Y_{1i}^* = X_{1i} + 6 + \varepsilon_i$, where ε_i is sampled from a standard

normal distribution. In the control group, the true outcome is generated as $Y_{0i}^* = X_{0i} + 1 + \varepsilon_i$. The individual treatment effect is 5 under this data-generating process.

In this case, we make the assumption that individuals receive the payoff based on the value of their reported outcome. For $k = \{0, 1\}$, the potential payoff an individual can receive by reporting the potential outcome Y_k is given by

$$V(X, Y_k) = 2XY_k - 0.2Y_k^2.$$

This payoff function allows heterogeneous payoffs for individuals based on their observed characteristics. We assume individuals incur a quadratic cost when they misreport,

$$C(W_k) = \frac{1}{2}W_k^2.$$

An individual seeks to maximize their utility $U = 2XY_k - 0.2Y_k^2 - \frac{1}{2}W_k^2$ by determining the optimal amount of misreporting, for $k = \{0, 1\}$. The optimal amount of misreporting in group k is given by:

$$W_k = 2X - 0.2Y_k. \tag{2.30}$$

To satisfy (2.30), we first generate $Y_{0i} = \frac{2X_{0i}+Y_{0i}^*}{1.2}$ for $i = 1, \ldots, n_0$, yielding $W_{0i} = 2X_{0i} - 0.2Y_{0i}$. Similarly, Y_{1i} is generated via $Y_{1i} = \frac{2X_{1i}+Y_{1i}^*}{1.2}$ for $i = 1, \ldots, n_1$, resulting in $W_{1i} = 2X_{1i} - 0.2Y_{i1}$. We conduct 1,000 Monte Carlo simulations. The sample size for the treatment group is 1,000, and the sample size for the control group is 2,000.

To estimate ATU, we follow the estimation steps in section 2.3.2. We adopt sieve least square estimator $\widehat{\mu}^{*d}(x)$ with quadratic spline basis for $d = 1, \ldots, d_Y$. Specifically, we estimate

$$\widehat{\mu}^{*d}(x) = \sum_{i=1}^{n_1} Y_{1i}^{*d}\, p^{J_{n_1}}(X_{1i})' \left(P_1' P_1\right)^{-1} p^{J_{n_1}}(x),$$

for some integer J_{n_1}. Let $\widetilde{J}_{n_1} = J_{n_1} - 3$,

$$p^{J_{n_1}}(x) = \left(1, x, x^2, (x - t_1)_+^2, \ldots, (x - t_{\widetilde{J}_{n_1}})_+^2\right)',$$

where t_j is the knots for the spline basis, and $(\cdot)_+$ denote the positive part. $P_1 = (p^{J_{n_1}}(X_{11}), ..., \ p^{J_{n_1}}(X_{1n_1}))'$, where $(X_{11}, \ldots, X_{1n_1})$ denotes the observations of covariates in the treatment group. Similarly, the estimator for $h^d(x, y)$ can be written as

$$\widehat{h}^d(z) = \sum_{i=1}^{n_1} W_{1i}^d \ p_z^{J_{n_1}}(Z_{1i})' \ (P_{Z1}' P_{Z1})^{-1} \ p_z^{J_{n_1}}(z), \quad \text{for } d = 1, \ldots, d_Y,$$

where $z = (x, y)$, $Z_{1i} = (X_{1i}, Y_{1i})$, $p_z^{J_{n_1}}(z) = \left(1, z, z^2, (z - t_1)_+^2, \ldots, (z - t_{\widetilde{J}_{n_1}})_+^2\right)'$, $P_{Z1} = (p_Z^{J_{n_1}}(Z_{11}), ..., \ p_Z^{J_{n_1}}(Z_{1n_1}))'$.

Table 2.1 compares the performance of the plugged-in estimator with and without accounting for strategic reporting, i.e., using the estimator in the baseline scenario in Section 2.3.2. The results indicate that ignoring strategic reporting introduces significant bias in the estimator.

	Ignore Strategic Reporting	Correct for Strategic Reporting
Bias	-0.834	-0.000
Variance	0.001	0.002
MSE	0.696	0.002

Table 2.1: Performance of Estimators (Scenario 1)

2.6.2 Scenario 2: Strategic Behavior toward the Rank of the Reported Outcome

In this subsection, we explore a DGP in which individuals strategically report the outcome when there are incentives linked to the rank of the reported outcome. The generation of covariates and the true outcome in both the treatment and control groups is the same as described in subsection 2.6.1. To integrate the strategic behavior due to incentives linked to

the rank of the reported outcome, we modify the equation used in subsection 2.6.1. Rather than use equation (2.30), we use the following equation for the DGP of the reported outcome:

$$W_{1i} = 2X_{1i} + 0.2Y_{1i} \quad \text{for } i = 1, \ldots, n_1; \quad W_{0j} = 2X_{0j} + 0.2T_o(Y_{0j}) \quad \text{for } j = 1, \ldots, n_0, \tag{2.31}$$

where $T_o(Y_{0j}) = Y_{0j} + 3.5$. Under (2.31), we have $E[W_1|X = x, Y_1 = T_o(y), D = 1] = E[W_0|X = x, Y_0 = y, D = 0]$ and thus (2.9) holds. We conduct 1,000 Monte Carlo simulations. The sample size for the treatment group is 1,000, and the sample size for the control group is 2,000.

To estimate T_o, we follow the estimator described in section 2.3.2 case 1. We then follow the same procedure as in section 2.6.1 to estimate $\widehat{\mu}^{*d}(x)$ and $\widehat{h}^d(x)$ for $d = 1, \ldots, d_Y$.

Table 2.2 compares the performance of the estimator with and without accounting for strategic misreporting behavior due to the incentives tied to the rank of the reported outcome. The results again indicate that ignoring strategic misreporting behavior by using the estimator in the baseline scenario introduces significant bias into the estimator. Compared to the results in Table 2.1, the estimator that corrects for strategic reporting exhibits higher variance. This increased variance is attributable to the need for estimating an additional nuisance function: optimal transport map $\widehat{T}(\cdot)$. Incorporating this function into the estimation process introduces an additional source of variability, thereby elevating the overall variance of the estimator.

	Ignore Strategic Reporting	Correct for Strategic Reporting
Bias	0.374	0.001
Variance	0.003	0.004
MSE	0.143	0.004

Table 2.2: Performance of Estimators (Example 2)

2.7 Empirical Example

To illustrate the method developed in previous sections, we consider an empirical example that revisited the treatment effect on criminal behaviors in (Blattman et al., 2016). The study's experimental design involved assigning participants to one of four groups: cash, therapy, both, or neither. The findings, based on self-reported survey data, suggest that the treatment, which involves both cash and therapy, reduces self-reported criminal behavior. However, there is a potential issue that respondents have incentives to misreport their outcomes. To address this concern, a validation sample from both treatment and control groups was gathered, aiming to mitigate the bias associated with the use of self-reported outcomes; see Blattman et al. (2016) for further details.

In our empirical example, we focus on the outcomes that involve sensitive behavior, which include the frequency of instances where respondents have engaged in theft, gambling, marijuana use, instances of sleeping outdoors within the past two weeks, and the total number of the above instances. We study the ATU across three interventions: (1) participants receive a cash incentive of 200 dollars; (2) participants receive an 8-week therapy program that tries to reduce self-destructive beliefs or behaviors and promote positive ones; (3) participants receive a combination of both interventions. The treatment variable, D, serves as an indicator of whether participants receive the specified treatment. The outcome variable Y is a 5-dimensional vector representing the frequency of these sensitive behaviors. The vector of covariates, X, encompasses factors such as age, marital status or cohabitation with a partner, the number of women supported, and the count of children under 15 years of age, etc. For a full list and details regarding the covariates, refer to Table 1 in Blattman et al. (2016).

We first provide the baseline estimator for ATU using the survey dataset with the self-reported outcomes. To compute the baseline estimator, one needs to estimate the function $\boldsymbol{\mu_o}$ as described in (2.11). We adopted a linear sieve estimator which is easy to implement in two steps: Firstly, constructing the sieve basis and; secondly, employing an Ordinary Least Squares (OLS) for estimation. In the first step, we calculate a basis vector $p^{J_{n_1}}(X_i)$,

$i = 1, \ldots, n$ for all the observations, where J_{n_1} is calculated based on n_1. We then perform an OLS estimation of reported outcomes in the treatment group on sieve-basis vectors $p^{J_{n_1}}(X_{1i})$ for $i = 1, \ldots, n_1$ using covariates in the treatment group. The predicted value $\widehat{Y}_{1j}$ was obtained utilizing the sieve basis vector $p^{J_{n_1}}(X_{0j})$ for $j = 1, \ldots, n_0$, which used covariates in the control group. The baseline estimator $\widehat{\tau}_{\mathrm{T},baseline}$ in (2.12), can be expressed using the alternative notation described above as follows:

$$\widehat{\tau}_{\mathrm{T},baseline} = \frac{1}{n_0}\sum_{j=1}^{n_0}\widehat{Y}_{1j} - \frac{1}{n_0}\sum_{j=1}^{n_0} Y_{0j}, \tag{2.32}$$

where Y_{0j} represents the reported outcome in the control group and T denotes the treatment, which falls into one of three categories: therapy and cash, therapy only, and cash only.

Next, we calculate the ATU estimator for Scenario 1, which involves strategic behavior due to incentives linked to the reported outcome's value. To make full use of the survey data, equation (2.13) can be rewritten as:

$$\tau_{o,S1} = E\left(E\left(Y_1|X, D=1\right)|D=0\right) - E\left(E\left(W_1|X, D=1\right)|D=0\right) - E\left(Y_0|D=0\right) + E\left(\boldsymbol{h_o}(X, Y_0)|D=0\right). \tag{2.33}$$

Initially, we construct $\widehat{Y}_{1j}$ following the same steps as previously mentioned in calculating $\widehat{\tau}_{\mathrm{T},baseline}$. Then, we perform an OLS regression of measurement errors W_i on the basis vectors $p^{J_{n_1}}(X_{1i})$ using the treatment group's validation dataset. Due to the small sample size of validation dataset, and to prevent overfitting, a lasso regression on W_i on $p^{J_{n_1}}(X_{1i})$ is first conducted, selecting a subset of elements in the basis vectors.

The results in Section 2.7.1 utilize a lasso hyperparameter of $\alpha = 0.1$, while in Section 2.7.2, the value of the hyperparameter increases to $\alpha = 2$. This increase is driven by the expanded dimension of the basis vector in the latter case, leading to a higher α to mitigate overfitting by selecting a reduced number of basis elements. Nevertheless, testing with various α values, we find that our results are robust across different values of lasso hyperparameter.

We then run an OLS on the selected basis. Subsequently, we construct $\widehat{W}_{1j}$ using the

chosen basis vector in the control group. For constructing the predicted value $\widehat{W}_{0j}$, for each $Z_{1i} := (Y_{1i}, X_{1i})$ we calculate a basis vector $p_Z^{J_{n_1}}(Z_{1i})$. To avoid potential overfitting here as well, a lasso is first applied to the basis vectors, followed by an OLS on the lasso-selected basis. Finally, using the selected basis in the control group, we obtain $\widehat{W}_{0j}$. Using the above notation, equation (2.14) can be alternatively written as:

$$\widehat{\tau}_{\mathrm{T},S1} = \frac{1}{n_0}\sum_{j=1}^{n_0}\widehat{Y}_{1j} - \frac{1}{n_0}\sum_{j=1}^{n_0}\widehat{W}_{1j} - \frac{1}{n_0}\sum_{j=1}^{n_0}Y_{0j} + \frac{1}{n_0}\sum_{j=1}^{n_0}\widehat{W}_{0j}, \tag{2.34}$$

where T denotes the treatment, which falls into one of three categories: therapy and cash, therapy only, and cash only.

Finally, we estimate the ATU for Scenario 2, which focuses on strategic behavior due to incentives linked to the rank of the reported outcome. Initially, we construct $\widehat{Y}_{1j}$ and $\widehat{W}_{1j}$ using the same steps as in the previous scenario. To construct $\widehat{W}_{0j}$, we first need to estimate the optimal transport map (OT) for the reported outcomes in the treatment and control groups. To do so, we follow the strategy outlined in 2.3.2 Case 2. We also provide the OT estimator using the strategy outlined 2.3.2 Case 1 as a robustness check in the appendix.

We start by calculating the sample averages for Stealing, Marijuana, Gambling, and Homelessness outcomes in both the treatment and control groups. This is represented as:

$$\widehat{\mu}_{k,s} = \left(\frac{1}{n_k}\sum_{i=1}^{n_k}Y_{ki,Stealing}, \frac{1}{n_k}\sum_{i=1}^{n_1}Y_{ki,Marijuana}, \frac{1}{n_k}\sum_{i=1}^{n_1}Y_{ki,Gambling}, \frac{1}{n_k}\sum_{i=1}^{n_1}Y_{ki,Homeless}\right)', \quad \text{for } k \in \{0,1\}.$$

We then estimate the variance-covariance matrix. Let

$$Y_{ki,s} := (Y_{ki,Stealing}, Y_{ki,Marijuana}, Y_{ki,Gambling}, Y_{ki,Homeless})' \quad \text{for } k \in \{0,1\},$$

the estimator for the variance-covariance matrix can be written as

$$\widehat{\Sigma}_{k,s} = \frac{1}{n_k}\sum_{i=1}^{n_k}\left(Y_{ki,s} - \widehat{\mu}_{k,s}\right)\left(Y_{ki,s} - \widehat{\mu}_{k,s}\right)' \quad \text{for } k \in \{0,1\}.$$

The OT estimator for these four outcomes is given by:

$$\widehat{T}_s(Y_{0i,s}) = \widehat{\mu}_{1,s} + \widehat{A}\left(Y_{0i,s} - \widehat{\mu}_{0,s}\right).$$

For the composite outcome of all sensitive behaviors (Stealing, Marijuana, Gambling, and Homelessness), we use the estimation strategy from 2.3.2 Case 1:

$$\widehat{T}_{All}\left(Y_{0i,All}\right) = \widehat{F}^{-1}_{Y_{1,All}|D=1}\left(\widehat{F}_{Y_{0,All}|D=0}\left(Y_{0i,All}\right)\right), \text{ for } i = 1, \ldots, n_0,$$

where $\widehat{F}_{Y_{k,All}|D=k}$ denotes the empirical CDF for all sensitive behaviors in both groups. The empirical CDF's specific formulation is detailed in Section 2.3.2. Similar to calculating the estimator in Scenario 1, we use $Z_{1i} := (Y_{1i}, X_{1i})$ to create a basis function $p_Z^{J_{n_1}}(Z_{1i})$. To mitigate potential overfitting, a lasso is initially applied to select the basis vectors, followed by an OLS on the chosen basis. We then employ the transported $\widehat{T}_s(Y_{0j,s})$ and $\widehat{T}_{All}(Y_{0j,All})$, together with X_{0j} to construct the selected basis in the control group, and use the result from previous OLS estimation to predict value $\overline{W}_{0j}$. Using the described notation, the estimator can be written as:

$$\widehat{\tau}_{\mathrm{T},S2} = \frac{1}{n_0}\sum_{j=1}^{n_0}\widehat{Y}_{1j} - \frac{1}{n_0}\sum_{j=1}^{n_0}\widehat{W}_{1j} - \frac{1}{n_0}\sum_{j=1}^{n_0}Y_{0j} + \frac{1}{n_0}\sum_{j=1}^{n_0}\overline{W}_{0j}. \tag{2.35}$$

2.7.1 OLS Results

This subsection presents results obtained by choosing the basis functions $p^{J_{n_1}}(x) = (1, x')'$ and $p_Z^{J_{n_1}}(z) = (1, z')'$. This selection yields results equivalent to applying an OLS estimator. The motivation for displaying OLS results is to compare our proposed estimator's performance with the one reported in Blattman et al. (2016) using OLS. Despite the difference in parameters (ATU vs. ATE), the usage of the above basis function allows for a meaningful comparison.

The results for the different treatments are presented in Tables 2.3, 2.4, and 2.5. Additionally, the ATE estimator from Blattman et al. (2016) is included in these tables. It is

noted that in our analysis, the result for the baseline ATU estimator in general aligns closely with the ATE estimator reported in their study in terms of the effect size and statistical significance.

For the result of the therapy and cash treatment that is shown in Table 2.3, we observed an overestimation of the ATU in Scenario 1, where the corrected ATU shows a smaller effect size. In contrast, the corrected ATU for Scenario 2 demonstrates a bigger effect size, indicating that misreporting behaviors towards the ranking of reported outcomes lead to an underestimation of the effect if only use the reported outcomes in this case.

Comparing these findings with those in Blattman et al. (2016), the corrected ATU for Scenario 2 appears to be consistent with their results. For the composite sensitive activity outcome, the corrected ATU inflates the effect size by -0.132, whereas in Blattman et al. (2016), the corrected ATE inflates it by -0.118. This pattern also shows in the Gambling and Homelessness categories. However, for Stealing and Marijuana categories, the corrected ATE in Blattman et al. (2016) shows no significance, but our corrected ATU in Scenario 2 shows significant bigger effect size. To summarize, our proposed estimator, which uses only half the validation data from the treatment group and omits the control group's validation data, yields results comparable to the corrected ATU in Scenario 2 but with increased statistical power for inference.

For the therapy treatment results displayed in Table 2.4 and the cash treatment results displayed in Table 2.5, the corrected ATU is not significant in both Scenario 1 and Scenario 2. This aligns with the finding in Blattman et al. (2016), where the corrected ATE was found to be insignificant.

2.7.2 Sieve Results

In this subsetion, we show the result for using the quardratic spline basis function for $d-$dimensional covariate $x = (x_1, \ldots, x_d)'$. To be specific, the basis function is given by

$$p^{J_{n_1}}(x) = (1, p^{J_{n_1},d}(x_1), \ldots, p^{J_{n_1},d}(x_d))'$$

		Total	Stealing	Marijuana	Gambling	Homeless
Baseline	ATU	-0.367***	-0.093***	-0.076***	-0.087***	-0.111***
		(0.051)	(0.020)	(0.024)	(0.021)	(0.021)
Scenario 1	Corrected ATU	-0.297***	-0.076	-0.073*	-0.039	-0.100
		(0.089)	(0.037)	(0.035)	(0.061)	(0.047)
	Difference	0.070	0.017	0.003	0.048	0.011
Scenario 2	Corrected ATU	-0.499***	-0.125*	-0.091*	-0.157*	-0.193**
		(0.125)	(0.049)	(0.049)	(0.069)	(0.062)
	Difference	-0.132	-0.032	-0.015	-0.070	-0.082
Results in Blattman et al. (2016)	ATE	-0.398***	-0.105***	-0.069*	-0.099***	-0.125***
		(0.090)	(0.027)	(0.040)	(0.026)	(0.029)
	Corrected ATE	-0.516***	-0.122	-0.048	-0.198**	-0.156*
		(0.196)	(0.077)	(0.074)	(0.096)	(0.083)
	Difference	-0.118	-0.017	0.021	-0.099	-0.031

Table 2.3: Therapy and Cash

Note: The standard errors are presented in parentheses. For the ATU estimator, the standard error calculation is based on a bootstrap method with 1000 repetitions. P values that are less than 0.001, 0.01, and 0.05 are represented by three asterisks, two asterisks, and one asterisk, respectively.

		Total	Stealing	Marijuana	Gambling	Homeless
Baseline	ATU	-0.175**	-0.046*	-0.027	-0.089***	-0.013
		(0.056)	(0.021)	(0.022)	(0.020)	(0.021)
Scenario 1	Corrected ATU	-0.017	0.007	-0.026	-0.012	0.025
		(0.117)	(0.051)	(0.026)	(0.052)	(0.049)
	Difference	0.158	0.053	0.001	0.077	0.038
Scenario 2	Corrected ATU	-0.127	-0.016	-0.040	-0.049	0.005
		(0.122)	(0.059)	(0.041)	(0.060)	(0.055)
	Difference	0.048	0.030	-0.013	0.040	0.018
Results in	ATE	-0.186**	-0.045	-0.022	-0.087***	-0.031
Blattman		(0.092)	(0.028)	(0.040)	(0.026)	(0.031)
et al. (2016)	Corrected ATE	-0.190	-0.041	-0.031	-0.131	0.007
		(0.198)	(0.082)	(0.065)	(0.106)	(0.086)
	Difference	-0.004	0.004	-0.009	-0.044	0.038

Table 2.4: Therapy Only

		Total	Stealing	Marijuana	Gambling	Homeless
Baseline	ATU	-0.082	-0.034	0.000	0.022	-0.070***
		(0.057)	(0.023)	(0.023)	(0.023)	(0.024)
Scenario 1	Corrected ATU	0.015	0.000	-0.010	0.068	-0.057*
		(0.091)	(0.047)	(0.035)	(0.043)	(0.026)
	Difference	0.097	0.034	-0.010	0.046	0.013
Scenario 2	Corrected ATU	0.004	-0.022	-0.011	0.071	-0.069
		(0.093)	(0.064)	(0.051)	(0.052)	(0.039)
	Difference	0.086	0.012	-0.011	0.049	0.001
Results in	ATE	-0.057	-0.031	0.012	0.025	-0.062**
Blattman		(0.095)	(0.029)	(0.040)	(0.029)	(0.031)
et al. (2016)	Corrected ATE	0.121	-0.014	0.058	0.072	0.001
		(0.195)	(0.087)	(0.074)	(0.093)	(0.076)
	Difference	0.178	0.017	0.046	0.047	0.063

Table 2.5: Cash Only

$$p^{J_{n_1},d}(x_d) = \left(x_d, x_d^2, (x_d - t_{x_d,1})_+^2, (x_d - t_{x_d,2})_+^2\right),$$

where $t_{x_d,1} = (x_d^{\max} - x_d^{\min})/3$, $t_{x_d,2} = 2(x_d^{\max} - x_d^{\min})/3$ are two knots, with $x_d^{\max}$ and $x_d^{\min}$ denotes the maximum and minimum value of the $\{x_{di}\}$ for $i = 1, \ldots, n_1$.

The results for the estimator are summarized in Tables 2.6, 2.7, and 2.8. Although the results are not statistically significant, they exhibit a consistent correction pattern. Comparing results from Tables 2.3 and 2.6 for the therapy and cash treatment, the correction — measured by the difference between baseline ATU and corrected ATU — is consistent in both size and direction. A similar pattern is observed for cash-only treatments in Tables 2.5 and 2.8. For therapy treatments, except for the Homeless outcome in Scenario 1, the correction pattern is consistent in Tables 2.4 and 2.7. Overall, the correction remains robust across different estimation strategies (OLS vs. quadratic spline estimator), although the sieve estimator shows less statistical power due to the high dimensional basis.

		Total	Stealing	Marijuana	Gambling	Homeless
Baseline	ATU	-0.499	-0.154	-0.415	0.077	-0.007
		(17.212)	(3.201)	(3.980)	(3.678)	(1.545)
Scenario 1	Corrected ATU	-0.489	-0.127	-0.414	-0.089	-0.001
		(4.405)	(5.737)	(4.358)	(3.117)	(2.386)
	Difference	0.010	0.027	0.001	(0.166)	0.006
Scenario 2	Corrected ATU	-0.686	-0.191	-0.414	-0.034	-0.082
		(11.162)	(4.065)	(6.327)	(12.386)	(1.593)
	Difference	-0.187	-0.037	0.001	-0.111	-0.075

Table 2.6: Therapy and Cash (quadratic spline)

2.8 Conclusion

In conclusion, we study the estimation of the ATU when there is strategic misreporting on the outcome. We found that when incentives are linked to the reported outcome, the patterns

		Total	Stealing	Marijuana	Gambling	Homeless
Baseline	ATU	-0.186	0.131	-0.065	-0.253	0.000
		(3.347)	(2.351)	(4.031)	(4.590)	(1.856)
Scenario 1	Corrected ATU	-0.026	0.198	0.064	-0.171	-0.034
		(0.708)	(0.277)	(0.296)	(0.301)	(0.282)
	Difference	0.160	0.067	0.129	0.082	-0.034
Scenario 2	Corrected ATU	-0.121	0.185	-0.080	-0.214	0.015
		(4.325)	(0.272)	(0.310)	(0.287)	(0.297)
	Difference	0.065	0.054	-0.015	0.039	0.015

Table 2.7: Therapy (quadratic spline)

		Total	Stealing	Marijuana	Gambling	Homeless
Baseline	ATU	-4.347	-1.783	-0.307	-1.520	-0.737
		(3.381)	(9.418)	(5.207)	(3.907)	(2.167)
Scenario 1	Corrected ATU	-4.305	-1.728	-0.319	-1.472	-0.725
		(21.975)	(4.889)	(6.314)	(7.054)	(3.684)
	Difference	0.042	0.055	-0.012	0.048	0.011
Scenario 2	Corrected ATU	-4.320	-1.747	-0.330	-1.467	-0.737
		(56.341)	(4.880)	(7.670)	(5.274)	(4.148)
	Difference	0.027	0.036	-0.023	0.053	0.000

Table 2.8: Cash (quadratic spline)

of misreporting can differ between the treatment and control groups. If the researchers omit this and use the reported outcome to estimate the ATU, they may have a biased estimator. We offer identification in two scenarios where incentives are linked to the reported outcome. We employ tools from optimal transport theory to assist with identification in the scenario where incentives are linked to the rank of the reported outcome. We provide estimators along with their asymptotic results in each of the scenarios. As an extension, we derive Neyman orthogonal moments within the DML framework and present the corresponding DML estimator. Our framework is applied to a self-reported criminal activity dataset, which illustrates the efficacy of the estimators that we propose.

Our work is not perfect and has its limitations. First, researchers must have prior knowledge to choose between estimators in different scenarios. Second, our estimator relies on a validation dataset, which is often criticized for potentially containing its own measurement errors. Third, violating our identification assumptions can lead to model misspecification and introduce bias into the estimator. In future work, we plan to extend the current framework in two ways. First, we aim to augment the framework to allow for partial identification, eliminating the need to choose between different scenarios and relaxing our current identification assumptions. Second, we intend to modify the existing framework to utilize a proxy for information, rather than relying on a validation dataset.

Chapter 3

COVID-19, URBAN TRANSPORTATION AND AIR POLLUTION

3.1 Introduction

Billions of individuals suffer from the collateral health and economic burden of severe air pollution (Ito and Zhang, 2020). In particular, poor air quality harms many societal and economic outcomes, including life expectancy (Ebenstein et al., 2017), infant mortality (Greenstone and Hanna, 2014), labor supply (Hanna and Oliva, 2015), and house demand (Huang and Lanz, 2018). Managing air pollution is more challenging for developing and industrializing economies, such as China and India, especially in urban areas. In 2017, the annual average exposure to fine particulate matter ($PM_{2.5}$) in China and India was around 20 times that of the United States, and 100% of their population was exposed to levels that exceeded the values in the World Health Organization (WHO) guidelines (WHO, 2021; World Bank, 2021).

As transportation systems are among the most significant contributors to urban air pollution (Diamond and Wood, 2020; Jabali et al., 2012), it is crucial to quantify the effects of urban transportation on air pollution. Understanding the environmental cost of increasing traffic volume contributes to managing the trade-off between environmental protection and economic development (Badia et al., 2021). Additionally, quantifying the heterogeneous effects of transportation subsectors (e.g., buses, subways, taxis, and private vehicles) on air pollution signifies because it enables fine-grained sustainable transportation policy making (Rivers et al., 2020; Park et al., 2022).

Quantifying such effects, however, is empirically challenging, as commuting decisions are endogenous to air pollution. High traffic volume increases air pollution, but air pollution, in

return, reduces the intention to go outdoors and traffic volume. Due to such simultaneity, the actual effects of urban transportation on air pollution are difficult to identify under normal conditions. COVID-19 and the resulting pandemic, however, are exogenous and unexpected by all participants in the transportation system. Therefore, the pandemic constitutes a unique negative shock on the transportation side but does not directly cause air pollution (e.g., $PM_{2.5}$, NO_2), enabling us to identify the one-sided effect of traffic on air pollution.

The literature on transportation and air pollution has estimated the effectiveness of green transportation measures, such as promoting electronic vehicles (Avci et al., 2015; Holland et al., 2016), imposing driving restrictions (Davis, 2008), implementing voluntary environmental programs (Scott et al., 2022), and building public transportation infrastructures (Sun et al., 2019). For a number of reasons, we need to advance the identification of the effect of transportation systems on air quality. First, it is argued that these measures are far from extensive or effective (Kelly and Zhu, 2016). The measures related to private vehicles, such as vehicle electrification and driving restrictions, are not as effective as expected in regard to their relationship with urban air pollution (Holland et al., 2016; Davis, 2008). This is possibly because such measures influence only the decisions of a small proportion of the population and do not introduce meaningful variation in the total transportation volume, which hinders researchers from observing significant effects of transportation on air quality. Although measures or events of large-scale public transportation, such as the opening of new metro routes, an increase in the number of buses, and strikes of public transit workers (Sun et al., 2019; Rivers et al., 2020), are more effective (Kelly and Zhu, 2016; Sun et al., 2019), these measures or events still affect behavior in only a few communities near the new routes and within one single city and, thus, are not able to generate a significant shock to traffic volume at a societal level, as compared to the COVID-19 pandemic. Second, these measures are potentially less exogenous than the pandemic, as poor air quality results in additional environmental protection measures, such as technology adoption or the implementation of regulatory measures for private vehicles (Drake and Spinler, 2013; Wang et al., 2013; Akyol and De Koster, 2013; Atasu et al., 2020). Third, the evaluated measures usually

induce meaningful variation in one or two subsystems of the transportation sector, e.g., buses, railways, taxis, private vehicles (Sun et al., 2019), which prevents us from comparing the effects of different subsystems. Further, the reduced demand in one subsystem may shift to other subsystems (Naumov et al., 2020), and thus the estimation may suffer from omitted variable biases. Fourth, these studies mainly provide estimates of the effects of transportation based on supply-side variations (e.g., change in the number of buses, a public transit strike), without investigating the substantial city- and time-varying variations on the demand side (e.g., change of the number of passengers) (Anderson, 2014). Unobserved demand-side response to the supply-side change may confound the identification. To sum up and to advance our understanding of the causal relationship between transportation and air pollution, we need a comprehensive investigation of the effect of transportation on air quality across different cities and transportation subsystems, taking advantage of the societal and exogenous shock, i.e., the pandemic, and making use of multi-source traffic-volume data which capturing the demand-side response to the shock.

We also note that the extant research has utilized the exogenous shock of COVID-19 to study the short-term effects of a lockdown on air pollution (e.g., Bauwens et al., 2020; Diamond and Wood, 2020; Almond et al., 2021; Wang and Yang, 2021). These effects, however, are argued to be negligible and of limited implication in the long run (Forster et al., 2020), The public discussion on air quality improvement by COVID-19 pandemic in social media attaches importance to long-term policy implications (Brimblecombe and Lai, 2020), because the lockdown will likely not be repeated in the near future.In this way, our study adds to this literature by demonstrating the mechanism of how the COVID-19 pandemic affected traffic or industrial production and then air quality. We argue that using COVID-19-related variables as instruments to quantify the causal relationships between transportation and air pollution is more economically meaningful, as it can offer richer implications for academia and policymakers in the long run.[1]

[1]Notably, although a few recent studies focus on the relationship between the pandemic and contributors to air pollution, such as transportation, industry, and telework (Badia et al., 2021; Tian et al., 2021;

This chapter aims to investigate the following research questions. First, what is the causal effect of urban transportation on air pollution? Second, to what extent is there endogeneity in the relationship between urban transportation and air pollution? Third, to what extent do the effects and endogeneity differ across heterogeneous subsystems of the transportation system, i.e., the public transport (railways, buses, and taxis), and the private transport subsectors? We collect data from each of 36 central cities (see Appendix C.1 for a full list of the cities) of China from January 2019 to April 2020 on (1) air pollution data from a public institution; (2) monthly passenger volume data (by taxi, bus, and railway), based on government disclosure; (3) monthly congestion data (as a proxy of private transportation, as we detail later) from a mobile map application; (4) COVID-19 infection data (number of confirmed, suspected, recovered, and deceased cases), based on government disclosure; (5) number of queries of COVID-19-related keywords to an online search engine; (6) green energy bus penetration rate in 2019; (7) control variables, including weather conditions and industrial production (measured by the gross domestic product [GDP] of the industry sector of the economy).

We use two sets of instruments for the passenger volume. The more severe the infection in a city, the lower the intention to travel. The infection is not directly linked to the air pollutants. As such, the city-level COVID-19 infection variables are ideal for instrumenting travel decisions. The second set of instruments is the number of COVID-19-related queries to the largest Chinese online search engine (*Baidu.com*). The number of queries for keywords such as "COVID-19" and "quarantine" can serve as a proxy of pandemic awareness and travel intention, correlating to travel decisions, but will not directly cause air pollution. To further ensure exclusion restriction, we adopt COVID-19-related variables in the previous month as instruments.

We leverage a two-stage ridge regression with city-month fixed-effects to identify the effects of transportation on air quality. Further, we implement a semi-parametric double

---

Heintzelman et al., 2021), they contain only correlational analyses without quantification of the causal effects.

machine learning (DML) model with the random forest algorithm to reduce model dependence. We compare the estimates of the ordinary, two-stage, and DML ridge regression models to understand the potential endogeneity. We find that, first, the DML ridge estimate of private transportation is much larger than the panel ridge estimate, while the DML ridge estimate of public transportation is only slightly larger than the panel estimate. These results demonstrate that, without addressing the endogeneity issue in the observational data, the effect of private vehicles on air pollution is likely to be underestimated significantly. The use of COVID-19-related instruments provides a solution to such empirical difficulties. The underestimation shows that air pollution reduces travel. Second, we find that the slight underestimation of the effect of public transportation is related to an underestimation of the effects of the bus and rail passenger volume and an overestimation of the effects of the taxi passenger volume. The underestimation and overestimation indicate that air pollution decreases the number of people who travel by bus and rail, but increases the number of people who travel by taxi, suggesting a demand-shifting effect of air pollution.

After addressing the issue of endogeneity, our models yield additional findings. We confirm that rail transportation is the most environmentally efficient public transportation mode (in regard to air pollution per capita). Further, an additional analysis shows that the adoption of new-energy buses effectively mitigates air pollution generated by bus transportation. Last, the estimation of the impact of the transportation sector on six primary pollutants shows that NO_2 and PM_{10} are affected by the transportation sector most, followed by $PM_{2.5}$ and CO, while SO_2 is less affected, and O_3 is even reduced by the transportation sector.

This study provides four primary contributions to the literature. First, we provide the first estimates of the causal effects of urban transportation on air pollution at the level of traffic volume. Rivers et al. (2020) emphasizes the importance of traffic-volume-level estimates obtained from demand-side variations in designing sustainable transportation policies. Still, these estimates are lacking for major urban transportation subsectors, likely due to (1) an absence of data on passenger volume of public transportation and (2) difficulty in measuring the traffic volume of private vehicles. Answering the call of Brandt and Dlugosch (2021),

Ketter et al. (2022) and He et al. (2022), this study leverages unprecedented availability of operational data enabled by mobile devices and the Internet of Things to obtain these estimates for improving transportation sustainability. The demand-side estimates also mitigate the generalizability issue in the previous estimation with the supply-side shocks: (1) The result for one measure (e.g., introducing a new rail) can not be easily generalized to estimate the effect of another measure (e.g., imposing driving restrictions), which hence hinders the ex-ante evaluation of a new policy; (2) A policy is usually evaluated in a specific context (i.e., city, state) (Rodrik, 2008). The effectiveness of a transportation policy would differ in other contexts, and therefore, context-specific investigations are needed. However, we can use our approach in combination with cross-city demand-level data to predict the size of the effect ex-ante if the impact of the new policy on demand for transportation sectors can be correctly predicted. Therefore, our traffic-volume-level estimates provide more generalizable references for sustainable transportation planning and evaluation.

Second, this study reveals some passenger behavioral responses to air pollution. The tested endogeneity suggests that citizens' "passive strategic behaviors" to air pollution, i.e., reducing traveling and migrating from mass transportation to taxis, dominate the previously identified "proactive strategic behaviors," such as choosing an environmentally efficient transportation mode (Mir et al., 2016; Culiberg et al., 2022; Flores and Jansson, 2021). Our findings suggest that health concerns, on average, dominate the motivations of environmental protection, similar to Jabali et al. (2012)'s finding that convenience dominates environmental motivations in electrical vehicles adoption. Transportation departments should take measures to promote eco-friendly transportation modes. These measures include providing incentives to commuters to adopt mass transportation through mobile apps,[2] improving the air quality inside subway cars and buses, and shortening the wait time and walking distance of mass transportation passengers.

Third, our findings provide important policy implications for ongoing discussions in energy

[2]Chinese mobile map applications, such as map.Baidu.com and amap.com, have started to promote green transportation among commuters in their platforms.

and environmental regulation in developing countries. The results on the heterogeneity of the transportation subsectors and the primary pollutants are particularly informative. The adoption of new-energy buses is shown to be effective in mitigating air pollution. Rail transportation is shown to generate the least air pollution per capita among three public transportation modes (statistically insignificant effect on air pollution), although building rail transportation requires a large investment and feasible geological conditions. Therefore, bus transportation as the second most environmentally efficient transportation is also preferable. For the primary pollutants, we expect that sustainable transportation operation policies would be most effective in reducing the concentration of NO_2 and PM_{10}. Additionally, the finding of the negative effect of private transportation on the concentrations of O_3 calls for future exploration of the underlying mechanism. This also shows the potential trade-off between O_3 and other pollutants when implementing green transportation policies. Moreover, we evaluate some green transportation policies in Section 3.6. We find that the Zhengzhou driving restriction implemented in December 2020 effectively reduces air pollution. We also find that expanding the configuration of public transportation in Chinese cities improves air quality. An analysis of implied mortality by transportation shows that China's air pollution-induced health problem is still worse than the world average.

Fourth, this work also contributes to the stream of literature on the COVID-19 pandemic and air quality and provides implications for a critical question in recent literature, which is to what extent the impact of the pandemic on air quality would sustain when the pandemic eases off or ends. The validation of the spread of COVID-19 as an instrument demonstrates the causal mechanism that the COVID-19 pandemic reduces transportation and, in turn, reduces air pollution. With our estimates and data on the change in transportation and air quality after the pandemic, our work provides implications for evaluating the pandemic's long-term effect on air quality (Forster et al., 2020) and establishing the causal structure among the pandemic, transportation, and air quality.

3.2 Data Collection

First, we collect air quality data from the city-level monthly air quality reports published by China National Environmental Monitoring Centre (CNEMC) (CNEMC, 2020). These reports contain six primary pollutants (i.e., CO, NO_2, O_3, $PM_{2.5}$, PM_{10}, and SO_2) and the synthesized air quality index of the main cities in China, collected since 2014. The concentration of CO is measured in milligrams per cubic meter (mg/m^3), and the concentration of the other pollutants is measured in micrograms per cubic meter ($\mu g/m^3$). The synthesized air quality index is a sum of the index of all six primary pollutants, for which the index of a pollutant is defined as the ratio of the pollutant's concentration to the baseline concentration level provided by China's National Ambient Air Quality Standard (NAAQS) (CNEMC, 2020). We collected air quality data from 36 central cities during our observational period from January 2019 to April 2020 and had balanced panel data for 576 samples.

Second, we obtain passenger volume data of public transportation in 36 cities from the website of the Ministry of Transport of the People's Republic of China (MOT) (MOT, 2021). The data include the monthly number of passengers for four urban public transportation modes: buses, railways, taxis, and ferries. Ferry transportation is not considered in this study because for the nine cities that have ferry transportation, the monthly average passenger volume is only 77.604 (in 10,000), dramatically less than that of the other transportation modes (See Table 3.1). Additionally, we collect the data on a ratio of the number of new-energy (electricity or hybrid)-powered buses to the total number of buses in a city in 2019, released by Chinese Academy of Transportation Sciences (2020).

Third, we collect data on the congestion index and the driving speed during peak traffic times from urban transportation reports published by *Map.Baidu.com* (Baidu Map, 2021), which has an approximate 33% share of China's web-mapping market. The congestion index is calculated as the ratio of total drive time to the duration of normal driving at peak traffic hours (07:00–09:00 and 17:00–19:00 during work days, adjusted by time zone in Urumqi and Lhasa) (Baidu Map, 2021). We utilize the congestion index and the driving speed during

Table 3.1: Variables and Summary Statistics

Variables	Description	Mean	Std.	Min	Max
The concentration of air pollutants					
$CO_{i,t}$	CO measured in mg/m^3	1.182	0.528	0.400	4.400
$NO2_{i,t}$	NO_2 in μg/m^3	35.377	13.045	5.000	82.000
$O3_{i,t}$	O_3 in μg/m^3	124.792	44.163	36.000	259.000
$PM25_{i,t}$	$PM_{2.5}$ in μg/m^3	40.056	24.587	7.000	164.000
$PM10_{i,t}$	PM_{10} in μg/m^3	66.757	31.309	19.000	232.000
$SO2_{i,t}$	SO_2 in μg/m^3	10.229	6.603	2.000	57.000
$Synindex_{i,t}$	The synthesized index of air pollutants	4.228	1.473	1.680	11.600
The volume of transportation					
$PublicVol_{i,t}$	The number of passenger visits travelling by bus, railway, taxi and ferry during a month in 10,000	13,336.284	14,010.568	0*	66,504
$BusVol_{i,t}$	The number of passenger visits travelling by bus during a month in 10,000	6,634.389	5,548.669	0*	29,998
$RailVol_{i,t}$	The number of passenger visits travelling by railway during a month in 10,000	4,614.168	7,986.854	0*	36,578
$TaxiVol_{i,t}$	The number of passenger visits travelling by taxi during a month in 10,000	1,977.240	1,525.445	0*	8,559
$CongIndex_{i,t}$	The ratio of the total travel time to the duration of normal driving	1.605	0.250	0.976	2.617
$CongSpeed_{i,t}$	The driving speed during traffic peak times in km/h	30.986	5.549	18.885	50.999
$NewEnergyBus_i$	The ratio of the number of electric or hybrid buses to the number of all buses in a city in 2019	0.472	0.227	0.101	0.978
COVID-19 Infection					
$Con_{i,t}$	The number of confirmed cases	291.566	3,576.461	0	50,333
$Sus_{i,t}$	The number of suspected cases	1.321	15.838	0	300
$Rec_{i,t}$	The number of recovered cases	214.071	2,817.175	0	46,464
$Dec_{i,t}$	The number of deceased cases	15.431	212.868	0	3,869
COVID-19-related queries to online search engine					
$COVIDTerm_{i,t}$	The number of queries regarding the names of COVID-19	389.519	979.167	0	7,437
$SpecializedRemedy_{i,t}$	The number of queries regarding exclusive remedy of COVID-19	326.969	839.278	0	7,165
$GenericRemedy_{i,t}$	The number of queries regarding non-exclusive remedy of COVID-19	941.554	977.682	14	7,944
$GenericTerm_{i,t}$	The number of queries regarding the generic terms for COVID pandemic and terms for other pandemics which are similar to COVID-19	3,444.832	7,194.674	81	80,040
Control variables for weather and industrial production					
$AirTem_{i,t}$	The degree Celsius of air temperature	136.320	103.547	-175.014	315.956
$DewPoint_{i,t}$	The degree Celsius of dew point, measuring humidity	58.433	118.672	-219.358	261.581
$WindSpeed_{i,t}$	The gap between airspeed and ground speed in 100 meters per second	0.231	0.151	-0.297	0.600
$Precip_{i,t}$	The ratio of rainy (snowy) days to total days	0.328	0.209	0.000	0.935
$Production_{i,t}$	GDP of the industry sector in 100 billion RMB	0.311	0.308	0.007	2.017

Note: i = city, t = month. The minimum values of public transportation ($PublicVol_{i,t}$, $BusVol_{i,t}$, $RailVol_{i,t}$, and $TaxiVol_{i,t}$), that is, zeros with an asterisk (*), pertain only to Wuhan in February, as public transportation in Wuhan was shut down from January 23, 2020 to March 25, 2020.

peak hours to represent the traffic volume of private vehicles, as the congestion variation in a specific city is caused mainly by the traffic volume of private vehicles (Li et al., 2022). The buses in China's cities run mainly in exclusive bus lanes, the length of which reaches 14,951.7 kilometers, as reported by MOT (2020). Taxis' contribution to the traffic congestion is also minor, as the number of taxis (about 1.4 million) is dramatically less than that of private vehicles (360 million) (MOT, 2020).

Fourth, we collect the data on COVID-19 infections that are available in the COVID-19/2019-nCoV Time Series Infection Data Warehouse on GitHub (GitHub, 2020). The data are drawn from the *DXY Global Pandemic Real-time Report*. DXY is a Chinese online healthcare company that collects real-time COVID-19 infection data from WHO, the Centers for Disease Control and Prevention (CDC), and local media reports. The reports contain the number of confirmed, suspected, recovered, and deceased cases worldwide. We extracted the monthly data of the 36 cities of interest during our observational period.

Fifth, we collect data on the queries of COVID-19-related keywords to the online search engine. The largest search engine platform in China, *Baidu.com*, publishes the number of queries of a keyword by its users. The number of queries is referred to as the *Baidu Index*, which resembles *Google Trends*. The search engine queries collected by *Google Trends* were found highly correlated to those of influenza epidemics, as people rely on search engines to acquire information about disease prevention and medical solutions (Ginsberg et al., 2009; Ru et al., 2021). Therefore the number of disease-related queries is a good indicator of people's awareness of and attention to a pandemic and, in our context, predicts people's willingness to travel. Following such research, we construct a list of 26 keywords that fall into four categories: (1) exclusively COVID-19-related terms, e.g., "COVID," "novel coronavirus", (2) exclusive COVID-19 remedies, e.g., "Remdesivir," "COVID-19 vaccine", (3) generic COVID-19-related terms, "pandemic," "pneumonia", and (4) generic COVID-19 remedies, e.g., "face mask," "sanitizer," (see the Appendix C.2 for a full list of the keywords). We collect the number of queries of the 26 keywords for each city and month of interest from the *Baidu Index*. In total, we have 14,976 city, month, and keyword combinations.

Sixth, we download data on weather conditions from the Integrated Surface Hourly data published by China's National Climatic Data Center (CMA, 2020). There are more than 400 stations in China that detect and report city-level weather condition data every 3 hours. We retrieve data on temperature, humidity, and wind speed for the 36 cities in our sample during our observational period. We then average all of these variables at the city-month level. Further, we collect precipitation data from *tianqihoubao.com*, a Chinese public weather condition database. We measure the precipitation with a ratio of rainy (and snowy) days to the total days in a month for each city.

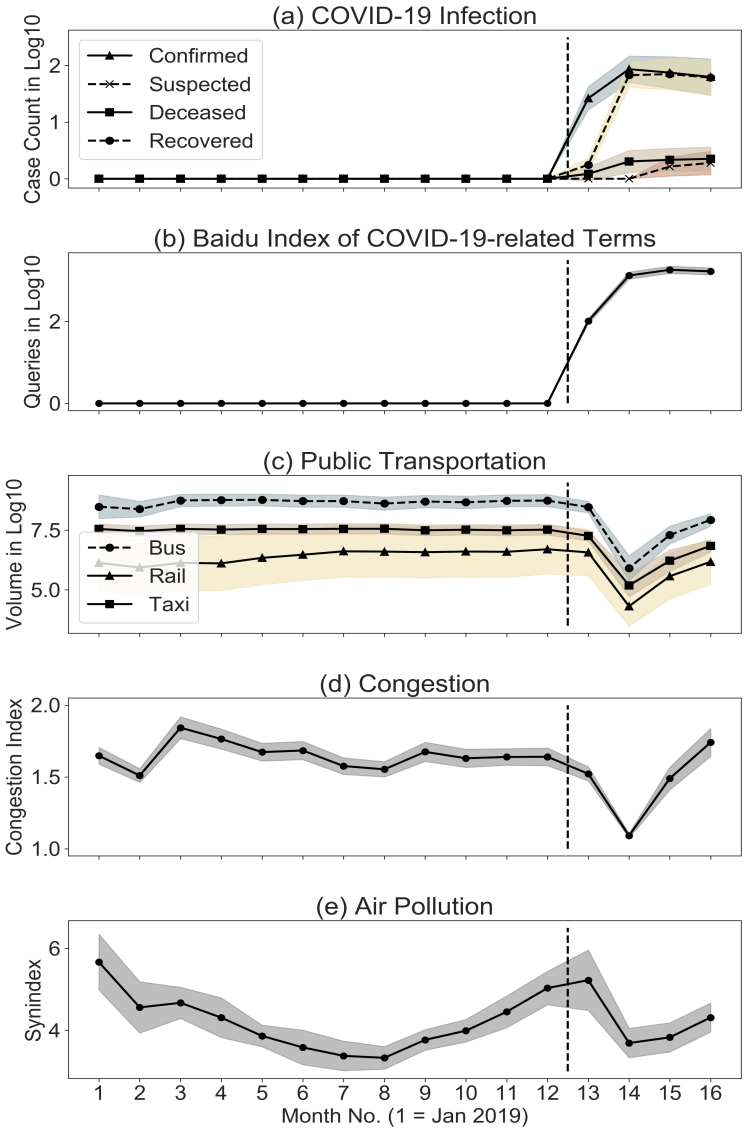
Finally, to measure the intensity of industrial production in the 36 cities, we collect the GDP of the economy's industry sector from each city's Bureau of Statistics. The transportation system and industrial production are two major sources of urban air pollution (Diamond and Wood, 2020). Therefore, it is necessary to control industrial production. The higher the GDP of the industry sector, the more intense the industrial production. We present all the variables with their definitions and summary statistics in Table 3.1.

3.3 Descriptive Analysis

We visualize the variation of our main variables in Figure 3.1. We denote Month 1 as January 2019. There are a total of 16 months in our observational period. In December 2019, 41 confirmed cases of COVID-19 emerged in Wuhan, Hubei, China (Huang et al., 2020). According to a WHO report, symptom onset of the 41 confirmed cases ranged from December 8, 2019 to January 2, 2020 (WHO, 2020). Therefore, the timepoint of the outbreak of COVID-19 in China is between December 2019 and January 2020. In Figure 3.1, we use vertical dashed lines to illustrate the timepoint. Figure 3.1(a) shows the log-transformed numbers of confirmed, suspected, recovered, and deceased cases. The figure shows that the number of confirmed cases increased sharply during the first two months after the virus outbreak . After that, the number started to decrease, indicating that the spread of the virus started to come under control.

Figure 3.1(b) shows the log-transformed number of queries of COVID-19-related terms to

Figure 3.1: Means of Main Variables (with 95% Confidence Intervals)



the online search engine *Baidu.com*. Along with the increase in confirmed cases, searching for "novel virus" online intensified to thousands per month per city, indicating the rapid rise of public awareness of the disease. In addition, the confidence intervals of the queries are much smaller than those of the confirmed cases. This indicates that, although the severity of the epidemic varied across cities, citizens in different cities paid similarly close attention to it. This suggests that the shock of the epidemic was effective across cities.

Figure 3.1(c) shows the log-transformed number of passengers who traveled by bus, rail, and taxi. Before the epidemic, the volume was relatively stable. After the outbreak of the epidemic, the number of passengers dropped sharply. For example, in November 2019, before the epidemic, the average number of passengers who traveled by bus was 542.3 million, whereas in February 2020, two months after the outbreak of the epidemic, when COVID-19 infection data peaked, the number dropped to the lowest level, 0.8 million, which means that only 0.15% of the traffic remained. After the end of February, as the virus came gradually under control and the number of confirmed cases started to decrease, the number of bus passengers resumed to 85.5 million by April. The patterns of railway and taxi passengers are similar. The large variation in the transportation systems provides an ideal identification source of effects on air quality. Further, we note that the transportation pattern is closely related to the COVID-19 infection data and COVID-19 online search data. This suggests that the severity and awareness of the epidemic shaped millions of traveling decisions. Therefore, as instruments, the COVID-19 infection and online query variables are highly relevant to the transportation volume. Figure 3.1(d) shows a very similar pattern in the congestion index.

Figure 3.1(e) shows the mean of the synthesized index of air pollutants ($Synindex_{i,t}$). On the one hand, we find that the pattern of air pollution levels after the outbreak of the epidemic resembles the pattern of transportation: The air pollution level reached its bottom in February and increased afterward. This shows a close relationship between transportation and air pollution. On the other hand, we find a seasonality of air pollution before the epidemic: In 2019, the level of air pollution decreased from January to August and then increased from September to December. This is likely because, in summer, the weather conditions,

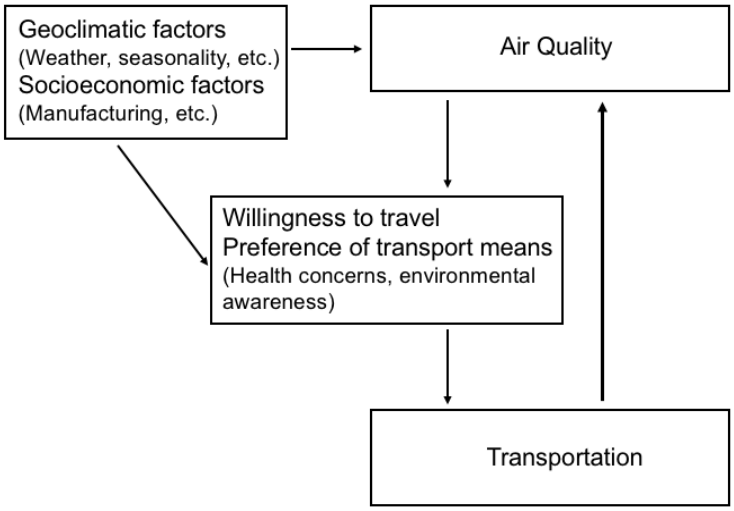
including the precipitation and wind speed, make it easier for air pollutants to be dismissed (Sun et al., 2019), whereas, in winter, more fossil fuels are consumed to increase the supply of heating, increasing the concentration of air pollutants. The seasonality of air pollution is a potential confounding factor because, from January to February 2019, there is a sharp decrease in the level of air pollution, which resembles the decrease from January to February 2020. The air pollution level decreased, however, in April 2019, whereas it increased in April 2020. In other words, the seasonality of air pollution cannot explain the rebounding of air pollution after March 2020. Instead, the rebounding resembles the change in transportation. Thus, we deduce that, although there is a natural trend of air pollution, transportation also plays an important role in shaping the pattern of air pollution. To separate the effect of transportation, we use month fixed-effects to control for the natural trend of air pollution.

3.4 Methods and Results

In the empirical analysis of causal effects of transportation on air quality, there are certain major specification challenges. First, the potential multicollinearity among the transportation subsectors will make estimators based on ordinary least squares (OLS) unstable and unreliable (Judge et al., 1988). Developed by Hoerl and Kennard (1970a,b), ridge regression modifies the OLS procedure by adding a small, positive increment into the diagonal of the ill-conditioned matrix, trading a small amount of bias in the coefficient estimates for a substantial reduction in coefficient sampling variance. As the ridge estimator is a consistent estimator, the small amount of bias tends to diminish when using large-scale observational data. The ridge estimator is often used to address the multicollinearity (Judge et al., 1988). Thus, we use ridge regression in this study. Second, the issue of endogeneity may hinder the identification of the causal effect between transportation and air quality.

In Figure 3.2, we present a customized conceptual model of the relationship between transportation and air quality to explain the potential sources of endogeneity. The volume of transportation is determined mainly by the willingness to travel and selection of transportation means. Geoclimatic and socioeconomic factors, such as the weather and industrial production,

Figure 3.2: Conceptual Model of the Relationship between Transportation and Air Quality



influence willingness to travel and selection of transportation means and, in turn, the volume of transportation and its subsectors. These geoclimatic and socioeconomic factors also can have an impact on air quality (Sun et al., 2019). Thus, we control for these factors. There are also some unobserved geoclimatic and socioeconomic factors, such as climate patterns and the city's industrial structure, that simultaneously affect transportation and air quality. Therefore, one source of endogeneity is the omitted variable bias. The time-invariant omitted geoclimatic and socioeconomic variables can be controlled by city fixed-effects (Wooldridge, 2010). To further control for time-variant unobserved variables, such as seasonality and holiday effects, we also control for month fixed-effects. Hence, a two-way fixed-effects ridge regression (fixed-effects RR) is the first method utilized.

The causal inference of transportation's effect on air quality also may be hindered by simultaneity biases (Wooldridge, 2010). Urban residents in China have become accustomed to checking the air quality index daily due to the serious air pollution. Poor air quality may reduce some individuals' willingness to travel due to health concerns and, in turn, transportation volume. Poor air quality also can cause some individuals to transfer to eco-friendly transportation modes, such as walking and bicycling, reducing the traffic volume of

vehicles that produce emissions. These reverse causalities may result in an underestimation of the effect of transportation on air quality. Moreover, conditional on these simultaneity biases between general transportation and air quality, the estimation of the effects of transportation subsectors may further be contaminated by the impact of air pollution on the selection of means of transportation.

Health concerns triggered by air pollution may decrease travelers' preference for mass transport, e.g., buses, railways, and increase the usage of taxis and private vehicles, as the indoor air pollution in mass transit systems worsens when the outdoor air quality index is of concern (Kadiyala and Kumar, 2012). Commuters also may tend to take taxis or drive private vehicles to minimize their time outdoors. In contrast, the environmental awareness triggered by air pollution may increase commuters' preference for mass transportation and decrease their preference for taxis and private vehicles. Further, the government provides incentives for citizens to travel by mass transportation to improve air quality. Therefore, these simultaneity biases may result in an underestimation or overestimation of the effects of different transportation modes, which we do not know ex ante. To address these simultaneity biases, instrument variable techniques are utilized, with a two-way fixed-effects specification maintained (Angrist and Krueger, 2001). As such, a fixed-effects two-stage ridge regression (two-stage RR) is the second method utilized.

The two-stage RR still faces the challenge of non-linearity in accurate learning of the causal effect. In particular, the linearity assumption of the relationships between air pollution, transportation, and controls in the two-stage RR may not hold. Geoclimatic and socioeconomic factors can have non-linear and interactive effects in predicting the dependent and independent variables. This motivates us to adopt a partial linear model in which the effects of controls are modeled in a non-parametric way (Robinson, 1988). Modern econometric research adopts machine-learning methods to provide flexible estimators of the complex and non-linear effects of controls (Athey and Wager, 2021; Farrell et al., 2021). Machine learning is especially useful in our context because it is difficult to determine a non-linear functional form for the effects of geoclimatic and socioeconomic factors. It has been shown, however, that naively fitting a

machine-learning model into a partial linear regression setting can lead to an estimator that is highly biased and is not root-N consistent (Chernozhukov et al., 2018). The DML method has been proposed to improve the accuracy and efficiency of the non-parametric causal inference when machine-learning methods are used to learn the function of controls (Chernozhukov et al., 2018). Therefore, a DML model combined with two-stage ridge regression (DML RR) is the third method utilized.

3.4.1 Fixed-effects Ridge Regression

We estimate a two-way fixed-effects model that can be expressed by the following equation:

$$\ln y_{i,t} = \beta \ln Vol_{i,t} + \gamma x_{i,t} + c_i + \tau_t + \epsilon_{i,t}, \tag{3.1}$$

where i is the city, t is the month, $y_{i,t}$ denotes one of the air pollution indexes, and $Vol_{i,t}$ is a vector of measures of transportation. Log functional form is adopted for both dependent variables $y_{i,t}$ and independent variables $Vol_{i,t}$. Hence, the coefficients of interest β are elasticities, i.e., the percentage of change in the air pollution index when there is a one-percent increase in the passenger volume of transportation or in the congestion index. In addition, $x_{i,t}$ is a vector of control variables; c_i is fixed city effects; τ_t is fixed month effects; and $\epsilon_{i,t}$ is an idiosyncratic error. Included in $y_{i,t}$ are the synthesized air quality index ($Synindex_{i,t}$) and six primary pollutants index (i.e., $CO_{i,t}$, $NO2_{i,t}$, $O3_{i,t}$, $PM25_{i,t}$, $PM10_{i,t}$, and $SO2_{i,t}$). Included in $Vol_{i,t}$ are the total volume of passengers by public transportation ($PublicVol_{i,t}$), the congestion index ($CongIndex_{i,t}$), and the volume of passengers by bus, rail, and taxi ($BusVol_{i,t}$, $RailVol_{i,t}$, and $TaxiVol_{i,t}$). The weather and manufacturing controls and a constant are included in $x_{i,t}$. Then, instead of using the ordinary panel model estimator, we estimate this model by the ridge regression estimator.

3.4.2 Two-Stage Ridge Regression

We then estimate our panel data specification with instrumental variable techniques, i.e., the two-stage RR, to address simultaneity biases. We use the spread of COVID-19 to construct instruments, arguing that it shifts the volume of transportation and can influence air quality only through transportation, conditional on the weather and manufacturing controls. Before the spread of COVID-19, air quality influenced people's willingness to travel and, in turn, the volume of transportation, resulting in reverse causality. Travel decisions were changed by the spread of COVID-19, either involuntarily or voluntarily, showing relevancy. During the pandemic in January and February 2020, local governments restricted trips within and between cities. At the post-pandemic stage, starting at the end of February, traveling restrictions were gradually removed, but voluntary travel decisions were still affected by concerns about the infectious disease. Therefore, the spread of COVID-19 highly impacts transportation volume and invalidates air pollution's effect on willingness to travel or selection of environmental transportation modes, blocking reverse causality. Further, the outbreak of COVID-19 is an exogenous shock and does not affect air quality directly. For reverse causality in estimating the impact of transportation subsectors on air quality, the spread of COVID-19 significantly reduces individuals' adoption of mass transport that results from environmental awareness or abandonment of mass transport due to health concerns, triggered by air pollution.

Exclusion Restriction of the Pandemic of COVID-19 as Instrument Variables: The exclusion conditions of the instrumental variables are further established with the argument that the pandemic's correlation with air quality occurs largely through transportation, with only a minor effect through industry, and little through other sectors. Emissions from transportation, industrial, residential, power, and agriculture sectors contribute to air pollution (Wang et al., 2020). The pandemic had a minor effect on industrial activities. Many manufacturing activities were seasonally halted during the Spring Festival, coinciding with the pandemic's outbreak in 2020. Some manufacturing subsectors that emit the greatest

amount of air pollutants, such as mining and gas production, are fundamental industries and ran as usual (Diamond and Wood, 2020). Some manufacturing subsectors, such as car and home appliance manufacturing, may have been halted for a short amount of time due to the pandemic but were quickly resilient (Yang et al., 2020; Chen et al., 2020). Therefore, the effect of the short halting of these subsectors is not likely to confound the effect of transportation on air pollution. Overall, research has concluded that manufacturing decreased by only about 10% to 20% during the pandemic (Diamond and Wood, 2020). Thus, controlling for manufacturing activities, industrial activities are not likely to invalidate the exclusion condition.

Further, the change of residential activities by the pandemic did not largely affect air pollution. Pandemic-related social distancing resulted mainly in an increase in two emission-related residential activities, cooking and heating. With 97.3% of the urban population in China having access to gas by 2019, cooking in Chinese central cities relies mostly on gas (National Bureau of Statistics, 2021), which is green energy that emits carbon dioxide but little of the six pollutants of interest in our study. Heating production for buildings in northern cities in China is based on a centralized heating system, was ensured during the pandemic, and fluctuates along with the temperature regardless of the pandemic (Chen et al., 2020). Coal burning for heating in rural areas around a city also can be controlled by the temperature variable. Third, power generation was ensured as a matter of priority and was largely unaffected by the pandemic (Diamond and Wood, 2020; Kroll et al., 2020). Fourth, agriculture was largely unaffected during the pandemic, and only a few rural areas around Wuhan suffered from a delay in the sowing of seeds (Daily, 2020; Kroll et al., 2020). In contrast to the above four sectors, which were little affected by the pandemic, transportation was largely affected and decreased by about 80% to 90% (Diamond and Wood, 2020; Kroll et al., 2020; Wang et al., 2020). Researchers also have found that air pollution improvement during the pandemic is highly correlated with reduced traffic and merely slightly correlated with reduced manufacturing (Diamond and Wood, 2020; Kroll et al., 2020; Wang et al., 2020). We explicitly control for manufacturing through secondary industry GDP and for

seasonality with month fixed-effects in our main model and conduct a robustness check with manufacturing instrumented. Therefore, with seasonality, holidays, meteorology, and manufacturing controlled for, we argue that it is reasonable to assume that the pandemic influences air quality only through transportation and serves as a valid instrument for transportation.

Apart from controlling alternative paths between COVID-19 and air quality in manufacturing and other residential activities, considering potential reverse effects of air quality on COVID-19 can further ensure exclusion restriction of the instruments. As found by He et al. (2020), air pollution leads to an increase in the growth rate of COVID-19 with a short-term delay. To block this simultaneity between COVID-19 and air pollution, we use COVID-19-related variables from the previous month as instruments. Conditional independence is satisfied because the previous month's COVID-19-related variables are unlikely to be affected by the current month's air pollution. Using COVID-19-related variables in the last month as instruments for transportation can also further block alternative processes in manufacturing and residential activities between COVID-19 and air quality.

Instruments for Transportation: The first set of instruments is the number of cases of infection of COVID-19, measured by the number of confirmed and suspected cases. The number of infections depicts the outbreak and the seriousness of COVID-19 directly. The number of recovered and deceased cases are omitted in our model due to their high multicollinearity with the number of confirmed and suspected cases. The second set of instruments is three variables that aggregate the Baidu Index, an aggregate measurement of individuals' search queries that reflect their awareness of and concerns about the COVID-19 pandemic (Ginsberg et al., 2009), into three groups of keywords related to the pandemic: exclusive terms for the COVID-19 pandemic, generic terms for the pandemic, and generic COVID-19 remedy-related keywords. The exclusive COVID-19 remedy-related keywords are omitted in our model due to their high multicollinearity with the existing instrument variables. Thus, there are five variables in the instrument set.

To apply the two-stage RR approach, we build the following fixed-effects two-stage model,

following Angrist and Krueger (2001). We estimate the following model:

$$\ln Vol_{i,t} = \delta z_{i,t-1} + \theta x_{i,t} + c_i + \tau_t + e_{i,t}, \tag{3.2}$$

$$\ln y_{i,t} = \beta \widehat{\ln Vol_{i,t}} + \gamma x_{i,t} + c_i + \tau_t + \epsilon_{i,t}, \tag{3.3}$$

by the ridge-regression estimator, where $z_{i,t-1}$ is a vector of the instrument variables and $x_{i,t}$ contains a vector of controls, including the constant. Equation (3.2) is the first-stage regression, and the relevance of the instruments is examined by the F-statistic of the regression. Based on the first-stage regression, we obtain the estimated $\widehat{\ln Vol_{i,t}}$. $\widehat{\ln Vol_{i,t}}$, is included in Equation (3.3), i.e., the second-stage regression. The exogeneity of the instruments is tested using the Sargan test of over-identification restrictions. As we detail later, our instruments pass both relevance and exogeneity tests.

3.4.3 Debiased (Double) Machine Learning Ridge Regression

Next, we estimate the causal effect of transportation on air quality using the DML method. Specifically, we utilize a DML estimator, using panel data, following Chernozhukov et al. (2018). Our model can be written into a partially linear IV regression model that takes the form:

$$\begin{aligned} \ln y_{i,t} - \beta_0 \ln Vol_{i,t} &= g_0(X_{i,t}) + \zeta, \quad \mathrm{E}[\zeta \mid z_{i,t-1}, X_{i,t}] = 0, \\ z_{i,t-1} &= m_0(X_{i,t}) + V_{i,t}, \quad \mathrm{E}[V_{i,t} \mid X_{i,t}] = 0, \end{aligned}$$

where $\ln y_{i,t}$ represents air quality, $\ln Vol_{i,t}$ indicates transportation for city $i = 1, \ldots, N$ and in month $t = 1, \ldots, T$, $X_{i,t}$ represents control variables for city i in month t which includes $x_{i,t}$, c_i and τ_t. $z_{i,t-1}$ represents COVID-19-related instrument variables for city i in month $t-1$. Thus, β_0 captures the effect of transportation on air quality.

A naive application of machine-learning methods to estimate g_0 and m_0 would involve, for example, the use of iterative methods that alternate between estimating g_0 and m_0 with

machine-learning models and some fixed initial β_0, with estimated g_0 and m_0 to estimate a new β_0 with two-stage least squares, and to eventually obtain converged estimates of β_0, g_0, and m_0. Such a naive application of machine-learning methods, however, can result in very high bias in finite-sample estimates because the estimator will generally have a slower than $1/\sqrt{n}$ rate of convergence (Chernozhukov et al., 2018). Alternatively, we can write the model in the following residualized form:

$$\begin{aligned}
W_{i,t} &= U_{i,t}\theta_0 + \zeta_{i,t}, \quad \mathrm{E}[\zeta_{i,t} \mid z_{i,t-1}, X_{i,t}] = 0, \\
W_{i,t} &= \ln y_{i,t} - \ell_0(X_{i,t}), \quad \ell_0(X_{i,t}) = \mathrm{E}[\ln y_{i,t} \mid X_{i,t}], \\
U_{i,t} &= \ln Vol_{i,t} - h_0(X_{i,t}), \quad h_0(X_{i,t}) = \mathrm{E}[\ln Vol_{i,t} \mid X_{i,t}], \\
V_{i,t} &= z_{i,t-1} - m_0(X_{i,t}), \quad m_0(X_{i,t}) = \mathrm{E}[z_{i,t-1} \mid X_{i,t}].
\end{aligned}$$

We use machine-learning methods and cross-validation to estimate ℓ_0, h_0, and m_0 by the following procedure: First, we divide the data into D evenly-sized folds among cities, with each fold as containing NT/D observations. Second, for each fold $d = 1, \ldots, D$, we run a random forest estimator (Breiman, 2001) on the other $D-1$ data folds to estimate the function ℓ_0, h_0 and m_0 by $\widehat{\ell}_0^{(-d(i))}$, $\widehat{h}_0^{(-d(i))}$ and $\widehat{m}_0^{(-d(i))}$. We can then obtain the estimated residuals as:

$$\begin{aligned}
\widehat{W}_{i,t} &= \ln y_{i,t} - \widehat{\ell}_0^{(-d(i))}(X_{i,t}), \\
\widehat{U}_{i,t} &= \ln Vol_{i,t} - \widehat{h}_0^{(-d(i))}(X_{i,t}), \\
\widehat{V}_{i,t} &= z_{i,t-1} - \widehat{m}_0^{(-d(i))}(X_{i,t}),
\end{aligned}$$

where $d(i) \in \{1, \ldots, D\}$ denotes the fold containing the ith city. The higher the parameter D, the more unbiased the estimator but the higher computational complexity. In practice, D is usually selected as 5 or 10 (Chernozhukov et al., 2018). To get a robust estimator, we select D as 20 in our estimation. Then, using $\widehat{W}_{i,t}$, $\widehat{U}_{i,t}$, and $\widehat{V}_{i,t}$, and by two-stage ridge estimation, we get our estimator $\widehat{\beta}_0$.

3.4.4 Results

To understand the impact of public and private transportation on air quality, we use the total volume of passenger travel by public transportation ($PublicVol_{i,t}$) and the congestion index ($CongIndex_{i,t}$) as the first set of independent variables of interest. To further compare the impacts of public transportation subsectors, we contain the volume of passenger travel by bus, rail, and taxi ($BusVol_{i,t}$, $RailVol_{i,t}$, and $TaxiVol_{i,t}$), and the congestion index ($CongIndex_{i,t}$) as the second set of independent variables of interest. Prior to ridge-regression analyses, multicollinearity tests are conducted. As reported in the Appendix C.5, the correlation coefficients, Farrar-Glauber tests, and cross-pair plot all show that the multicollinearity problem exists in the two model specifications that contain the two sets of independent variables of interest. The estimation of these models through the OLS method will be invalidated (Judge et al., 1988). Therefore, ridge regression is utilized to identify the effects of public and private transportation and the effects of public transportation subsectors. The ridge regression starts with the identification of the best value of the "shrinkage" parameter K. Following Hoerl and Kennard (1976), we optimize K with an iterative procedure. The results of the ridge regression of the above models are as presented below.

Impact of Public and Private Transportation on Overall Air Quality

To address the research questions on the causal effect of transportation on air quality and whether endogeneity exists in the estimation, we compared the results before and after instrument techniques were utilized. Table 3.2 presents the results of the three models when using the synthesized air pollution index as the dependent variable and the total passenger volume of public transportation and the congestion index as the independent variables. Column (1) of Table 3.2 provides the fixed-effects ridge estimates and, Columns (2) and (3) include the fixed-effects two-stage and DML ridge estimates, respectively. The fixed-effects RR estimate of public transportation is 0.046 ($p < 0.01$). Using the instrumental variables approach, the two-stage RR and DML RR estimates of public transportation are

0.052 ($p < 0.001$) and 0.039 ($p < 0.001$), respectively. The fixed-effects RR estimate of private transportation (the congestion index) is 0.089 ($p > 0.1$), and the two-stage and DML RR estimates of private transportation are instead 0.315 ($p < 0.05$) and 0.368 ($p < 0.001$), respectively. The magnitude of the DML RR estimate of private transportation is about four times larger than that of the fixed-effects RR estimate (0.368 vs. 0.089), showing the underestimation of the effect of private transportation on air pollution without instrument variable techniques and introducing non-linearity in the model. The underestimation is consistent with our prediction that there are (1) simultaneity bias that air pollution reduces, on average, the use of private transportation, and (2) non-linear effects of the control variables (which induce the difference between two-stage RR and DML RR). In contrary, the magnitude of the DML RR estimate of public transportation is slightly smaller (-15.22%, $t = 9.600$) than that of the fixed-effects RR estimate, suggesting the complexity in the simultaneity bias in the impact of public transportation on air quality. Such a slight overestimation of public transportation on air pollution before using instruments suggests that the increase of use of public transportation induced by poor air quality might exceed the decrease of use of public transportation.

Column (4) of Table 3.2 reports the DML RR estimate when using COVID-19 in the same month as instruments. As discussed above, air pollution could speed up the transmission of COVID-19 (He et al., 2020), which is opposite to the negative relationship between COVID-19 and air pollution, and using the COVID-19-related variables in the same month may result in an underestimation of the effect of interest (i.e., the invalidity of instruments). Such an underestimation is evidenced by the difference between estimates of private transportation (0.368 vs 0.269) in Columns (3) and (4) in Table 3.2. On the contrary, public transportation's effect on air pollution is slightly overestimated if using the COVID-19 of the same month as instruments (0.039 vs 0.048), also suggesting the complexity of interactions among public transportation, COVID-19, and air pollution. In Section 3.4.4, we will analyze the subsectors of public transportation.

Overall, the estimates represent the transportation elasticities of air quality; that is,

every 1% increase in the passenger volume of public transportation and the congestion index results in about a 0.039% (95% CI: 0.022%, 0.056%) and 0.368% (95% CI: 0.191%, 0.545%) increase in the synthesized air pollution index, respectively, the effect size of which is of economic significance. Comparing estimates from different models confirms the presence of endogeneity when using observational data to estimate transportation's air pollution effect and the necessity of adopting the previous month's COVID-19 as instruments.

Table 3.2: Results for Effects on Overall Air Quality (DV: ln $Synindex_{i,t}$)

	(1) Fixed-effects RR	(2) Two-stage RR	(3) DML RR	(4) DML RR (Same Month)[a]
ln $PublicVol_{i,t}$	0.046**	0.052***	0.039***	0.048***
	(0.015)	(0.011)	(0.009)	(0.009)
ln $CongIndex_{i,t}$	0.089	0.315*	0.368***	0.269**
	(0.178)	(0.124)	(0.090)	(0.104)
$AirTem_{i,t}$	-0.0001	0.001**	YES	YES
	(0.000)	(0.000)		
$DewPoint_{i,t}$	-0.001**	-0.002***	YES	YES
	(0.000)	(0.000)		
$Precip_{i,t}$	0.007	0.005	YES	YES
	(0.041)	(0.036)		
$WindSpeed_{i,t}$	-0.011	-0.076	YES	YES
	(0.090)	(0.076)		
$Production_{i,t}$	0.073	0.062	YES	YES
	(0.114)	(0.101)		
$Constant$	0.992***	0.000	YES	YES
	(0.139)	(0.007)		
Month Effects	YES	YES	YES	YES
City Effects	YES	YES	YES	YES
IV		YES	YES	YES
DML			YES	YES
K	0.057	0.077	0.213	0.201
F-statistic	18.511	11.982	11.676	14.778
N	576	576	576	576

Note: DV stands for dependent variable. Standard errors in parentheses.
+ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.
a. Using COVID-19 in the same month as instruments.

The diagnostic tests reported in Table C.1 in Appendix C.3 demonstrate that the instruments based on the spread of COVID-19 are neither weak nor invalid. The F-statistic for ln $PublicVol_{i,t}$ is high (113.390) and, for ln $CongIndex_{i,t}$, is 54.500, supporting the absence

of a weak instruments problem (Angrist and Krueger, 2001). The Sargan statistic is 0.792 ($p = 0.852$), and the p-value indicates that the instruments constructed based on COVID-19 cases and the Baidu Index are exogenous (Angrist and Krueger, 2001). The F-statistics and Sargan statistic for ln $BusVol_{i,t}$, ln $RailVol_{i,t}$, and ln $TaxiVol_{i,t}$ also show the validity and exogeneity of the instruments. The coefficients of the instrument variables are significant and confirm our argument that the number of COVID-19 cases and the number of COVID-19-related queries predict individuals' willingness to travel and their selection of transportation mode. Specifically, the number of confirmed cases, ln $Con_{i,t-1}$, is negatively related to the passenger volume of all transportation subsectors. There is a positive relationship between the search queries of COVID-19 terms and private transportation (the congestion index, ln $CongIndex_{i,t}$), showing the increase in private transportation adoption due to health concerns related to COVID-19.

Impact of Transportation Subsectors

The results of the transportation subsectors analyses are reported in Table 3.3. First, the DML ridge estimate for the railway passenger volume (0.005, $p > 0.1$, 95% CI: -0.004, 0.013) is not statistically significant, consistent with our expectation that travelling by rail is an environmentally friendly transportation mode. Second, the DML ridge estimate for the bus passenger volume (0.018, $p < 0.001$, 95% CI: 0.010, 0.025) is significantly larger than their fixed-effects counterparts (0.002, $p > 0.1$, 95% CI: -0.030, 0.034), demonstrating the underestimation before using the instrument variables. Third, the DML ridge estimate for the taxi passenger volume (0.023, $p < 0.001$, 95% CI: 0.013, 0.033) is significantly smaller than its fixed-effects counterpart (0.050, $p < 0.05$, 95% CI: 0.009, 0.091), unexpectedly showing the overestimation before using the instrument variables. As discussed above, conditional on the primary endogeneity caused by air pollution's reducing overall transportation, the estimation of the effects of transportation subsectors faces a secondary source of endogeneity, that air pollution increases or decreases passengers' preference for taxis due to health concerns or environmental awareness. The underestimation and the overestimation show that air pollution

reduces, on average, the passenger volume of buses but increases the passenger volume of taxis, implying that the health concerns evoked by air pollution outweigh the environmental awareness triggered by air pollution. In other words, the simultaneity biases in estimating the impact of mass transportation and taxi transportation are opposite, suggesting that air pollution shifts the demand for buses to taxis.

This surprising demand-shifting effect from bus to taxi may pose serious harm to air quality because travelling by taxi is the least environmentally efficient public transportation mode. Moreover, the coexistence of the underestimation and overestimation explains the slight underestimation of the air pollution effect of public transportation before using instrument variables.

The effect of public transportation on air quality may be heterogeneous, as cities differ in the penetration rate of new-energy buses. We do a post-hoc analysis to answer the policy question of the extent to which the new-energy-powered vehicle policy translates into mitigating air pollution from public transportation. We therefore add an interaction term between $NewEnergyBus_i$ (ranging from 10.1% to 97.8%) and ln $BusVol_{i,t}$, into the model and instrument it with the five instrument variables. The results are reported in Table C.2 in the Appendix C.4. The coefficient of the interaction term is negative and significant (-0.006, $p < 0.1$), demonstrating a mitigation effect of new energy buses on the air pollution effect of public transportation. In particular, when the penetration rate is 0, 1% increase in passenger volume by bus would lead to 0.014% increase in synthesized index. But for a city with an average new energy bus penetration rate (i.e., 47.2%), 1% increase in passenger volume by bus would only lead to 0.006% increase in synthesized index.

Impact of Transportation on Primary Air Pollutants

To further understand the heterogeneity of the transportation effects, we estimate the impact of the passenger volume of public transportation and the congestion index on the concentration of the six primary pollutants. Table 3.4 presents the DML RR estimates. All estimates are positive and significant, except the estimate for ln $O3_{i,t}$. The estimates

Table 3.3: Results for Effects of Transportation Subsectors (DV: ln $Synindex_{i,t}$)

	(1) Fixed-effects RR	(2) Two-stage RR	(3) DML RR
ln $BusVol_{i,t}$	0.002	0.025***	0.018***
	(0.016)	(0.006)	(0.004)
ln $RailVol_{i,t}$	-0.003	0.002	0.005
	(0.010)	(0.012)	(0.004)
ln $TaxiVol_{i,t}$	0.050*	0.036***	0.023***
	(0.021)	(0.007)	(0.005)
ln $CongIndex_{i,t}$	0.066	0.196+	0.243***
	(0.177)	(0.111)	(0.062)
$AirTem_{i,t}$	-0.0001	0.001**	YES
	(0.000)	(0.000)	
$DewPoint_{i,t}$	-0.001**	-0.002***	YES
	(0.000)	(0.000)	
$Precip_{i,t}$	0.006	0.003	YES
	(0.041)	(0.035)	
$WindSpeed_{i,t}$	-0.008	-0.057	YES
	(0.091)	(0.075)	
$Production_{i,t}$	0.074	0.070	YES
	(0.114)	(0.099)	
$Constant$	1.058***	0.000	YES
	(0.138)	(0.007)	
Month Effects	YES	YES	YES
City Effects	YES	YES	YES
IV		YES	YES
DML			YES
K	0.054	0.097	0.390
F-statistic	16.879	9.324	5.961
N	576	576	576

Note: DV stands for dependent variable. Standard errors in parentheses.
+ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

of public transportation for ln $NO2_{i,t}$ and ln $PM10_{i,t}$ are the largest, and the estimates for ln $SO2_{i,t}$ and ln $CO_{i,t}$ are the smallest. Specifically, these estimates show that a 1% increase in the volume of public transportation results in a 0.076%, 0.063%, 0.063%, 0.060%, and 0.015% increase in the concentration of NO_2, PM_{10}, CO, $PM_{2.5}$, and SO_2, respectively, and that the effect on the concentration of O_3 is negative (-0.036%). The estimates of the congestion index (private vehicles) on the pollutants are significantly larger than that of public transportation, but have the same pattern (i.e., the effects on NO_2 and PM_{10} are the largest, and the effect on O_3 is negative). These heterogeneous effects of public and private transportation are consistent with the previous finding that transportation is the major source of NO_2, and power generation and heavy industry are the major sources of $PM_{2.5}$ and SO_2 (Diamond and Wood, 2020). The existing research also finds that the COVID-19 pandemic cut transportation dramatically but only slightly influenced power generation and heavy industry, therefore decreasing the concentration of NO_2 but only minimally changing the aerosol optical depth, which is affected mainly by the concentration of $PM_{2.5}$ and SO_2 (Diamond and Wood, 2020). In addition, the estimates for ln $O3_{i,t}$ are negative though insignificant for private transportation. This counterintuitive result may be explained by the fact that O_3 reacts easily with primary emissions of motor vehicles (Zhao et al., 2019).

3.5 Robustness Check

We utilize an alternative measure, driving speed during peak hours, for private transportation to check the robustness of our results. Tables C.5 and C.6 in the Appendix C.6 present the results of the analyses for the synthesized index of air pollution, transportation subsectors, and primary pollutants when using driving speed during peak hours to measure congestion (i.e., the proxy of the volume of private transportation). The fixed-effects, two-stage, and DML RR estimates of driving speed are all negative, and the magnitude of the DML RR estimate (-0.248, $p < 0.001$) is significantly larger than that of the fixed-effects RR estimate (-0.115, $p > 0.1$). The DML estimates of driving speed in the subsector and pollutant analyses also are negative and significant, except for the estimates for O_3. Generally, the results for the air

Table 3.4: Results for the Effects on Primary Air Pollutants (N = 576)

	(1) ln $CO_{i,t}$	(2) ln $NO2_{i,t}$	(3) ln $O3_{i,t}$	(4) ln $PM25_{i,t}$	(5) ln $PM10_{i,t}$	(6) ln $SO2_{i,t}$
ln $PublicVol_{i,t}$	0.063**	0.076***	-0.036*	0.060**	0.063***	0.015+
	(0.019)	(0.017)	(0.016)	(0.018)	(0.012)	(0.009)
ln $CongIndex_{i,t}$	0.409+	0.675**	-0.165	0.427*	0.658***	0.188*
	(0.237)	(0.210)	(0.184)	(0.195)	(0.134)	(0.088)
Controls	YES	YES	YES	YES	YES	YES
Month Effects	YES	YES	YES	YES	YES	YES
City Effects	YES	YES	YES	YES	YES	YES
IV	YES	YES	YES	YES	YES	YES
DML	YES	YES	YES	YES	YES	YES
K	0.052#	0.044	0.085#	0.130#	0.138	0.477#
F-statistic	10.245	28.055	2.957	6.617	18.750	2.332

Note: Standard errors in parentheses. + $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. The K with a "#" is obtained following Hoerl and Kennard (1970a,b), because the procedure following Hoerl and Kennard (1976) fails to converge.

pollution effects of private transportation are robust. All estimates of public transportation and its subsectors are almost the same as the estimates when using the congestion index to measure private transportation. Therefore, the results for public transportation and its subsectors also remain robust.

Because manufacturing also could be endogenous due to air pollution contingency plans, we did an analysis with manufacturing also as instrumented. The results are reported in Tables C.7 and C.8 in the e-companion. Further, we add more control variables to check the robustness of the results. Air pollution worsens during the lunar new year celebration due to fireworks in the suburbs and rural areas (fireworks are prohibited in all Chinese cities). The lunar new year in 2019 was in February, whereas the lunar new year in 2020 was in January. To better control for this holiday effect, we add a dummy, which takes the value of 1 when the lunar new year is in month t, into all of the models as a control variable. The results are reported in Table C.9 in the e-companion.

We also do a few robustness checks with subsamples. Wuhan is a potential outlier; therefore, we rerun our models without Wuhan. In addition, citizens in Wuhan and in cities near Wuhan may have moved out of these cities at the beginning of the pandemic before

lockdown policies were implemented, resulting in a dramatic change in population density and a potential overestimation. Hence, we remove Wuhan and eight nearby cities.[3] The results are reported in Tables C.10 and C.11 in the e-companion. Moreover, one could be concerned that a cross-city spillover effect is a potential confounding factor. That is, the change in air quality in one city may have an impact on the air quality of nearby cities. We argue that this is not likely because the 36 cities in our sample are large distances from each other (the minimal linear distance between the center points of cities is 96 kilometers, and the average linear distance is 1,646 kilometers). To further mitigate this concern, we aggregate geographically close cities as a city cluster and obtain an alternative sample that contains 11 individual cities and 10 city clusters.[4] The monthly statistic of a city cluster is the mean of the statistics of all of the cities contained in the cluster. As such, a cross-city spillover effect would not affect our estimation, as it would happen only within the city clusters. The results are reported in Tables C.12 and C.13 in the e-companion. The results of all the above robustness checks are consistent with our main results.

3.6 Discussion

3.6.1 Policy Implications

Evaluation of Zhengzhou Driving Restriction Policy: With these estimates, we also can estimate air quality change in response to a policy. For example, we can re-evaluate how much air quality was improved by a traffic restriction based on even- and odd-numbered license plates in Zhengzhou (a city in central China) in December 2020. As reported by Baidu Map (2021), the restriction increased by 18.6% passenger visits that involved traveling by bus and by 10.6% passenger visits that used rail travel, and the congestion index decreased by

[3]Eight nearby cities are Hefei, Nanchang, Changsha, Guangzhou, Shenzhen, Hangzhou, Ningbo, and Zhengzhou.

[4]Cities in one city cluster are contained in a pair of square brackets. 21 cities and city clusters are listed below: [Beijing, Tianjin, Shijiazhuang], [Shanghai, Hangzhou, Ningbo], [Guangzhou, Shenzhen], [Chengdu, Chongqing], [Nanjing, Hefei], [Xiamen, Fuzhou], Zhengzhou, [Shenyang, Dalian, Harbin, Changchun], [Wuhan, Nanchang, Changsha], Xi'an, [Qingdao, Jinan], Guiyang, Haikou, Hohhot, Kunming, Lhasa, [Lanzhou, Yinchuan], Nanning, Taiyuan, Urumqi, Xining.

18.34%. Approximately, the synthesized air pollution index decreased by the traffic restriction is 0.172 units (relative change: 4.07%, 95% CI: 0.087, 0.257).[5] As noted, such a decrease will cause a city with an average level of air pollution to be ranked 11 positions higher. Zhengzhou's rank for air quality improved from 151 in November 2020 to 128 in December 2020, which is consistent with our estimation.[6]

Counterfactual Analysis of Shifting Private Transportation to Public Transportation: As one strategy to reduce air pollution is to shift passengers from private transportation to public transportation, we interpret the estimates in a situation in which public transportation increases and private transportation decreases. The estimates show that, for every 10% increase in the passenger volume of public transportation (about 4.43 million passenger visits every day), the synthesized air pollution index increases by 0.016 units (relative change: 0.39%; 95% CI: 0.009, 0.024), while, for every 10% decrease in the congestion index, the synthesized air pollution index decreases by 0.156 units (relative change: 3.68%; 95% CI: 0.081, 0.230), which causes a city with an average level of air pollution to be ranked 11 places higher in air quality among a group of 168 Chinese cities (CNEMC, 2020), making the city more livable for current residents and increasing population inflow. The local government can evaluate the effectiveness of a policy or a new bus/rail route in reducing air pollution easily when it observes the change in passenger volume of public transportation and the congestion index by the policy or route. In addition, the government can easily estimate how many private transportation passenger visits correspond to a 1% congestion index decrease by using the data on speed and vehicle volume, detected by traffic sensors and cameras (Frangoul, 2021), and then compare the environmental efficiency of public and private transportation and evaluate green measures of diverting passengers of private

[5]Notably, the report, without addressing endogeneity or controlling for seasonality and other factors, shows that the air quality index did not change significantly after the restriction was imposed. This highlights the importance of leveraging causal inference to evaluate the impact of a policy.

[6]Zhengzhou's rank for air quality was 159 in November 2019 and 143 in December 2019. The air quality reports for November 2019, December 2019, November 2020, and December 2020 are available at http://www.cnemc.cn/jcbg/kqzlzkbg/.

transportation to public transportation. Based on the calculation below[7], if 1% decrease in congestion index corresponds to less than 2.48 million[8] private transport passenger visits every day, shifting passengers from private transportation to public transportation will improve air quality. The congestion index is very responsive to public transportation passenger volume, as indicated by Anderson (2014). Overall, shifting passengers from private transportation to public transportation in Chinese cities is likely to improve air quality. In comparison, Rivers et al. (2020) find that the current configuration of public transits in North American cities counter-intuitively increases the concentration of NO_2 and has no statistically significant effect on CO or $PM_{2.5}$ and concludes that expanding the current configuration of public transit in North American cities will not improve air quality. The counterfactual analysis shows that expanding the current configuration of public transits in Chinese cities will realize improvement in air quality.

Implied Mortality Decrease by Shifting Private Transportation to Public Transportation: These results also demonstrate economic significance, especially for substituting private transportation with public transportation. The estimates show that a 10% increase in public transportation passenger volume will result in a 0.269 $\mu g/m^3$ (relative change: 0.760%; 95% CI: 0.151, 0.386) increase in the concentration of NO_2, 0.421 $\mu g/m^3$ (relative change: 0.630%; 95% CI: 0.259, 0.582) increase in the concentration of PM_{10}, 0.007 mg/m^3 (relative change: 0.630%; 95% CI: 0.003, 0.012) increase in the concentration of CO, 0.240 $\mu g/m^3$ (relative change: 0.600%; 95% CI: 0.101, 0.380) increase in the concentration of $PM_{2.5}$, 0.015 $\mu g/m^3$ (relative change: 0.150%; 95% CI: -0.002, 0.033) increase in the concentration of

[7]If a policy or a new bus/rail route shifts about 0.443 million passenger visits every day (1% of public transportation passenger visits each day) from private transportation to public transportation, and a 1% congestion index decrease corresponds to a million private transportation passenger visits every day, the synthesized air pollution index will decrease by $(0.368 - 0.039 \times (a/0.443))\%$.

[8]According to Baidu Map (2021), the driving restriction in Zhengzhou in December 2020 decreased the congestion index by 18.34%. According to MOT (2021), the passenger volume of public transportation increased by 21.06 million from November 2020 to December 2020. Therefore, 1% decrease in congestion index corresponded to an increase in 1.15 million public transportation passengers in Zhengzhou, less than 2.48 million passengers change in public transportation, which is a turning point in decreasing air pollution by shifting private transportation to public transportation.

SO_2, and 0.449 $\mu g/m^3$ (relative change:-0.360%; 95% CI: 0.058, 0.841) decrease in the daily maximum 8-hour O_3 concentration, yielding a 0.001—0.004-point increase in relative risk of all non-accidental mortality[9], at least a 0.197% increase in labor outflow.[10] The estimate also shows a 10% decrease in the congestion index will result in a 2.388 $\mu g/m^3$ (relative change: 6.750%; 95% CI: 0.931, 3.845) decrease in the concentration of NO_2, 4.393 $\mu g/m^3$ (relative change: 6.580%; 95% CI: 2.633, 6.153) decrease in the concentration of PM_{10}, 0.048 mg/m^3 (relative change: 4.090%; 95% CI: -0.007, 0.103) decrease in the concentration of CO, 1.710 $\mu g/m^3$ (relative change: 4.270%; 95% CI: 0.180, 3.241) decrease in the concentration of $PM_{2.5}$, 0.192 $\mu g/m^3$ (relative change: 1.880%; 95% CI: 0.016, 0.368) decrease in the concentration of SO_2, and 2.064 $\mu g/m^3$ (relative change: -1.654%; 95% CI: -2.453, 6.582) increase in the daily maximum 8-hour O_3 concentration, yielding a 0.006—0.037-point decrease in the relative risk of all non-accidental mortality[9] and roughly a 1.585% decrease in labor outflow.[10] If we assume that 10% increase in public transport passenger volume corresponds to 10% decrease in the congestion index, the demand-shifting from private transportation to public transportation will result in 0.002-0.036 point decrease in relative risk of all non-accidental mortality, which is about 200—3,600 people will die in a city with a population of 10 million.[11] If we assume

[9]The relative risk is the ratio of the probability of a disease or health outcome (e.g., premature death, asthma attack, hospital admission) in an exposed group to an air pollutant to the probability of the outcome in an unexposed group. WHO (2021) reports meta-analytic effect estimates of relative risk of 1.02 per 10 $\mu g/m^3$ NO_2 in all-cause non-accidental mortality, relative risk of 1.04 per 10 $\mu g/m^3$ PM_{10} in all-cause non-accidental mortality, relative risk of 1.08 per 10 $\mu g/m^3$ $PM_{2.5}$ in all-cause non-accidental mortality, and relative risk of 1.01 in all-cause non-accidental mortality per 10 $\mu g/m^3$ increase in the peak-season average of the daily maximum 8-hour mean O_3 concentration when establishing long-term air quality guideline levels with annual means of concentrations of these air pollutants. For short-term 24-hour means of concentrations of these air pollutants, WHO (2021) reports relative risk of 1.0072 per 10 $\mu g/m^3$ NO_2 in all-cause non-accidental mortality, relative risk of 1.0041 per 10 $\mu g/m^3$ PM_{10} in all-cause non-accidental mortality, relative risk of 1.052 in hospital admissions and mortality from myocardial infarction per 1 mg/m^3 CO, relative risk of 1.0065 per 10 $\mu g/m^3$ $PM_{2.5}$ in all-cause non-accidental mortality, relative risk of 1.0059 per 10 $\mu g/m^3$ SO_2 in daily mortality, and relative risk of 1.0043 in all-cause non-accidental mortality per 10 $\mu g/m^3$ increase in 8-hour maximum O_3 concentration. Because the measurement of the concentration of air pollutants in this study is the monthly mean value, we calculate the range of relative risk for our estimates using these values of relative risk based on the annual mean and 24-hour statistics reported by WHO (2021).

[10]According to Chen et al. (2017), a 10% increase in the annual concentration of $PM_{2.5}$ and SO_2 results in about a reduction in population of 2.7 and 2.3 per 100 inhabitants due to migration, respectively.

[11]China has 17 cities with a population of more than 10 million in 2021

that half of the world population is urbanized and apply the estimate of the interval of risk of non-accidental mortality reduced by shifting private transportation to public transportation, obtained in our data set, the life saved by shifting 10% private transportation passenger volume to public transportation passenger volume is about 60,000—1080,000, averagely 570,000. According to Anenberg et al. (2019), 385,000 deaths resulted from transportation tailpipe globally in 2015. Comparison of the deaths caused by 10% transportation using estimate from China with the deaths caused by all transportation globally, we find that China's air pollution-induced health problem is worse than the world average level.

3.6.2 Limitations and Future Research

This study has a few limitations that may open promising avenues for future research. First, ride sharing is considered private transportation, and the heterogeneity between rides via self-owned cars and rides via carpooling is not considered due to data limitations in this study. Future work may obtain data of the volumes of cars that accommodate single passenger or multiple passengers to understand the heterogeneous effect of ride sharing on air quality. Second, the potential demand-shifting effects between private and public transportation by air pollution remain unknown due to the unavailability of passenger volume data for private vehicles. Future work can investigate air pollution's effect on demand transfer between public and private transportation with more complete passenger volume data or individual-level transportation mode data. Our findings also suggest that studies on green transportation mode choice behaviors may suffer from the reverse causality from air quality. For example, when investigating the air pollution effect of motorized and human-powered transportation modes, air pollution may shift passengers between the two modes. Environmental awareness facilitates the shift from motorized to human-powered modes, whereas health concerns about exposure to polluted air have the opposite effect. In addition, future research can further investigate the demand-shifting effect among different transportation subsectors as related to

(https://finance.ifeng.com/c/8FuAunSSgGm).

air pollution.

3.7 Conclusion

Leveraging the outbreak of the COVID-19 pandemic, this chapter constructs instruments for transportation and estimates the impact of transportation and its subsectors on air quality. The fixed-effects RR, two-stage RR, and DML RR estimates demonstrate the presence of strong endogeneity in using the observational data. The tested endogeneity suggests that air pollution reduces transportation and shifts passengers from mass transportation to taxis. Therefore, sustainable operation of transportation in industrializing and developing countries should further promote passengers' usage of the bus and rail transportation, improve air quality in bus and rail systems, and reduce traffic congestion. Considering such passenger behavioral responses to air pollution in decisions to travel and choices of transportation modes are also needed for related future research. The estimates for six primary pollutants additionally show that sustainable operation of transportation will be most effective in reducing the concentration of NO_2 and PM_{10}. The finding of the negative effect of private transportation on the concentrations of O_3 calls for future exploration of the underlying mechanism. This shows the potential trade-off between O_3 and other pollutants when implementing green transportation policies. The DML estimates can be adopted in sustainable transportation planning and policy evaluation. In all, this chapter provides estimates for the passenger-volume-level impact of public transportation on air quality as well as the air pollution effect of private transportation in the dimension of congestion, giving implications for green transportation in both a policy identification stage and an ex-ante policy assessment stage.

BIBLIOGRAPHY

Derya Eren Akyol and René B.M. De Koster. Non-dominated time-window policies in city distribution. *Production and Operations Management*, 22(3):739–751, 2013.

Douglas Almond, Xinming Du, Valerie J. Karplus, and Shuang Zhang. Ambiguous air pollution effects of China's COVID-19 lock-down. *AEA Papers and Proceedings*, 111: 376–380, may 2021.

Michael L Anderson. Subways, strikes, and slowdowns: The impacts of public transit on traffic congestion. *American Economic Review*, 104(9):2763–96, 2014.

James Andreoni, Brian Erard, and Jonathan Feinstein. Tax compliance. *Journal of economic literature*, 36(2):818–860, 1998.

Susan C Anenberg, Joshua Miller, Daven K Henze, Ray Minjares, and Pattanun Achakulwisut. The global burden of transportation tailpipe emissions on air pollution-related mortality in 2010 and 2015. *Environmental Research Letters*, 14(9):094012, sep 2019.

Joshua D. Angrist and Alan B. Krueger. Instrumental variables and the search for identification: From supply and demand to natural experiments. *The Journal of Economic Perspectives*, 15(4):69–85, 2001.

Joshua D Angrist, Stacey H Chen, and Brigham R Frandsen. Did vietnam veterans get sicker in the 1990s? the complicated effects of military service on self-reported health. *Journal of public Economics*, 94(11-12):824–837, 2010.

Atalay Atasu, Charles J. Corbett, Ximin Huang, and L. Beril Toktay. Sustainable operations management through the perspective of *Manufacturing & Service Operations Management*. *Manufacturing and Service Operations Management*, 22(1):146–157, 2020.

Susan Athey and Guido W Imbens. Identification and inference in nonlinear difference-in-differences models. *Econometrica*, 74(2):431–497, 2006.

Susan Athey and Stefan Wager. Policy learning with observational data. *Econometrica*, 89 (1):133–161, 2021.

Buket Avci, Karan Girotra, and Serguei Netessine. Electric vehicles with a battery switching station: Adoption and environmental impact. *Management Science*, 61(4):772–794, apr 2015.

Alba Badia, Johannes Langemeyer, Xavier Codina, Joan Gilabert, Nacho Guilera, Veronica Vidal, Ricard Segura, Mar Vives, and Gara Villalba. A take-home message from COVID-19 on urban air pollution reduction through mobility limitations and teleworking. *npj Urban Sustainability*, 1(1):35, dec 2021.

Baidu Map. 2020 China urban transportation report. Technical report, Baidu, Beijing, China, 2021. Accessed March 19, 2021, https://jiaotong.baidu.com/reports/.

M Bauwens, S Compernolle, T Stavrakou, J-F Müller, J Van Gent, H Eskes, Pieternel Felicitas Levelt, R van der A, JP Veefkind, J Vlietinck, et al. Impact of coronavirus outbreak on NO_2 pollution assessed using TROPOMI and OMI observations. *Geophysical Research Letters*, 47(11):e2020GL087978, 2020.

Roland Bénabou and Jean Tirole. Intrinsic and extrinsic motivation. *The review of economic studies*, 70(3):489–520, 2003.

Hugo Benítez-Silva, Moshe Buchinsky, Hiu Man Chan, Sofia Cheidvasser, and John Rust. How large is the bias in self-reported disability? *Journal of Applied Econometrics*, 19(6): 649–670, 2004.

Nicole Black, David W Johnston, and Agne Suziedelyte. Justification bias in self-reported disability: New evidence from panel data. *Journal of Health Economics*, 54:124–134, 2017.

Christopher Blattman, Julian Jamison, Tricia Koroknay-Palicz, Katherine Rodrigues, and Margaret Sheridan. Measuring the measurement error: A method to qualitatively validate survey data. *Journal of Development Economics*, 120:99–112, 2016.

Tobias Brandt and Oliver Dlugosch. Exploratory data science for discovery and ex-ante assessment of operational policies: Insights from vehicle sharing. *Journal of Operations Management*, 67(3):307–328, 2021.

Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

Peter Brimblecombe and Yonghang Lai. Effect of Fireworks, Chinese New Year and the COVID-19 Lockdown on Air Pollution and Public Attitudes. *Aerosol and Air Quality Research*, 20(11):2318–2331, 2020.

Guillaume Carlier, Victor Chernozhukov, and Alfred Galichon. Vector quantile regression: An optimal transport approach. *The Annals of Statistics*, 44(3):1165 – 1192, 2016. doi: 10.1214/15-AOS1401. URL https://doi.org/10.1214/15-AOS1401.

Kai Chen, Meng Wang, Conghong Huang, Patrick L Kinney, and Paul T Anastas. Air pollution reduction and mortality benefit during the COVID-19 outbreak in China. *The Lancet Planetary Health*, 4(6):e210–e212, jun 2020.

Ruidi Chen and Ioannis Ch. Paschalidis. Distributionally robust learning, 2021.

Shuai Chen, Paulina Oliva, and Peng Zhang. The effect of air pollution on migration: evidence from china. 2017. NBER working paper, National Bureau of Economic Research, Cambridge, MA.

Xiaohong Chen. Large sample sieve estimation of semi-nonparametric models. *Handbook of econometrics*, 6:5549–5632, 2007.

Xiaohong Chen and Zhipeng Liao. Sieve semiparametric two-step gmm under weak dependence. *Journal of Econometrics*, 189(1):163–186, 2015.

Xiaohong Chen and Xiaotong Shen. Sieve extremum estimates for weakly dependent data. *Econometrica*, pages 289–314, 1998.

Xiaohong Chen, Han Hong, and Elie Tamer. Measurement error models with auxiliary data. *The Review of Economic Studies*, 72(2):343–366, 2005.

Xiaohong Chen, Yanqin Fan, and Wendao Xue. Identification and estimation of treatment effects in the limited overlap region. 2023.

Victor Chernozhukov, Alfred Galichon, Marc Hallin, Marc Henry, et al. Monge–kantorovich depth, quantiles, ranks and signs. *The Annals of Statistics*, 45(1):223–256, 2017.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 01 2018.

Pierre-André Chiappori, Carlo V Fiorio, Alfred Galichon, and Stefano Verzillo. Assortative matching on income. *JRC Working Papers in Economics and Finance*, 2022.

Chinese Academy of Transportation Sciences. 2019 China new energy bus penetration report. Technical report, 2020. URL https://www.thepaper.cn/newsDetail_forward_8358283.

CMA. Data Center of National Meteorological Sciences of China, 2020. Accessed March 19, 2021, http://data.cma.cn/.

CNEMC. Monthly air quality report for Chinese central cities. Technical report, China National Environmental Monitoring Centre, Beijing, China, 2020. Accessed March 19, 2021, http://www.cnemc.cn/jcbg/kqzlzkbg/.

Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy. Optimal transport for domain adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 39(9):1853–1865, 2016.

Richard K Crump, V Joseph Hotz, Guido W Imbens, and Oscar A Mitnik. Dealing with limited overlap in estimation of average treatment effects. *Biometrika*, 96(1):187–199, 2009.

JA Cuesta-Albertos, C Matrán-Bea, and A Tuero-Diaz. On lower bounds for thel 2-wasserstein metric in a hilbert space. *Journal of Theoretical Probability*, 9(2):263–283, 1996.

Barbara Culiberg, Hichang Cho, Mateja Kos Koklic, and Vesna Zabkar. From car use reduction to ride-sharing: The relevance of moral and environmental identity. *Journal of Consumer Behaviour*, pages 1–12, 2022.

E. Daily. Spring farming is busy at villages without epidemic. Technical report, 2020. URL http://www.xinhuanet.com/local/2020-03/01/c_1125645258.htm.

Lucas W. Davis. The effect of driving restrictions on air quality in Mexico City. *Journal of Political Economy*, 116(1):38–81, 2008.

Rajeev H Dehejia and Sadek Wahba. Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American statistical Association*, 94 (448):1053–1062, 1999.

Alex Delalande and Quentin Merigot. Quantitative stability of optimal transport maps under variations of the target measure. *arXiv preprint arXiv:2103.05934*, 2021.

Michael S Diamond and Robert Wood. Limited regional aerosol and cloud microphysical changes despite unprecedented decline in nitrogen oxide pollution during the February 2020 COVID-19 shutdown in China. *Geophysical Research Letters*, 47(17):e2020GL088913, 2020.

Iván Díaz and Mark J van der Laan. Sensitivity analysis for causal inference under unmeasured confounding and measurement error problems. *The international journal of biostatistics*, 9 (2):149–160, 2013.

David F. Drake and Stefan Spinler. Sustainable operations management: An enduring stream

or a passing fancy? *Manufacturing and Service Operations Management*, 15(4):689–700, 2013.

Avraham Ebenstein, Maoyong Fan, Michael Greenstone, Guojun He, and Maigeng Zhou. New evidence on the impact of sustained exposure to air pollution on life expectancy from China's Huai River Policy. *Proceedings of the National Academy of Sciences*, 114(39): 10384–10389, 2017.

Ivar Ekeland, Alfred Galichon, and Marc Henry. Comonotonic measures of multivariate risks. *Mathematical Finance: An International Journal of Mathematics, Statistics and Financial Economics*, 22(1):109–132, 2012.

Yanqin Fan, Marc Henry, Brendan Pass, and Jorge A. Rivero. Lorenz map, inequality ordering and curves based on multidimensional rearrangements, 2022.

Yanqin Fan, Hyeonseok Park, and Gaoqian Xu. Quantifying distributional model risk in marginal problems via optimal transport, 2023.

Donald E. Farrar and Robert R. Glauber. Multicollinearity in Regression Analysis: The Problem Revisited. *The Review of Economics and Statistics*, 49(1):92, feb 1967.

Max H Farrell, Tengyuan Liang, and Sanjog Misra. Deep neural networks for estimation and inference. *Econometrica*, 89(1):181–213, 2021.

Urs Fischbacher and Franziska Föllmi-Heusi. Lies in disguise—an experimental study on cheating. *Journal of the European Economic Association*, 11(3):525–547, 2013.

Rémi Flamary, Karim Lounici, and André Ferrari. Concentration bounds for linear monge mapping estimation and optimal transport domain adaptation. *arXiv preprint arXiv:1905.10155*, 2019.

Phil Justice Flores and Johan Jansson. The role of consumer innovativeness and green

perceptions on green innovation use: The case of shared e-bikes and e-scooters. *Journal of Consumer Behaviour*, 20(6):1466–1479, 2021.

Piers M Forster, Harriet I Forster, Mat J Evans, Matthew J Gidden, Chris D Jones, Christoph A Keller, Robin D Lamboll, Corinne Le Quéré, Joeri Rogelj, Deborah Rosen, et al. Current and future global climate impacts resulting from COVID-19. *Nature Climate Change*, 10:913–919, 2020.

Anmar Frangoul. How traffic sensors and cameras are transforming city streets, feb 2021. URL https://www.cnbc.com/2021/02/22/how-traffic-sensors-and-cameras-are-transforming-city-streets.html.

Alex Frankel and Navin Kartik. Muddled information. *Journal of Political Economy*, 127(4): 1739–1776, 2019.

David A Freedman. Statistical models and shoe leather. *Sociological methodology*, pages 291–313, 1991.

VN Gabushin. Inequalities for the norms of a function and its derivatives in metric l_p. *Matematicheskie Zametki*, 1(3):291–298, 1967.

Alfred Galichon. *Optimal Transport Methods in Economics*. Princeton University Press, 2016.

Alfred Galichon. *Optimal transport methods in economics*. Princeton University Press, 2018.

Alfred Galichon and Marc Henry. Dual theory of choice with multivariate risks. *Journal of Economic Theory*, 147(4):1501–1516, 2012.

Matthias Gelbrich. On a formula for the l2 wasserstein metric between measures on euclidean and hilbert spaces. *Mathematische Nachrichten*, 147(1):185–203, 1990.

Ben Gillen, Erik Snowberg, and Leeat Yariv. Experimenting with measurement error: Techniques with applications to the caltech cohort study. *Journal of Political Economy*, 127(4):1826–1863, 2019.

Jeremy Ginsberg, Matthew H Mohebbi, Rajan S Patel, Lynnette Brammer, Mark S Smolinski, and Larry Brilliant. Detecting influenza epidemics using search engine query data. *Nature*, 457(7232):1012–1014, 2009.

GitHub. COVID-19/2019-nCoV Time Series Infection Data Warehouse, 2020. Accessed March 19, 2021, https://github.com/BlankerL/DXY-COVID-19-Data/blob/master/README.en.md.

Michael Greenstone and Rema Hanna. Environmental regulations, air and water pollution, and infant mortality in India. *American Economic Review*, 104(10):3038–3072, 2014.

Jinyong Hahn. On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, 66(2):315–331, 1998. ISSN 00129682, 14680262.

Rema Hanna and Paulina Oliva. The effect of pollution on labor supply: Evidence from a natural experiment in Mexico City. *Journal of Public Economics*, 122:68–79, 2015.

Keegan Harris, Dung Daniel T Ngo, Logan Stapleton, Hoda Heidari, and Steven Wu. Strategic instrumental variable regression: Recovering causal relationships from strategic responses. In *International Conference on Machine Learning*, pages 8502–8522. PMLR, 2022.

Guojun He, Yuhang Pan, and Takanao Tanaka. The causal effect of air pollution on COVID-19 transmission: Evidence from China. *medRxiv*, 2020. URL https://www.medrxiv.org/content/early/2020/10/21/2020.10.19.20215236.

Long He, Sheng Liu, and Zuo-Jun Max Shen. Smart urban transport and logistics: A business analytics perspective. *Production and Operations Management*, 31(10):3771–3787, 2022.

James Heckman, Hidehiko Ichimura, Jeffrey Smith, and Petra Todd. Characterizing selection bias using experimental data. *Econometrica*, 66(5):1017–1098, 1998a. ISSN 00129682, 14680262.

James J Heckman, Hidehiko Ichimura, and Petra Todd. Matching as an econometric evaluation estimator. *The review of economic studies*, 65(2):261–294, 1998b.

Asrah Heintzelman, Gabriel Filippelli, and Vijay Lulla. Substantial Decreases in U.S. Cities' Ground-Based NO2 Concentrations during COVID-19 from Reduced Transportation. *Sustainability*, 13(16):9030, aug 2021.

Keisuke Hirano, Guido W Imbens, and Geert Ridder. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4):1161–1189, 2003.

Arthur E. Hoerl and Robert W. Kennard. Ridge regression: Applications to nonorthogonal problems. *Technometrics*, 12(1):69–82, 1970a.

Arthur E. Hoerl and Robert W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970b.

Arthur E. Hoerl and Robert W. Kennard. Ridge regression iterative estimation of the biasing parameter. *Communications in Statistics - Theory and Methods*, 5(1):77–88, jan 1976.

Stephen P. Holland, Erin T. Mansur, Nicholas Z. Muller, and Andrew J. Yates. Are there environmental benefits from driving electric vehicles? The importance of local factors. *American Economic Review*, 106(12):3700–3729, 2016.

Bengt Holmström. Moral hazard and observability. *The Bell journal of economics*, pages 74–91, 1979.

Chaolin Huang, Yeming Wang, Xingwang Li, Lili Ren, Jianping Zhao, Yi Hu, Li Zhang, Guohui Fan, Jiuyang Xu, Xiaoying Gu, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *The Lancet*, 395(10223):497–506, 2020.

Pengrun Huang and Maggie Makar. Conditional differential measurement error: partial identifiability and estimation. In *NeurIPS workshop on causal machine learning for real world impact*, 2022.

Xuan Huang and Bruno Lanz. The value of air quality in Chinese cities: Evidence from labor and property market outcomes. *Environmental and Resource Economics*, 71(4):849–874, dec 2018.

Jan-Christian Hütter and Philippe Rigollet. Minimax rates of estimation for smooth optimal transport maps. *arXiv preprint arXiv:1905.05828*, 2019.

Koichiro Ito and Shuang Zhang. Willingness to pay for clean air: Evidence from air purifier markets in China. *Journal of Political Economy*, 128(5):1627–1672, 2020.

O. Jabali, T. Van Woensel, and A.G. de Kok. Analysis of travel times and co2 emissions in time-dependent vehicle routing. *Production and Operations Management*, 21(6):1060–1074, 2012.

Zhichao Jiang and Peng Ding. Measurement errors in the binary instrumental variable model. *Biometrika*, 107:238–245, 3 2020. ISSN 14643510. doi: 10.1093/biomet/asz060.

George G. Judge, Carter Hill, William E. Griffiths, Helmut Lutkepohl, and Tsoung-Chao Lee. *Introduction to the Theory and Practice of Econometrics*. John Wiley & Sons, New York, 2nd edition, 1988.

Akhil Kadiyala and Ashok Kumar. An examination of the sensitivity of Sulfur Dioxide, Nitric Oxide, and Nitrogen Dioxide concentrations to the important factors affecting air quality inside a public transportation bus. *Atmosphere*, 3(2):266–287, jun 2012.

Frank J. Kelly and Tong Zhu. Transport solutions for cleaner air. *Science*, 352(6288):934–936, 2016.

Wolfgang Ketter, Karsten Schroer, and Konstantina Valogianni. Information systems research for smart sustainable mobility: A framework and call for action. *Information Systems Research*, 2022.

Jesse H. Kroll, Colette L. Heald, Christopher D. Cappa, Delphine K. Farmer, Juliane L. Fry, Jennifer G. Murphy, and Allison L. Steiner. The complex chemical effects of COVID-19 shutdowns on air quality. *Nature Chemistry*, 12(9):777–779, 2020.

Edward P Lazear and Sherwin Rosen. Rank-order tournaments as optimum labor contracts. *Journal of political Economy*, 89(5):841–864, 1981.

Arthur Lewbel. Estimation of average treatment effects with misclassification. *Econometrica*, 75(2):537–551, 2007.

Ziru Li, Chen Liang, Yili Hong, and Zhongju Zhang. How do on-demand ridesharing services affect traffic congestion? the moderating role of urban compactness. *Production and Operations Management*, 31(1):239–258, 2022.

Thomas E MaCurdy. The use of time series processes to model the error structure of earnings in a longitudinal data analysis. *Journal of econometrics*, 18(1):83–114, 1982.

César Martinelli and Susan Wendy Parker. Deception and misreporting in a social program. *Journal of the European Economic Association*, 7(4):886–908, 2009.

Robert J McCann. Existence and uniqueness of monotone measure-preserving maps. 1995.

Daniel L Millimet. The elephant in the corner: a cautionary tale about measurement error in treatment effects models. In *Missing data methods: Cross-sectional methods and applications*. Emerald Group Publishing Limited, 2011.

Hussein M. Mir, Koorosh Behrang, Mohammad T. Isaai, and Pegah Nejat. The impact of outcome framing and psychological distance of air pollution consequences on transportation mode choice. *Transportation Research Part D: Transport and Environment*, 46:328–338, jul 2016.

Gaspard Monge. Mémoire sur la théorie des déblais et des remblais. *Mem. Math. Phys. Acad. Royale Sci.*, pages 666–704, 1781.

MOT. Statistic report on China's transportation sector in 2019. Technical report, MOT, Beijing, China, 2020. Accessed March 19, 2021, http://xxgk.mot.gov.cn/jigou/zhghs/202005/t20200512{_}3374322.html.

MOT. Transportation intelligence and data, 2021. Accessed March 19, 2021, https://www.mot.gov.cn/tongjishuju/chengshikeyun/index.html.

National Bureau of Statistics. Data Query for Percentage of Urban Population with Access to Gas, 2021. URL http://data.stats.gov.cn.proxy.stats.gov.cn/easyquery.htm?proxy=https_&cn=C01&zb=A0B0A&sj=2019.

Sergey Naumov, David R. Keith, and Charles H. Fine. Unintended consequences of automated vehicles and pooling for urban transportation systems. *Production and Operations Management*, 29(5):1354–1371, 2020.

Rachel C Nethery, Fabrizia Mealli, and Francesca Dominici. Estimating population average causal effects in the presence of non-overlap: The effect of natural gas compressor station exposure on cancer mortality. *The annals of applied statistics*, 13(2):1242, 2019.

Whitney K Newey. The asymptotic variance of semiparametric estimators. *Econometrica: Journal of the Econometric Society*, pages 1349–1382, 1994.

Hyunwoo Park, Christian C. Blanco, and Elliot Bendoly. Vessel sharing and its impact on maritime operations and carbon emissions. *Production and Operations Management*, 31(7): 2925–2942, 2022.

Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

Aram-Alexandre Pooladian and Jonathan Niles-Weed. Entropic estimation of optimal transport maps, 2022.

Nicholas Rivers, Soodeh Saberian, and Brandon Schaufele. Public transit and air pollution: Evidence from canadian transit strikes. *Canadian Journal of Economics/Revue canadienne d'économique*, 53(2):496–525, 2020.

Peter M Robinson. Root-n-consistent semiparametric regression. *Econometrica*, 56(4):931–954, 1988.

Dani Rodrik. The new development economics: we shall experiment, but how shall we learn? 2008. HKS working paper, Harvard Kennedy School, Cambridge, MA.

P. R. Rosenbaum and D. B. Rubin. Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society. Series B (Methodological)*, 45(2):212–218, 1983a. ISSN 00359246.

Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983b.

Hong Ru, Endong Yang, and Kunru Zou. Combating the COVID-19 pandemic: The role of the SARS imprint. *Management Science*, 67(9):5606–5615, sep 2021.

Filippo Santambrogio. Optimal transport for applied mathematicians. *Birkäuser, NY*, 55 (58-63):94, 2015.

Alex Scott, Ming Li, David E. Cantor, and Thomas M. Corsi. Do voluntary environmental programs matter? evidence from the epa smartway program. *Journal of Operations Management*, pages 1–21, 2022.

Xiaotong Shen. On methods of sieves and penalization. *The Annals of Statistics*, 25(6): 2555–2591, 1997.

Di Shu and Grace Y Yi. Causal inference with measurement error in outcomes: Bias analysis and estimation methods. *Statistical methods in medical research*, 28(7):2049–2068, 2019.

Ari Stern. L^ p change of variables inequalities on manifolds. *arXiv preprint arXiv:1004.0401*, 2010.

Chuanwang Sun, Wenyue Zhang, Xingming Fang, Xiang Gao, and Meilian Xu. Urban public transport and air quality: Empirical study of China cities. *Energy Policy*, 135:110998, 2019.

Xuelin Tian, Chunjiang An, Zhikun Chen, and Zhiqiang Tian. Assessing the impact of COVID-19 pandemic on urban transportation and air quality in Canada. *Science of The Total Environment*, 765:144270, apr 2021.

Tyler J VanderWeele and Yige Li. Simple sensitivity analysis for differential measurement error. *American journal of epidemiology*, 188(10):1823–1829, 2019.

Cédric Villani. *Optimal transport: old and new*, volume 338. Springer, 2009.

Pengfei Wang, Kaiyu Chen, Shengqiang Zhu, Peng Wang, and Hongliang Zhang. Severe air pollution events not avoided by reduced anthropogenic activities during COVID-19 outbreak. *Resources, Conservation and Recycling*, 158:104814, jul 2020.

Qiang Wang and Xuan Yang. How do pollutants change post-pandemic? Evidence from changes in five key pollutants in nine Chinese cities most affected by the COVID-19. *Environmental Research*, 197:111108, jun 2021.

Wenbin Wang, Mark E. Ferguson, Shanshan Hu, and Gilvan C. Souza. Dynamic capacity investment with two competing technologies. *Manufacturing and Service Operations Management*, 15(4):616–629, 2013.

WHO. Novel coronavirus – China, 2020. Accessed March 19, 2021, https://www.who.int/csr/don/12-january-2020-novel-coronavirus-china/en/.

WHO. WHO global air quality guidelines: particulate matter ($PM_{2.5}$ and PM_{10}), ozone,

nitrogen dioxide, sulfur dioxide and carbon monoxide. Technical report, 2021. URL https://www.who.int/publications/i/item/9789240034228?ua=1.

Jeffrey M. Wooldridge. *Econometric Analysis of Cross Section and Panel Data*. MIT Press, Cambridge, MA, 2010.

World Bank. $PM_{2.5}$ air pollution, population exposed to levels exceeding WHO guideline value (% of total), 2021. URL https://data.worldbank.org/indicator/EN.ATM.PM25.MC.ZS?view=chart.

Hongkang Yang and Esteban G Tabak. Clustering, factor discovery and optimal transport. *Information and Inference: A Journal of the IMA*, 12 2020. ISSN 2049-8772. iaaa040.

Zhiming Yang, Fujun Li, and Wei Lin. Manufacturing is the first of work resumption-a quick survey on top 100 manufacturers of China. Technical report, Economic Information Daily, 2020. URL http://www.jjckb.cn/2020-03/03/c_138837044.htm.

Yifan Yu, Lin Jia, and Yong Tan. Emotion ai meets strategic users. 2023. URL https://ssrn.com/abstract=4218083.

Suping Zhao, Ye Yu, Dahe Qin, Daiying Yin, Longxiang Dong, and Jianjun He. Analyses of regional pollution and transportation of $PM_{2.5}$ and ozone in the city clusters of Sichuan Basin, China. *Atmospheric Pollution Research*, 10(2):374–385, 2019.

Øystein Daljord, Guillaume Pouliot, Junji Xiao, and Mandy Hu. The black market for beijing license plates, 2021.

Appendix A

APPENDICES FOR IDENTIFICATION AND ESTIMATION OF TREATMENT EFFECTS IN THE NON-OVERLAP REGION

A.1 Proofs in Section 1.4

Proposition 10. Under Assumptions 1, 3, 4 and 17, if $\frac{J_{n_0}}{n_0} \to 0$ and $J_{n_0} \to \infty$, then

$$\left\|\widehat{h}(\widehat{T}(\cdot),\cdot) - h(T_o(\cdot),\cdot)\right\|_{\infty,\omega} = o_p(1) \text{ and } \sup_{\beta\in\mathcal{B}}\left\|\widehat{h}(\widehat{T}(\cdot),\beta) - h(T_o(\cdot),\beta)\right\|_{2,1} = o_p(1).$$

Proof of Proposition 10. We note that

$$\left\|\widehat{h}(\widehat{T}(\cdot),\cdot) - h\left(T_o(\cdot),\cdot\right)\right\|_{\infty,\omega} \leq \left\|\widehat{h} - h\right\|_{\infty,\omega} + \left\|h(\widehat{T}(\cdot),\cdot) - h\left(T_o(\cdot),\cdot\right)\right\|_{\infty,\omega}.$$

By Proposition A1 in Chen, Hong, and Tamer (2005), under Assumptions 3 and 4 we have $\left\|\widehat{h} - h\right\|_{\infty,\omega} = o_p(1)$. By definition and Assumption 17,

$$\begin{aligned}
&\left\|h\left(\widehat{T}(\cdot),\cdot\right) - h\left(T(\cdot),\cdot\right)\right\|_{\infty,\omega} \\
=& \sup_{x\in\mathcal{X}_1,\beta\in\mathcal{B}}\left|h(\widehat{T}\left(x\right),\beta)\left[1+|\widehat{T}\left(x\right)|^2\right]^{-\omega/2} - h(T_o\left(x\right),\beta)\left[1+|T_o\left(x\right)|^2\right]^{-\omega/2}\right| \\
=& \sup_{\beta\in\mathcal{B}}\left\{\sup_{x\in\mathcal{X}_1}\left|h(\widehat{T}\left(x\right),\beta)\left[1+|\widehat{T}\left(x\right)|^2\right]^{-\omega/2} - h(T_o\left(x\right),\beta)\left[1+|T_o\left(x\right)|^2\right]^{-\omega/2}\right|\right\} \\
=& \sup_{\beta\in\mathcal{B}}\left\{\|h\left(\widehat{T}\left(\cdot\right),\beta\right) - h\left(T_o\left(\cdot\right),\beta\right)\|_{\infty,\omega}\right\} = o_p(1).
\end{aligned}$$

Let $\widehat{g}(\cdot,\beta) := \widehat{h}(\widehat{T}(\cdot),\beta)$ and $g(\cdot,\beta) := h(T_o(\cdot),\beta)$. Now following the proof of Proposition

A1 in Chen, Hong, and Tamer (2005), we obtain

$$\begin{aligned}\sup_{\beta\in\mathcal{B}}\|\widehat{g}(\cdot,\beta)-g(\cdot,\beta)\|_{2,1} &= \sup_{\beta\in\mathcal{B}}\sqrt{\int[\widehat{g}(x,\beta)-g(x,\beta)]^2 f_{X_1}(x)dx} \\ &\leq \sqrt{\left(\|\widehat{g}-g\|_{\infty,\omega}\right)^2\int\left(1+x'x\right)^{\omega}f_{X_1}(x)dx} \\ &= o_p(1) \quad \text{(by Assumption 4.2).}\end{aligned}$$

□

Proof of Theorem 1. All conditions of Lemma 5.2 in Newey (1994) are satisfied under Assumption 1-4.3 and Proposition 10, then $\widehat{\beta}-\beta_{o\mathcal{S}}=o_p(1)$. □

In the following we denote

$$\widehat{R}\equiv\frac{1}{n_1}\sum_{i=1}^{n_1}\left[\left(\widehat{h}(\widehat{T}(X_{1i}),\beta_{o\mathcal{S}})-\widehat{h}\left(T_o(X_{1i}),\beta_{o\mathcal{S}}\right)\right)-\left(h(\widehat{T}(X_{1i}),\beta_{o\mathcal{S}})-h\left(T_o(X_{1i}),\beta_{o\mathcal{S}}\right)\right)\right]I\left\{X_{1i}\in\mathcal{S}\right\}. \tag{A.1}$$

Lemma 4. Under Assumptions 3, 4, 17 and 9.1, we have

1.

$$\int\left\{\widehat{h}\left(T_o(x),\beta_{o\mathcal{S}}\right)I\left\{x\in\mathcal{S}\right\}-h\left(T_o(x),\beta_{o\mathcal{S}}\right)I\left\{x\in\mathcal{S}\right\}\right\}d\left[\widehat{F}_{X_1}(x)-F_{X_1}(x)\right]=o_p\left(n_1^{-1/2}\right), \tag{A.2}$$

$$\int\left\{\widehat{h}\left(\widehat{T}(x),\beta_{o\mathcal{S}}\right)I\left\{x\in\mathcal{S}\right\}-h\left(T_o(x),\beta_{o\mathcal{S}}\right)I\left\{x\in\mathcal{S}\right\}\right\}d\left[\widehat{F}_{X_1}\left(x\right)-F_{X_1}\left(x\right)\right]=o_p\left(n_1^{-1/2}\right), \tag{A.3}$$

$$\int\left\{h\left(\widehat{T}(x),\beta_{o\mathcal{S}}\right)I\left\{x\in\mathcal{S}\right\}-h\left(T_o(x),\beta_{o\mathcal{S}}\right)I\left\{x\in\mathcal{S}\right\}\right\}d\left[\widehat{F}_{X_1}\left(x\right)-F_{X_1}\left(x\right)\right]=o_p\left(n_1^{-1/2}\right). \tag{A.4}$$

2.

$$\widehat{R} = o_p\left(n_1^{-1/2}\right). \tag{A.5}$$

Proof.

To prove for (A.2), we let $N_{[]}\left(\varepsilon, \Lambda_c^\gamma\left(\mathcal{X}_1, \omega_1\right), \|\cdot\|_{2,1}\right)$ denote the $\|\cdot\|_{2,1}$-covering number of $\Lambda_c^\gamma\left(\mathcal{X}_1, \omega_1\right)$ with bracketing (i.e., the minimal number of N for which there exist ε-brackets $\left\{[l_j, u_j] : \|l_j - u_j\|_{2,1} \leq \varepsilon, \|l_j\|_{2,1}, \|u_j\|_{2,1} < \infty,\ j = 1, \ldots, N\right\}$ to cover $\Lambda_c^\gamma\left(\mathcal{X}_1, \omega_1\right)$. We also let $N\left(\varepsilon, \Lambda_c^\gamma\left(\mathcal{X}_1, \omega_1\right), \|\cdot\|_{\infty,\omega}\right)$ denote the $\|\cdot\|_{\infty,\omega}$-covering number of $\Lambda_c^\gamma\left(\mathcal{X}_1, \omega_1\right)$ (i.e., the minimal number of N for which there exist ε-balls $\left\{g : \|g - u_j\|_{\infty,\omega} \leq \varepsilon\right\}, j = 1, \ldots, N$ to cover $\Lambda_c^\gamma\left(\mathcal{X}_1, \omega_1\right)$). By Assumptions 3.1 and 9.1 with $\omega > \omega_1 + \gamma$, we have

$$\log N_{[]}\left(\delta, \Lambda_c^\gamma\left(\mathcal{X}_1, \omega_1\right), \|\cdot\|_{2,1}\right) \leq \log N\left(\delta, \Lambda_c^\gamma\left(\mathcal{X}_1, \omega_1\right), \|\cdot\|_{\infty,\omega}\right) \leq \text{const.} \left(\frac{c}{\delta}\right)^{d/\gamma},$$

see Chen et al. (1997) and Blundell et al. (2003). Thus under Assumption 9.1 ($\gamma > d/2$),

$$\int_0^1 \sqrt{\log N_{[]}\left(\delta, \Lambda_c^\gamma\left(\mathcal{X}_1, \omega_1\right), \|\cdot\|_{2,1}\right)}\, d\delta < \infty$$

and the class $\{\widetilde{g}\left(\cdot, \beta_{o\mathcal{S}}\right) : \widetilde{g}\left(\cdot, \beta_{o\mathcal{S}}\right) \in \Lambda_c^\gamma\left(\mathcal{X}_1, \omega_1\right)\}$ is a F_{X_1}-Donsker class. $\widehat{g}\left(x, \beta_{o\mathcal{S}}\right) \in \Lambda_c^\gamma\left(\mathcal{X}_1, \omega_1\right)$, $\gamma > d/2$ with probability approaching one as $n_0 \to \infty$ imply that

$$\sup_{\widetilde{g}(\cdot,\beta_{o\mathcal{S}}) \in \Lambda_c^\gamma(\mathcal{X}_1,\omega_1):\ \|\widetilde{g}(\cdot,\beta_{o\mathcal{S}}) - g(\cdot,\beta_{o\mathcal{S}})\|_{2,1} = o(1)} \left| \int \left[\widetilde{g}\left(x, \beta_{o\mathcal{S}}\right) - g\left(x, \beta_{o\mathcal{S}}\right)\right] d\left[\widehat{F}_{X_1}(x) - F_{X_1}(x)\right] \right| = o_p\left(n_1^{-1/2}\right),$$

this together with Assumption 3.2 established (A.2):

$$\int \left[\widehat{h}\left(T_o(x), \beta_{o\mathcal{S}}\right) I\left\{x \in \mathcal{S}\right\} - h\left(T_o(x), \beta_{o\mathcal{S}}\right) I\left\{x \in \mathcal{S}\right\}\right] d\left[\widehat{F}_{X_1}(x) - F_{X_1}(x)\right] = o_p\left(\frac{\sqrt{\lambda}}{\sqrt{n_0}}\right) = o_p\left(n_0^{-1/2}\right).$$

We can then establish equation (A.2) by Assumption 17.1, Assumption 4.5 and

$$\left\|\widehat{h}\left(T_o(\cdot), \beta_{o\mathcal{S}}\right) I\left\{x \in \mathcal{S}\right\} - g\left(\cdot, \beta_o\right)\right\|_{2,1} = \left\|\left(\widehat{h}\left(\cdot, \beta_{o\mathcal{S}}\right) - h\left(\cdot, \beta_{o\mathcal{S}}\right)\right) I\left\{x \in \mathcal{S}\right\}\right\|_{2,D=0} \leq$$

$\left\|\widehat{h}\left(\cdot,\beta_{o\mathcal{S}}\right)-h\left(\cdot,\beta_{o\mathcal{S}}\right)\right\|_{2,D=0}\left\|I\left\{\cdot\in\mathcal{S}\right\}\right\|_{2,D=0}$
$=o_p(1)$ by the existence of T_o and Monge-Ampere equation. We can establish (A.3) by Assumption 4.2 and Proposition 10. We can establish (A.4) by Assumption 17.1, Assumption 4.2 and Assumption 4.5.

The proof of second part (A.5) is shown as below:

$$
\begin{aligned}
&\frac{1}{n_p}\sum_{i=1}^{n_p}\left[\left(\widehat{h}\left(\widehat{T}(X_{1i}),\beta_{o\mathcal{S}}\right)-\widehat{h}\left(T_o(X_{1i}),\beta_o\right)\right)-\left(h\left(\widehat{T}(X_{1i}),\beta_o\right)-h\left(T_o(X_{1i}),\beta_{o\mathcal{S}}\right)\right)\right]I\left\{X_{1i}\in\mathcal{S}\right\}\\
&\quad=\int\left[\widehat{h}\left(\widehat{T}(x),\beta_{o\mathcal{S}}\right)-h\left(T_o(x),\beta_{o\mathcal{S}}\right)\right]I\left\{x\in\mathcal{S}\right\}d\widehat{F}_{X_1}(x)\\
&\qquad-\int\left[\widehat{h}\left(T_o(x),\beta_{o\mathcal{S}}\right)-h\left(T_o(x),\beta_o\right)\right]I\left\{x\in\mathcal{S}\right\}d\widehat{F}_{X_1}(x)\\
&\qquad-\int\left[h\left(\widehat{T}(x),\beta_{o\mathcal{S}}\right)-h\left(T_o(x),\beta_{o\mathcal{S}}\right)\right]I\left\{x\in\mathcal{S}\right\}d\widehat{F}_{X_1}(x)\\
&\quad=\int\left[\widehat{h}\left(\widehat{T}(x),\beta_{o\mathcal{S}}\right)-h\left(T_o(x),\beta_{o\mathcal{S}}\right)\right]I\left\{x\in\mathcal{S}\right\}dF_{X_1}(x)\\
&\qquad-\int\left[\widehat{h}\left(T_o(x),\beta_{o\mathcal{S}}\right)-h\left(T_o(x),\beta_o\right)\right]I\left\{x\in\mathcal{S}\right\}dF_{X_1}(x)\\
&\qquad-\int\left[h\left(\widehat{T}(x),\beta_{o\mathcal{S}}\right)-h\left(T_o(x),\beta_{o\mathcal{S}}\right)\right]I\left\{x\in\mathcal{S}\right\}dF_{X_1}(x)\\
&\qquad+\int\left[\widehat{h}\left(\widehat{T}(x),\beta_{o\mathcal{S}}\right)-h\left(T_o(x),\beta_{o\mathcal{S}}\right)\right]I\left\{x\in\mathcal{S}\right\}d\left[\widehat{F}_{X_1}(x)-F_{X_1}(x)\right]\\
&\qquad-\int\left[\widehat{h}\left(T_o(x),\beta_{o\mathcal{S}}\right)-h\left(T_o(x),\beta_{o\mathcal{S}}\right)\right]I\left\{x\in\mathcal{S}\right\}d\left[\widehat{F}_{X_1}(x)-F_{X_1}(x)\right]\\
&\qquad-\int\left[h\left(\widehat{T}(x),\beta_{o\mathcal{S}}\right)-h\left(T_o(x),\beta_{o\mathcal{S}}\right)\right]I\left\{x\in\mathcal{S}\right\}d\left[\widehat{F}_{X_1}(x)-F_{X_1}(x)\right]\\
&\quad=\int\left[\left\{\widehat{h}\left(\widehat{T}(x),\beta_{o\mathcal{S}}\right)-\widehat{h}\left(T_o(x),\beta_{o\mathcal{S}}\right)\right\}I\left\{x\in\mathcal{S}\right\}\right.\\
&\qquad\left.-\left\{h\left(\widehat{T}(x),\beta_{o\mathcal{S}}\right)-h\left(T_o(x),\beta_{o\mathcal{S}}\right)\right\}I\left\{x\in\mathcal{S}\right\}\right]dF_{X_1}(x)+o_p\left(n_1^{-1/2}\right)\\
&\quad=\int\left[\left\{\widehat{h}\left(\widehat{T}(x),\beta_{o\mathcal{S}}\right)-h\left(\widehat{T}(x),\beta_{o\mathcal{S}}\right)\right\}-\left\{\widehat{h}\left(T_o(x),\beta_{o\mathcal{S}}\right)-h\left(T_o(x),\beta_{o\mathcal{S}}\right)\right\}\right]I\left\{x\in\mathcal{S}\right\}dF_{X_p}(x)\\
&\qquad+o_p\left(n_1^{-1/2}\right)
\end{aligned}
$$

$$
\begin{aligned}
&= \int \frac{\partial \left\{ \widehat{h}\left(T_o(x), \beta_{o\mathcal{S}}\right) - h\left(T_o(x), \beta_{o\mathcal{S}}\right) \right\}}{\partial T} \left[\widehat{T}(x) - T_o(x) \right] I\left\{ x \in \mathcal{S} \right\} dF_{X_1}(x) + o_p\left(n_1^{-1/2}\right) \\
&\leq \left\| \frac{\partial \left\{ \widehat{h}\left(T_o(\cdot), \beta_{o\mathcal{S}}\right) - h\left(T_o(\cdot), \beta_{o\mathcal{S}}\right) \right\}}{\partial T} \right\|_{2,1} \cdot \left\| \widehat{T}(\cdot) - T_o(\cdot) \right\|_{2,1} \cdot \int I\{x \in \mathcal{S}\} dF_{X_1} + o_p\left(n_1^{-1/2}\right) \\
&= o_p\left(n_1^{-1/2}\right)
\end{aligned}
$$

where the last equality is obtained by Assumption 18. □

Proof of Theorem 2.

First we show the asymptotic linear representation

$$
\begin{aligned}
&\frac{1}{n_1} \sum_{i=1}^{n_1} \widehat{h}(\widehat{T}\left(X_{1i}\right), \beta_{o\mathcal{S}}) I\left\{ X_{1i} \in \mathcal{S} \right\} - E\left[h(T_o(X_1), \beta_{o\mathcal{S}}) I\left\{ X_1 \in \mathcal{S} \right\} \right] \\
&= \frac{1}{n_1} \sum_{i=1}^{n_1} \left[h\left(T_o(X_{1i}), \beta_{o\mathcal{S}}\right) I\left\{ X_{1i} \in \mathcal{S} \right\} + \varphi_1\left(X_{1i}\right) \right] - E\left[h(T_o(X_1), \beta_{o\mathcal{S}}) I\left\{ X_1 \in \mathcal{S} \right\} \right] \\
&+ \frac{1}{n_0} \sum_{j=1}^{n_0} \left[\varphi_0\left(X_{0j}\right) + \left(m\left(Y_{0j}, \beta_{o\mathcal{S}}\right) - h\left(X_{0j}, \beta_{o\mathcal{S}}\right) \right) I\left\{ X_{0j} \in T_0(\mathcal{S}) \right\} \right] + o_p\left(n_1^{-1/2}\right). \qquad \text{(A.6)}
\end{aligned}
$$

We do this by using the following decomposition:

$$
\begin{aligned}
&\frac{1}{n_1} \sum_{i=1}^{n_1} \widehat{h}(\widehat{T}(X_{1i}), \beta_{o\mathcal{S}}) I\left\{ X_{1i} \in \mathcal{S} \right\} - E\left[h(T_o(X_1), \beta_{o\mathcal{S}}) I\left\{ X_{1i} \in \mathcal{S} \right\} \right] \\
&= \frac{1}{n_1} \sum_{i=1}^{n_1} \left(\widehat{h}(\widehat{T}(X_{1i}), \beta_{o\mathcal{S}}) I\left\{ X_{1i} \in \mathcal{S} \right\} - h(\widehat{T}(X_{1i}), \beta_{o\mathcal{S}}) I\left\{ X_{1i} \in \mathcal{S} \right\} \right) \\
&\quad + \frac{1}{n_1} \sum_{i=1}^{n_1} \left(h(\widehat{T}(X_{1i}), \beta_{o\mathcal{S}}) I\left\{ X_{1i} \in \mathcal{S} \right\} - h\left(T_o(X_{1i}), \beta_{o\mathcal{S}}\right) I\left\{ X_{1i} \in \mathcal{S} \right\} \right) \\
&\quad + \frac{1}{n_1} \sum_{i=1}^{n_1} h\left(T_o(X_{1i}), \beta_{o\mathcal{S}}\right) I\left\{ X_{1i} \in \mathcal{S} \right\} - E\left[h(T_o(X_1), \beta_{o\mathcal{S}}) I\left\{ X_{1i} \in \mathcal{S} \right\} \right] \\
&= I_n + \frac{1}{n_1} \sum_{i=1}^{n_1} \left[h\left(T_o(X_{1i}), \beta_{o\mathcal{S}}\right) I\left\{ X_{1i} \in \mathcal{S} \right\} + \varphi_1\left(X_{1i}\right) \right] + \frac{1}{n_0} \sum_{j=1}^{n_0} \varphi_0\left(X_{0j}\right) + o_p\left(n_1^{-1/2}\right)
\end{aligned}
$$

(A.7)

where the last equality is due to Assumption 7 and the definition of

$$\begin{aligned} I_n &= \frac{1}{n_1}\sum_{i=1}^{n_1}\left(\widehat{h}(\widehat{T}(X_{1i}),\beta_{o\mathcal{S}})I\left\{X_{1i}\in\mathcal{S}\right\} - h(\widehat{T}(X_{1i}),\beta_{o\mathcal{S}})I\left\{X_{1i}\in\mathcal{S}\right\}\right) \\ &= \frac{1}{n_1}\sum_{i=1}^{n_1}\left(\widehat{h}(T_o(X_{1i}),\beta_{o\mathcal{S}})I\left\{X_{1i}\in\mathcal{S}\right\} - h(T_o(X_{1i}),\beta_{o\mathcal{S}})I\left\{X_{1i}\in\mathcal{S}\right\}\right) + \widehat{R}, \end{aligned}$$

where $\widehat{R}$ is defined in A.1. Since $\widehat{R} = o_p\left(n_1^{-1/2}\right)$ by Lemma 4 it suffices to show that

$$\begin{aligned} &\frac{1}{n_1}\sum_{i=1}^{n_1}\left(\widehat{h}(T_o(X_{1i}),\beta_{o\mathcal{S}})I\left\{X_{1i}\in\mathcal{S}\right\} - h(T_o(X_{1i}),\beta_{o\mathcal{S}})I\left\{X_{1i}\in\mathcal{S}\right\}\right) \\ = &\int\left[\widehat{h}\left(T_o(x),\beta_{o\mathcal{S}}\right) - h\left(T_o\left(x\right),\beta_{o\mathcal{S}}\right)\right]I\left\{x\in\mathcal{S}\right\}dF_{X_1}(x) \\ &+\int\left[\widehat{h}\left(T_o(x),\beta_{o\mathcal{S}}\right) - h\left(T_o\left(x\right),\beta_{o\mathcal{S}}\right)\right]I\left\{x\in\mathcal{S}\right\}d\left[\widehat{F}_{X_1}(x) - F_{X_1}(x)\right] \\ \equiv &\; II_n + III_n, \end{aligned}$$

where $\widehat{F}_{X_1}(x) = \frac{1}{n}\sum_{i=1}^{n_1} I\{X_{1i}\le x\}$ is the emprical distribution based on $X_{11},\ldots,X_{1n_1}$. We note that similar to that in Chen, Hong, and Tamer (2005) $III_n = o_p\left(n_1^{-1/2}\right)$ by stochastic equicontinuity. It suffices to show that

$$\begin{aligned} II_n &= \int\left[\widehat{h}\left(T_o(x),\beta_{o\mathcal{S}}\right) - h\left(T_o\left(x\right),\beta_{o\mathcal{S}}\right)\right]I\left\{x\in\mathcal{S}\right\}dF_{X_1}(x) \\ &= \frac{1}{n_0}\sum_{j=1}^{n_0}\left[\left(m\left(Y_{0j},\beta_{o\mathcal{S}}\right) - h\left(X_{0j},\beta_{o\mathcal{S}}\right)\right)I\left\{X_{0j}\in T_o(\mathcal{S})\right\}\right] + o_p\left(n_1^{-1/2}\right) \end{aligned}$$

Similar to Chen, Hong, and Tamer (2005), we can show that the last equality above is shown by establishing the following steps:

$$\begin{aligned} &\int_{\mathcal{X}_1}\left[\widehat{h}\left(T_o(x),\beta_{o\mathcal{S}}\right) - h\left(T_o(x),\beta_{o\mathcal{S}}\right)\right]I\left\{x\in\mathcal{S}\right\}f_{X_1}(x)dx \\ &= \int_{\mathcal{X}_0}\left[\widehat{h}\left(x,\beta_{o\mathcal{S}}\right) - h\left(x,\beta_{o\mathcal{S}}\right)\right]I\left\{x\in T_o(\mathcal{S})\right\}f_{X_0}(x)\,dx \qquad \text{(By Monge-Ampere equation)} \end{aligned}$$

$$
\begin{aligned}
&:= \int_{\mathcal{X}_0} \left[\widehat{h}\left(x, \beta_{o\mathcal{S}}\right) - h\left(x, \beta_{o\mathcal{S}}\right)\right] v^*(x) f_{X_0}(x)\, dx \\
&= \frac{1}{n_0} \sum_{j=1}^{n_0} \left[m\left(Y_{0j}, \beta_o\right) - h\left(X_{0j}, \beta_{o\mathcal{S}}\right)\right] v^*\left(X_{0j}\right) + o_p(1)
\end{aligned}
$$

with

$$
v^*(\cdot) \equiv I\left\{\cdot \in T_o(\mathcal{S})\right\},
$$

where the last equality above is shown below by following Chen, Hong, and Tamer (2005). The representer in our case is an indicator function. To be self-contained, we provide it here. Recall that $\widehat{h}\left(\cdot, \beta_{o\mathcal{S}}\right)$ is the sieve LS estimator of $h\left(\cdot, \beta_{o\mathcal{S}}\right) = E\left[m\left(Y_0, \beta_{o\mathcal{S}}\right) \mid X_0 = \cdot\right] \in \Lambda_c^\gamma\left(\mathcal{X}_0, \omega_1\right)$ using the control sample $\mathcal{X}_0$. That is, $\widehat{h}(\cdot, \beta_o)$ solves

$$
\inf_{\widetilde{h} \in \Psi_n} \frac{1}{n_0} \sum_{j=1}^{n_0} \left[m\left(Y_{0j}, \beta_{o\mathcal{S}}\right) - \widetilde{h}\left(X_{0j}, \beta_o\right)\right]^2,
$$

where Ψ_n increases with auxiliary sample size n_0, and is dense in $\Lambda_c^\gamma\left(\mathcal{X}_0, \omega_1\right)$ as $k_{nv} \to \infty$. In the following we denote the loss function

$$
L_n\left(\widetilde{h}\right) = -\frac{1}{2n_0} \sum_{j=1}^{n_0} \left[m\left(Y_{0j}, \beta_{o\mathcal{S}}\right) - \widetilde{h}\left(X_{0j}, \beta_o\right)\right]^2.
$$

We also let ε_n be any positive sequence with $\varepsilon_n = o\left(n_0^{-1/2}\right)$.
Then by definition, $\widehat{h}$ maximize $L_n\left(\widetilde{h}\right)$,

$$
\begin{aligned}
0 &\leq L_n(\widehat{h}) - L_n\left(\widehat{h} \pm \varepsilon_n \Pi_n v^*\right) \\
&= -\frac{1}{2n_0} \sum_{i=1}^{n_0} \left[m\left(Y_{0j}, \beta_o\right) - \widehat{h}\left(X_{0j}, \beta_{o\mathcal{S}}\right)\right]^2 + \frac{1}{2n_0} \sum_{i=1}^{n_0} \left[m\left(Y_{0j}, \beta_{o\mathcal{S}}\right) - \widehat{h}\left(X_{0j}, \beta_{o\mathcal{S}}\right) \mp \varepsilon_n \Pi_n v^*\left(X_{0j}\right)\right]^2 \\
&= -\frac{1}{2n_0} \sum_{i=1}^{n_0} \left[-\varepsilon_n^2 \left(\Pi_n v^*\left(X_{0j}\right)\right)^2 \pm 2\varepsilon_n \Pi_n v^*\left(X_{0j}\right) \left(m\left(Y_{0j}, \beta_{o\mathcal{S}}\right) - \widehat{h}\left(X_{0j}, \beta_{o\mathcal{S}}\right)\right)\right] \\
&= \varepsilon_n \frac{1}{2n_0} \sum_{i=1}^{n_0} \left(\Pi_n v^*\left(X_{0j}\right)\right)^2 \mp \frac{1}{n_0} \sum_{i=1}^{n_0} \Pi_n v^*\left(X_{0j}\right) \left[m\left(Y_{0j}, \beta_{o\mathcal{S}}\right) - \widehat{h}\left(X_{0j}, \beta_{o\mathcal{S}}\right)\right]
\end{aligned}
$$

$$
\begin{aligned}
&= \varepsilon_n \frac{1}{2n_0} \sum_{i=1}^{n_0} \left(\Pi_n v^*\left(X_{0j}\right)\right)^2 \mp E\left[\left(\Pi_n v^*\left(X_0\right) - v^*\left(X_0\right)\right)\left(\widehat{h}\left(X_0, \beta_{o\mathcal{S}}\right) - h\left(X_0, \beta_{o\mathcal{S}}\right)\right)\right] \\
&\quad \mp E\left[v^*\left(X_0\right)\left(h\left(X_0, \beta_o\right) - \widehat{h}\left(X_0, \beta_{o\mathcal{S}}\right)\right)\right] \\
&\quad \mp \frac{1}{n_0} \sum_{i=1}^{n_0} \Bigg\{ \left(\Pi_n v^*\left(X_{0j}\right) - v^*\left(X_{0j}\right)\right)\left(m\left(Y_{0j}, \beta_{o\mathcal{S}}\right) - h\left(X_{0j}, \beta_{o\mathcal{S}}\right)\right) \\
&\qquad\qquad - E\left[\left(\Pi_n v^*\left(X_0\right) - v^*\left(X_0\right)\right)\left(m\left(Y_0, \beta_{o\mathcal{S}}\right) - h\left(X_0, \beta_{o\mathcal{S}}\right)\right)\right] \Bigg\} \\
&\quad \mp \frac{1}{n_0} \sum_{i=1}^{n_0} \left\{ \Pi_n v^*\left(X_{0j}\right)\left(h\left(X_{0j}, \beta_{o\mathcal{S}}\right) - \widehat{h}\left(X_{0j}, \beta_{o\mathcal{S}}\right)\right) - E\left[\Pi_n v^*\left(X_0\right)\left(h\left(X_0, \beta_{o\mathcal{S}}\right) - \widehat{h}\left(X_0, \beta_{o\mathcal{S}}\right)\right)\right] \right\} \\
&\quad \mp \frac{1}{n_0} \sum_{i=1}^{n_0} \left\{ v^*\left(X_{0j}\right)\left(m\left(Y_{0j}, \beta_{o\mathcal{S}}\right) - h\left(X_{0j}, \beta_o\right)\right) - E\left[v^*\left(X_0\right)\left(m\left(Y_0, \beta_{o\mathcal{S}}\right) - h\left(X_0, \beta_{o\mathcal{S}}\right)\right)\right] \right\}.
\end{aligned}
$$

Then we establish the following (A.8)-(A.11)

$$
E\left[\left(\Pi_n v^*\left(X_0\right) - v^*\left(X_0\right)\right)\left(\widehat{h}\left(X_0, \beta_{o\mathcal{S}}\right) - h\left(X_0, \beta_{o\mathcal{S}}\right)\right)\right] = o_p\left(n_0^{-1/2}\right) \tag{A.8}
$$

$$
\begin{aligned}
&\frac{1}{n_0} \sum_{i=1}^{n_0} \Bigg\{ \left(\Pi_n v^*\left(X_{0j}\right) - v^*\left(X_{0j}\right)\right)\left(m\left(Y_{0j}, \beta_{o\mathcal{S}}\right) - h\left(X_{0j}, \beta_{o\mathcal{S}}\right)\right) \\
&\qquad - E\left[\left(\Pi_n v^*\left(X_0\right) - v^*\left(X_0\right)\right)\left(m\left(Y_0, \beta_{o\mathcal{S}}\right) - h\left(X_0, \beta_{o\mathcal{S}}\right)\right)\right] \Bigg\} = o_p\left(n_0^{-1/2}\right)
\end{aligned} \tag{A.9}
$$

$$
\frac{1}{n_0} \sum_{i=1}^{n_0} \left\{ \Pi_n v^*\left(X_{0j}\right)\left(h\left(X_{0j}, \beta_{o\mathcal{S}}\right) - \widehat{h}\left(X_{0j}, \beta_o\right)\right) - E\left[\Pi_n v^*\left(X_0\right)\left(h\left(X_0, \beta_{o\mathcal{S}}\right) - \widehat{h}\left(X_0, \beta_{o\mathcal{S}}\right)\right)\right] \right\} = o_p\left(n_0^{-1/2}\right. \tag{A.10}
$$

$$
\frac{1}{2n_0} \sum_{i=1}^{n_0} \left(\Pi_n v^*\left(X_{0j}\right)\right)^2 = O_p\left(1\right) \tag{A.11}
$$

Since the representer in our case is a constant function 1 when $X_0 \in T_o\left(\mathcal{S}\right)$ and is a constant function 0 otherwise, the proof is much simplified. We only need to show (A.10) as below.

By Assumption 3.1, Assumption 4 and Proposition A1 in Chen, Hong, and Tamer (2005), let

$\mathcal{F}_n = \left\{\widetilde{h}\left(\cdot, \beta_{o\mathcal{S}}\right) : \widetilde{h}\left(\cdot, \beta_{o\mathcal{S}}\right) \in \Lambda_c^\gamma\left(\mathcal{X}_0, \omega_1\right)\right\}$, we have $\log N_{[]}\left(\delta, \mathcal{F}_n, \|\cdot\|_{2,0}\right) \leq \text{const.} \cdot \left(\frac{c}{\delta}\right)^{d/\gamma}$ for

any $\delta > 0$. Applying Theorem 3 in Chen and Shen (1998) with $\delta_n = (n_0)^{-\gamma/(2\gamma+d)}$, we have

$$\sup_{\widetilde{h}\in\mathcal{F}_n:\left\|\widetilde{h}-h(\cdot,\beta_{o\mathcal{S}})\right\|_{2,0}\leq\delta_n}\left|\frac{1}{\sqrt{n_0}}\sum_{j=1}^{n_0}\left(\left[\widetilde{h}\left(X_{0j},\beta_{o\mathcal{S}}\right)-h\left(X_{0j},\beta_{o\mathcal{S}}\right)\right]I\left\{X_{0j}\in T_o(\mathcal{S})\right\}\right.\right.$$
$$\left.\left.-E\left[\left(\widetilde{h}\left(X_0,\beta_{o\mathcal{S}}\right)-h\left(X_0,\beta_{o\mathcal{S}}\right)\right)I\left\{X_0\in T_o(\mathcal{S})\right\}\right]\right)\right|=O_p\left((n_0)^{-\frac{2\gamma-d}{2(2\gamma+d)}}\right)=o_p(1).$$

Thus we get

$$\frac{1}{n_0}\sum_{i=1}^{n_0}\left[\left(h\left(X_{0j},\beta_o\right)-\widehat{h}\left(X_{0j},\beta_{o\mathcal{S}}\right)\right)I\left\{X_{0j}\in T_o(\mathcal{S})\right\}\right]-E\left[\left(\widehat{h}\left(X_0,\beta_{o\mathcal{S}}\right)-h\left(X_0,\beta_{o\mathcal{S}}\right)\right)I\left\{X_0\in T_o(\mathcal{S})\right\}\right]$$
$$=o_p\left(n_0^{-1/2}\right)$$

and

$$\begin{aligned}0&\leq L_n(\widehat{h})-L_n\left(\widehat{h}\pm\varepsilon_n\right)\\&=\mp\frac{1}{n_0}\sum_{i=1}^{n_0}\left[m\left(Y_{0j},\beta_o\right)-h\left(X_{0j},\beta_{o\mathcal{S}}\right)\right]I\left\{X_{0j}\in T_o(\mathcal{S})\right\}\\&\quad\pm E\left[\left(\widehat{h}\left(X_{0j},\beta_{o\mathcal{S}}\right)-h\left(X_{0j},\beta_{o\mathcal{S}}\right)\right)I\left\{X_0\in T_o(\mathcal{S})\right\}\right]+o_p\left(n_0^{-1/2}\right).\end{aligned}$$

As a result, it holds that

$$\begin{aligned}&\frac{1}{n_0}\sum_{i=1}^{n_0}\left[m\left(Y_{0j},\beta_{o\mathcal{S}}\right)-h\left(X_{0j},\beta_{o\mathcal{S}}\right)\right]I\left\{X_{0j}\in T_o(\mathcal{S})\right\}+o_p\left(n_0^{-1/2}\right)\\&=E\left[\left(\widehat{h}\left(X_{0j},\beta_{o\mathcal{S}}\right)-h\left(X_{0j},\beta_{o\mathcal{S}}\right)\right)I\left\{X_0\in T(\mathcal{S})\right\}\right].\end{aligned}$$

Thus we have

$$\begin{aligned}&\int\left[\widehat{h}\left(T_o\left(x\right),\beta_{o\mathcal{S}}\right)-h\left(T_o\left(x\right),\beta_{o\mathcal{S}}\right)\right]I\left\{x\in\mathcal{S}\right\}f_{X_1}(x)\,dx\\&=\int\left[\widehat{h}\left(x,\beta_{o\mathcal{S}}\right)-h\left(x,\beta_o\right)\right]I\left\{x\in T_o(\mathcal{S})\right\}f_{X_0}(x)\,dx\end{aligned}$$

$$= E\left[\left(\widehat{h}\left(X_0, \beta_{o\mathcal{S}}\right) - h\left(X_0, \beta_{o\mathcal{S}}\right)\right) I\left\{X_0 \in T(\mathcal{S})\right\}\right]$$
$$= \frac{1}{n_0}\sum_{i=1}^{n_0}\left[m\left(Y_{0j}, \beta_o\right) - h\left(X_{0j}, \beta_{o\mathcal{S}}\right)\right] I\left\{X_{0j} \in T_o(\mathcal{S})\right\} + o_p\left(n_0^{-1/2}\right).$$

Then we want to prove the second part of theorem 2.

Recall that $\widehat{\beta}$ is an estimator that solves

$$\frac{1}{n_1}\sum_{i=1}^{n_1}\widehat{h}(\widehat{T}(X_{1i}), \widehat{\beta}) I\left\{X_{1i} \in \mathcal{S}\right\} = 0,$$

Applying mean value expansion gives us,

$$\frac{1}{n_1}\sum_{i=1}^{n_1}\left(\widehat{h}(\widehat{T}(X_{1i}), \beta_{o\mathcal{S}}) I\left\{X_{1i} \in \mathcal{S}\right\} + \frac{\partial\widehat{h}(\widehat{T}(X_{1i}), \widetilde{\beta})}{\partial\beta'} I\left\{X_{1i} \in \mathcal{S}\right\}\left(\widehat{\beta} - \beta_{o\mathcal{S}}\right)\right) = 0,$$

where $\widetilde{\beta}$ is a convex combination of $\widehat{\beta}_l$ and $\beta_{o\mathcal{S}}$, and we can write

$$\widehat{\beta} - \beta_{o\mathcal{S}} = -\left(\frac{1}{n_1}\sum_{i=1}^{n_1}\frac{\partial\widehat{h}(\widehat{T}(X_{1i}), \widetilde{\beta})}{\partial\beta'} I\left\{X_{1i} \in \mathcal{S}\right\}\right)^{-1}\left(\frac{1}{n_1}\sum_{i=1}^{n_1}\widehat{h}(\widehat{T}(X_{1i}), \beta_{o\mathcal{S}}) I\left\{X_{1i} \in \mathcal{S}\right\}\right).$$

In the following we shall establish

$$\sup_{|\beta - \beta_{o\mathcal{S}}| = o(1)}\left|\frac{1}{n_1}\sum_{i=1}^{n_1}\frac{\partial\widehat{h}\left(\widehat{T}(X_{1i}), \beta\right)}{\partial\beta'} I\left\{X_{1i} \in \mathcal{S}\right\} - E\left[\frac{\partial h\left(T_o(X_1), \beta_{o\mathcal{S}}\right)}{\partial\beta'} I\left\{X_1 \in \mathcal{S}\right\}\right]\right| = o_p(1).$$

We can write the left hand side of the above equation as

$$\sup_{|\beta - \beta_{o\mathcal{S}}| = o(1)}\left|\frac{1}{n_1}\sum_{i=1}^{n_1}\frac{\partial\widehat{h}\left(\widehat{T}\left(X_{1i}\right), \beta\right)}{\partial\beta'} I\left\{X_{1i} \in \mathcal{S}\right\} - E\left[\frac{\partial h\left(T_o\left(X_1\right), \beta_{o\mathcal{S}}\right)}{\partial\beta'} I\left\{X_1 \in \mathcal{S}\right\}\right]\right|$$
$$\leq \left|\frac{1}{n_1}\sum_{i=1}^{n_1}\frac{\partial\widehat{h}\left(\widehat{T}\left(X_{1i}\right), \beta\right)}{\partial\beta'} I\left\{X_{1i} \in \mathcal{S}\right\} - \frac{1}{n_1}\sum_{i=1}^{n_1}\frac{\partial h\left(\widehat{T}\left(X_{1i}\right), \beta\right)}{\partial\beta'} I\left\{X_{1i} \in \mathcal{S}\right\}\right| \tag{A.12}$$

$$+\sup_{|\beta-\beta_{o\mathcal{S}}|=o(1)}\left|\frac{1}{n_1}\sum_{i=1}^{n_1}\frac{\partial h\left(\widehat{T}\left(X_{1i}\right),\beta\right)}{\partial\beta'}I\left\{X_{1i}\in\mathcal{S}\right\}-\frac{1}{n_1}\sum_{i=1}^{n_1}\frac{\partial h\left(\widehat{T}\left(X_{1i}\right),\beta_{o\mathcal{S}}\right)}{\partial\beta'}I\left\{X_{1i}\in\mathcal{S}\right\}\right| \tag{A.13}$$

$$+\left|\frac{1}{n_1}\sum_{i=1}^{n_1}\frac{\partial h\left(\widehat{T}\left(X_{1i}\right),\beta_{o\mathcal{S}}\right)}{\partial\beta'}I\left\{X_{1i}\in\mathcal{S}\right\}-E\left[\frac{\partial h\left(T\left(X_1\right),\beta_{o\mathcal{S}}\right)}{\partial\beta'}I\left\{X_1\in\mathcal{S}\right\}\right]\right|. \tag{A.14}$$

Equation (A.12) is $o_p\left(1\right)$ because by Assumption 8.4 we have equation (A.12) is bounded by

$$\begin{aligned}
&\frac{1}{n_1}\sum_{i=1}^{n_1}\left|\left(\frac{\partial\widehat{h}\left(\widehat{T}\left(X_{1i}\right),\beta\right)}{\partial\beta'}-\frac{\partial h\left(\widehat{T}\left(X_{1i}\right),\beta\right)}{\partial\beta'}\right)I\left\{X_{1i}\in\mathcal{S}\right\}\right| \\
&\leq\frac{1}{n_1}\sum_{i=1}^{n_1}b(\widehat{T}\left(X_{1i}\right)I\left\{X_{1i}\in\mathcal{S}\right\})\left[\left\|\widehat{h}-h\right\|_{\infty,\omega}\right]^{\epsilon} \\
&=\left\{E\left[b(\widehat{T}\left(X_1\right))I\left\{X_1\in\mathcal{S}\right\}\right]+o_p(1)\right\}\left[\left\|\widehat{h}-h\right\|_{\infty,\omega}\right]^{\epsilon}=o_p\left(1\right).
\end{aligned}$$

By Assumption 8.3, we have $\frac{\partial h(X_{1i},\beta)}{\partial\beta'}$ is continuous in $\beta\in\mathcal{B}$ with $|\beta-\beta_{o\mathcal{S}}|\leq\delta$. Thus we can bound equation (A.13) by

$$\frac{1}{n_1}\sum_{i=1}^{n_1}\sup_{|\beta-\beta_o|=o(1)}\left|\left(\frac{\partial h\left(\widehat{T}\left(X_{1i}\right),\beta\right)}{\partial\beta'}-\frac{\partial h\left(\widehat{T}\left(X_{1i}\right),\beta_{o\mathcal{S}}\right)}{\partial\beta'}\right)I\left\{X_{1i}\in\mathcal{S}\right\}\right|=o(1). \tag{A.15}$$

The last term (A.14) is $o_p\left(1\right)$ by Assumption 3.1 and Assumption 7

$$\begin{aligned}
&\left|\frac{\partial}{\partial\beta'}\left(\frac{1}{n_1}\sum_{i=1}^{n_1}h\left(\widehat{T}\left(X_{1i}\right),\beta_{o\mathcal{S}}\right)I\left\{X_{1i}\in\mathcal{S}\right\}-E\left[h\left(T_o\left(X_1\right),\beta_{o\mathcal{S}}\right)I\left\{X_1\in\mathcal{S}\right\}\right]\right)\right| \\
&=\left|\frac{\partial}{\partial\beta'}\left(E\left[\left(h\left(T(X_{1i}),\beta_{o\mathcal{S}}\right)I\left\{X_{1i}\in\mathcal{S}\right\}-E\left[h\left(T(X_1),\beta_{o\mathcal{S}}\right)I\left\{X_1\in\mathcal{S}\right\}\right]\right)+\psi_1\left(X_{1i},\beta_{o\mathcal{S}}\right)\right]\right. \\
&\quad\left.+E\left[\psi_0\left(X_0,\beta_{o\mathcal{S}}\right)\right]+o_p\left(1\right)\right)\Bigg|.
\end{aligned}$$

Using the linear expression (A.6) in Theorem 2 and Assumption 3 ($\lambda=\lim_{n_0\to\infty}\left(n_0/n_1\right)\in$

$[0,\infty)$), we have

$$\begin{aligned}
&\sqrt{n_0}\frac{1}{n_1}\sum_{i=1}^{n_1}\widehat{h}(\widehat{T}(X_{1i}),\beta_o)I\left\{X_{1i}\in\mathcal{S}\right\}\\
&=\sqrt{\frac{n_0}{n_1}}\frac{1}{\sqrt{n_1}}\sum_{i=1}^{n_1}\left[h\left(T_o(X_{1i}),\beta_{o\mathcal{S}}\right)I\left\{X_{1i}\in\mathcal{S}\right\}+\psi_1\left(X_{1i},\beta_{o\mathcal{S}}\right)\right]\\
&\quad+\frac{1}{\sqrt{n_0}}\sum_{j=1}^{n_0}\left[\psi_0\left(X_{0j},\beta_{o\mathcal{S}}\right)+\left(m\left(Y_{0j},\beta_{o\mathcal{S}}\right)-h\left(X_{0j},\beta_{o\mathcal{S}}\right)\right)I\left\{X_{0j}\in T_o(\mathcal{S})\right\}\right]+o_p\left(1\right),
\end{aligned}$$

and by Assumption 4.3, Assumption 7 and Assumption 8,

$$\begin{aligned}
&\frac{1}{\sqrt{n_1}}\sum_{j=1}^{n_1}\left(h\left(T_o(X_{1i}),\beta_{o\mathcal{S}}\right)I\left\{X_{1i}\in\mathcal{S}\right\}+\varphi_1\left(X_{1i}\right)\right)\xrightarrow{d}\mathcal{N}\left(0,\Omega_1\right),\\
&\frac{1}{\sqrt{n_0}}\sum_{j=1}^{n_0}\left[\varphi_0\left(X_{0j}\right)+\left(m\left(Y_{0j},\beta_{o\mathcal{S}}\right)-h\left(X_{0j},\beta_{o\mathcal{S}}\right)\right)I\left\{X_{0j}\in T_o(\mathcal{S})\right\}\right]\xrightarrow{d}\mathcal{N}\left(0,\Omega_0\right),
\end{aligned}$$

where $\Omega_1=E\left[h\left(T_o(X_1),\beta_{o\mathcal{S}}\right)I\left\{X_1\in\mathcal{S}\right\}+\varphi_1\left(X_1\right)\right]^2$,
$\Omega_0=E\left[\varphi_0\left(X_0\right)+\left(m\left(Y_{0j},\beta_{o\mathcal{S}}\right)-h\left(X_{0j},\beta_{o\mathcal{S}}\right)\right)I\left\{X_{0j}\in T_o(\mathcal{S})\right\}\right]^2$. □

Proofs of Corollaries 1 and 2.

Since $\widehat{\tau}-\tau_{o\mathcal{S}}=[\widehat{\kappa}-\kappa_{o\mathcal{S}}]-[\widehat{\beta}-\beta_{o\mathcal{S}}]$, the proofs of both Corollaries are straightforward. By Assumption 2.2 and 3, sample moment converges to the true moment almost surely, we have $\widehat{\kappa}-\kappa_{o\mathcal{S}}=o_p\left(1\right)$ and thus $\widehat{\tau}-\tau_{o\mathcal{S}}=o_p\left(1\right)$.

Next under mild conditions we have:

$$\widehat{\kappa}-\kappa_{o\mathcal{S}}=-\left[M+o_p\left(1\right)\right]^{-1}\left[\frac{1}{n_1}\sum_{i=1}^{n_1}m\left(Y_{1i};\kappa_{o\mathcal{S}}\right)I\left\{X_{1i}\in\mathcal{S}\right\}\right],$$

Thus

$$\sqrt{n_1}\left(\widehat{\tau}-\tau_{o\mathcal{S}}\right)=G^{-1}\left\{\frac{1}{\sqrt{n_1}}\sum_{i=1}^{n_1}\left[h\left(T_o(X_{1i}),\beta_{o\mathcal{S}}\right)I\left\{X_{1i}\in\mathcal{S}\right\}+\varphi_1\left(X_{1i}\right)-GM^{-1}m\left(Y_{1i};\kappa_{o\mathcal{S}}\right)I\left\{X_{1i}\in\mathcal{S}\right\}\right]\right.$$

$$+\frac{\sqrt{n_1}}{n_0}\sum_{j=1}^{n_0}\left[\varphi_0\left(X_{0j}\right)+\left(m\left(Y_{0j};\beta_{o\mathcal{S}}\right)-h\left(X_{0j},\beta_{o\mathcal{S}}\right)\right)I\left\{X_{0j}\in T_o(\mathcal{S})\right\}\right]\Bigg\}+o_p\left(1\right)$$

$$\xrightarrow{d}\mathcal{N}(0,V_\tau)$$

where $V_\tau=G^{-2}\left[\Omega_{\tau,1}+\lambda\Omega_0\right]$ and,

$$\Omega_{\tau,1}=E\left[-\frac{G}{M}m\left(Y_1;\kappa_{o\mathcal{S}}\right)I\left\{X_1\in\mathcal{S}\right\}+h\left(T_o(X_1),\beta_{o\mathcal{S}}\right)I\left\{X_1\in\mathcal{S}\right\}+\psi_1\left(X_1,\beta_{o\mathcal{S}}\right)\right]^2,$$

$$\Omega_0=E\left[\varphi_0\left(X_0\right)+\left(m\left(Y_0;\beta_{o\mathcal{S}}\right)-h\left(X_0,\beta_{o\mathcal{S}}\right)\right)I\left\{X_0\in T_o(\mathcal{S})\right\}\right]^2.$$

□

A.2 Proofs in Section 1.5.1

Before proving the theorem, we first introduce the following lemma.

Lemma 5. Suppose $h(x,\beta_{o\mathcal{S}})\in\Lambda^\gamma\left(\mathcal{X},\omega_1\right)$ and $f_X\in[\underline{f}_X,\overline{f}_X]$ is bounded away from zero. For all $0\leq\eta\leq 5/7$,

$$\sup_q n^\eta\cdot\left|h\left(\widehat{F}_X^{-1}(q),\beta_{o\mathcal{S}}\right)-h\left(F_X^{-1}(q),\beta_{o\mathcal{S}}\right)+\frac{h_1\left(F_X^{-1}(q),\beta_{o\mathcal{S}}\right)}{f_X\left(F_X^{-1}(q)\right)}\left(\widehat{F}_X\left(F_X^{-1}(q)\right)-q\right)\right|\xrightarrow{p}0.$$

Remark. This lemma is similar to Lemma A.6 in Athey and Imbens (2006), the difference is we're proving for $h(\cdot,\beta_{o\mathcal{S}})$.

Proof of Proposition 1.

Under Condition 3, we have

$$\begin{aligned}\sup_{x\in\mathcal{X}_1}\left|\widehat{T}\left(x\right)-T_o\left(x\right)\right|&=\sup_{x\in\mathcal{X}_1}\left|\widehat{F}_{X_0}^{-1}\left(\widehat{F}_{X_1}\left(x\right)\right)-F_{X_0}^{-1}\left(F_{X_1}\left(x\right)\right)\right|\\&=\sup_{x\in\mathcal{X}_1}\left|\widehat{F}_{X_0}^{-1}\left(\widehat{F}_{X_1}\left(x\right)\right)-\widehat{F}_{X_0}^{-1}\left(F_{X_1}\left(x\right)\right)+\widehat{F}_{X_0}^{-1}\left(F_{X_1}\left(x\right)\right)-F_{X_0}^{-1}\left(F_{X_1}\left(x\right)\right)\right|\\&\leq\sup_{x\in\mathcal{X}_1}\left|\widehat{F}_{X_0}^{-1}\left(\widehat{F}_{X_1}\left(x\right)\right)-\widehat{F}_{X_0}^{-1}\left(F_{X_1}\left(x\right)\right)\right|+\sup_{x\in\mathcal{X}_1}\left|\widehat{F}_{X_0}^{-1}\left(F_{X_1}\left(x\right)\right)-F_{X_0}^{-1}\left(F_{X_1}\left(x\right)\right)\right|\end{aligned}$$

$$= O_p\left(n^{-1/2}\right).$$

The last equation holds by applying Lemma A.2 and Lemma A.3 in Athey and Imbens (2006). By Assumption 11, the first part of Proposition 1 is proved.

Now let's prove the second part of the proposition. We can get the convergence rate of $\left\|\widehat{h}\left(\cdot,\beta_{o\mathcal{S}}\right)-h\left(\cdot,\beta_{o\mathcal{S}}\right)\right\|_{2,0}$ by Proposition A1.(ii) in Chen, Hong, and Tamer (2005), all the assumption of the Proposition A1.(ii) are satisfied given our Assumption 3.2, 4 and 9. Using Monge-Ampere equation, we obtain

$$\left\|\widehat{h}\left(T_o\left(\cdot\right),\beta_{o\mathcal{S}}\right)-h\left(T_o\left(\cdot\right),\beta_{o\mathcal{S}}\right)\right\|_{2,1}=\left\|\widehat{h}\left(\cdot,\beta_{o\mathcal{S}}\right)-h\left(\cdot,\beta_{o\mathcal{S}}\right)\right\|_{2,0}=O_p\left(\sqrt{\frac{k_{n_0}}{n_0}}+\left(k_{n_0}\right)^{-\gamma/d_X}\right)$$
$$=O_p\left(n^{-\gamma/(2\gamma+d_X)}\right).$$

By Taylor expansion, we have

$$h\left(T_o\left(x\right)+\delta,\beta_{o\mathcal{S}}\right)=h\left(T_o\left(x\right),\beta_{o\mathcal{S}}\right)+h_1\left(T_o\left(x\right),\beta_{o\mathcal{S}}\right)\delta+\frac{1}{2}h_{11}\left(T_o\left(x\right),\beta_{o\mathcal{S}}\right)\delta^2+o\left(\delta^2\right),$$

and similarly,

$$\widehat{h}\left(T_o\left(x\right)+\delta,\beta_{o\mathcal{S}}\right)=\widehat{h}\left(T_o\left(x\right),\beta_{o\mathcal{S}}\right)+\widehat{h}_1\left(T_o\left(x\right),\beta_{o\mathcal{S}}\right)\delta+\frac{1}{2}\widehat{h}_{11}\left(T_o\left(x\right),\beta_{o\mathcal{S}}\right)\delta^2+o\left(\delta^2\right),$$

where $h_1\left(x,\beta_{o\mathcal{S}}\right)=\frac{\partial}{\partial x}h\left(x,\beta_{o\mathcal{S}}\right), h_{11}\left(x,\beta_{o\mathcal{S}}\right)=\frac{\partial^2}{\partial x^2}h\left(x,\beta_{o\mathcal{S}}\right), \widehat{h}_1\left(x,\beta_{o\mathcal{S}}\right)=\frac{\partial}{\partial x}\widehat{h}\left(x,\beta_{o\mathcal{S}}\right), \widehat{h}_{11}\left(x,\beta_{o\mathcal{S}}\right)=\frac{\partial^2}{\partial x^2}\widehat{h}\left(x,\beta_{o\mathcal{S}}\right).$

For $\delta>0$ we have

$$\begin{aligned}&\left\|\widehat{h}_1\left(T_o\left(x\right),\beta_{o\mathcal{S}}\right)-h_1\left(T_o\left(x\right),\beta_{o\mathcal{S}}\right)\right\|_{2,1}\\&\leq\left\|\frac{1}{\delta}\left(\widehat{h}\left(T_o\left(x\right)+\delta,\beta_{o\mathcal{S}}\right)-h\left(T_o\left(x\right)+\delta,\beta_{o\mathcal{S}}\right)\right)-\frac{1}{\delta}\left(\widehat{h}\left(T_o\left(x\right),\beta_{o\mathcal{S}}\right)-h\left(T_o\left(x\right),\beta_{o\mathcal{S}}\right)\right)\right.\\&\quad\left.-\frac{\delta}{2}\left(\widehat{h}_{11}\left(T_o\left(x\right),\beta_{o\mathcal{S}}\right)-h_{11}\left(T_o\left(x\right),\beta_{o\mathcal{S}}\right)\right)\right\|_{2,1}+o_p\left(\delta\right)\end{aligned}$$

$$
\begin{aligned}
&\leq \left\| \frac{1}{\delta} \left(\widehat{h} \left(T_o \left(x \right) + \delta, \beta_{o\mathcal{S}} \right) - h \left(T_o \left(x \right) + \delta, \beta_{o\mathcal{S}} \right) \right) \right\|_{2,1} \\
&\quad + \left\| \frac{1}{\delta} \left(\widehat{h} \left(T_o \left(x \right), \beta_{o\mathcal{S}} \right) - h \left(T_o \left(x \right), \beta_{o\mathcal{S}} \right) \right) \right\|_{2,1} \\
&\quad + \left\| \frac{\delta}{2} \left(\widehat{h}_{11} \left(T_o \left(x \right), \beta_{o\mathcal{S}} \right) - h_{11} \left(T_o \left(x \right), \beta_{o\mathcal{S}} \right) \right) \right\|_{2,1} + o_p \left(\delta \right) \\
&= O_p \left(n^{-\gamma/(2\gamma + d_X)} / \delta \right) + O_p \left(\delta \right) + o_p \left(\delta \right).
\end{aligned}
$$

Choose $\delta = O_p \left(n^{-\gamma/(4\gamma + 2d_X)} \right)$, we have

$$
\left\| \widehat{h}_1 \left(T_o \left(x \right), \beta_{o\mathcal{S}} \right) - h_1 \left(T_o \left(x \right), \beta_{o\mathcal{S}} \right) \right\|_{2,1} = O_p \left(n^{-\gamma/(4\gamma + 2d_X)} \right). \tag{A.16}
$$

Hence, for any $\gamma > 0$, we have

$$
\begin{aligned}
&\left\| \frac{\partial \left\{ \widehat{h} \left(T_o(\cdot), \beta_{o\mathcal{S}} \right) - h \left(T_o(\cdot), \beta_{o\mathcal{S}} \right) \right\}}{\partial T} \right\|_{2,1} \cdot \left\| \widehat{T}(\cdot) - T_o(\cdot) \right\|_{2,1} \\
&\leq \left\| \frac{\partial \left\{ \widehat{h} \left(T_o(\cdot), \beta_{o\mathcal{S}} \right) - h \left(T_o(\cdot), \beta_{o\mathcal{S}} \right) \right\}}{\partial T} \right\|_{2,1} \cdot \sup_{x \in \mathcal{X}_1} \left| \widehat{T} \left(x \right) - T_o \left(x \right) \right| = o_p \left(n_1^{-1/2} \right).
\end{aligned}
$$

□

Proof of Proposition 2.

We can write

$$
\begin{aligned}
&\frac{1}{n_1} \sum_{i=1}^{n_1} h \left(\widehat{T} \left(X_{1i} \right), \beta_{o\mathcal{S}} \right) I \left\{ X_{1i} \in \mathcal{S} \right\} - E \left[h \left(T \left(X_1 \right), \beta_{o\mathcal{S}} \right) I \left\{ X_1 \in \mathcal{S} \right\} \right] \\
&= \frac{1}{n_1} \sum_{i=1}^{n_1} h \left(\widehat{F}_{X_0}^{-1} \left(\widehat{F}_{X_1}(X_{1i}) \right), \beta_{o\mathcal{S}} \right) I \left\{ X_{1i} \in \mathcal{S} \right\} - E \left[h \left(F_{X_0}^{-1} \left(F_{X_1}(X_1) \right), \beta_{o\mathcal{S}} \right) I \left\{ X_1 \in \mathcal{S} \right\} \right] \\
&= \frac{1}{n_1} \sum_{i=1}^{n_1} h \left(\widehat{F}_{X_0}^{-1} \left(\widehat{F}_{X_1}(X_{1i}) \right), \beta_{o\mathcal{S}} \right) I \left\{ X_{1i} \in \mathcal{S} \right\} - \frac{1}{n_1} \sum_{i=1}^{n_1} h \left(F_{X_0}^{-1} \left(\widehat{F}_{X_1}(X_{1i}) \right), \beta_o \right) I \left\{ X_{1i} \in \mathcal{S} \right\}
\end{aligned} \tag{A.17}
$$

$$+\frac{1}{n_1}\sum_{i=1}^{n_1} h\left(F_{X_0}^{-1}\left(\widehat{F}_{X_1}(X_{1i})\right),\beta_{o\mathcal{S}}\right) I\left\{X_{1i}\in\mathcal{S}\right\} - \frac{1}{n_1}\sum_{i=1}^{n_1} h\left(F_{X_0}^{-1}\left(F_{X_1}(X_{1i})\right),\beta_{o\mathcal{S}}\right) I\left\{X_{1i}\in\mathcal{S}\right\} \tag{A.18}$$

$$+\frac{1}{n_1}\sum_{i=1}^{n_1} h\left(F_{X_0}^{-1}\left(F_{X_1}(X_{1i})\right),\beta_{o\mathcal{S}}\right) I\left\{X_{1i}\in\mathcal{S}\right\} - E\left[h\left(F_{X_0}^{-1}\left(F_{X_1}(X_1)\right),\beta_{o\mathcal{S}}\right) I\left\{X_1\in\mathcal{S}\right\}\right]. \tag{A.19}$$

Recall the notation

$$Q(x_0, X_1, \beta_{o\mathcal{S}}) = -\frac{h_1\left(T_o\left(X_1\right),\beta_o\right)}{f_{X_0}\left(T_o\left(X_1\right)\right)}\left[I\left\{x_0 \le T_o\left(X_1\right)\right\} - F_{X_1}(X_1)\right] I\left\{X_1\in\mathcal{S}\right\} \quad x_0\in\mathcal{X}_0,$$

$$q(x_0,\beta_{o\mathcal{S}}) = E\left[Q\left(x_0, X_1, \beta_{o\mathcal{S}}\right)\right] \quad x_0\in\mathcal{X}_0.$$

and $T_o(x_1) = F_{X_0}^{-1}\left(F_{X_1}(x_1)\right)$ in the univariate case. First consider (A.17), we have

$$\begin{aligned}
&\frac{1}{n_1}\sum_{i=1}^{n_1} h\left(\widehat{F}_{X_0}^{-1}\left(\widehat{F}_{X_1}\left(X_{1i}\right)\right),\beta_{o\mathcal{S}}\right) I\left\{X_{1i}\in\mathcal{S}\right\} - \frac{1}{n_1}\sum_{i=1}^{n_1} h\left(F_{X_0}^{-1}\left(\widehat{F}_{X_1}\left(X_{1i}\right)\right),\beta_o\right) I\left\{X_{1i}\in\mathcal{S}\right\} \\
&= \frac{1}{n_1}\sum_{i=1}^{n_1} h\left(\widehat{F}_{X_0}^{-1}\left(\widehat{F}_{X_1}\left(X_{1i}\right)\right),\beta_{o\mathcal{S}}\right) I\left\{X_{1i}\in\mathcal{S}\right\} - \frac{1}{n_1}\sum_{i=1}^{n_1} h\left(F_{X_0}^{-1}\left(\widehat{F}_{X_1}\left(X_{1i}\right)\right),\beta_o\right) I\left\{X_{1i}\in\mathcal{S}\right\} \\
&\quad + \frac{1}{n_1}\frac{1}{n_0}\sum_{i=1}^{n_1}\sum_{j=1}^{n_0} \frac{h_1\left(F_{X_0}^{-1}\left(\widehat{F}_{X_1}\left(X_{1i}\right)\right),\beta_{o\mathcal{S}}\right)}{f_0\left(F_{X_0}^{-1}\left(\widehat{F}_{X_1}\left(X_{1i}\right)\right)\right)}\left(I\left\{X_{0j}\le F_{X_0}^{-1}\left(\widehat{F}_{X_1}\left(X_{1i}\right)\right)\right\} - \widehat{F}_{X_1}\left(X_{1i}\right)\right) I\left\{X_{1i}\in\mathcal{S}\right\}
\end{aligned} \tag{A.20}$$

$$- \frac{1}{n_1}\frac{1}{n_0}\sum_{i=1}^{n_1}\sum_{j=1}^{n_0} \frac{h_1\left(F_{X_0}^{-1}\left(\widehat{F}_{X_1}\left(X_{1i}\right)\right),\beta_{o\mathcal{S}}\right)}{f_0\left(F_{X_0}^{-1}\left(\widehat{F}_{X_1}\left(X_{1i}\right)\right)\right)}\left(I\left\{X_{0j}\le F_{X_0}^{-1}\left(\widehat{F}_{X_1}\left(X_{1i}\right)\right)\right\} - \widehat{F}_{X_1}\left(X_{1i}\right)\right) I\left\{X_{1i}\in\mathcal{S}\right\}$$

$$- \frac{1}{n_0}\frac{1}{n_1}\sum_{j=1}^{n_0}\sum_{i=1}^{n_1} Q\left(X_{0j}, X_{1i}, \beta_{o\mathcal{S}}\right) \tag{A.21}$$

$$+ \frac{1}{n_0}\sum_{j=1}^{n_0} q\left(X_{0j},\beta_{o\mathcal{S}}\right) + o_p\left(n_1^{-1/2}\right). \tag{A.22}$$

The absolute value of (A.20) can be bounded by

$$
\begin{aligned}
&\frac{1}{n_1}\sum_{i=1}^{n_1}\left|h\left(\widehat{F}_{X_0}^{-1}\left(\widehat{F}_{X_1}\left(X_{1i}\right)\right),\beta_{o\mathcal{S}}\right)I\left\{X_{1i}\in\mathcal{S}\right\}-h\left(F_{X_0}^{-1}\left(\widehat{F}_{X_1}\left(X_{1i}\right)\right),\beta_{o\mathcal{S}}\right)I\left\{X_{1i}\in\mathcal{S}\right\}\right.\\
&\qquad\left.+\frac{1}{n_0}\sum_{j=1}^{n_0}\frac{h_1\left(F_{X_0}^{-1}\left(\widehat{F}_{X_1}\left(X_{1i}\right)\right),\beta_o\right)}{f_0\left(F_{X_0}^{-1}\left(\widehat{F}_{X_1}\left(X_{1i}\right)\right)\right)}\left(I\left\{X_{0j}\le F_{X_0}^{-1}\left(\widehat{F}_{X_1}\left(X_{1i}\right)\right)\right\}-\widehat{F}_{X_1}\left(X_{1i}\right)\right)I\left\{X_{1i}\in\mathcal{S}\right\}\right|\\
&\le\sup_q\left|\left[h\left(\widehat{F}_{X_0}^{-1}(q),\beta_o\right)I\left\{q\in\widehat{F}_{X_1}\left(\mathcal{S}\right)\right\}-h\left(F_{X_0}^{-1}(q),\beta_{o\mathcal{S}}\right)I\left\{q\in\widehat{F}_{X_1}\left(\mathcal{S}\right)\right\}\right.\right.\\
&\qquad\left.\left.+\frac{1}{n_0}\sum_{j=1}^{n_0}\frac{h_1\left(F_{X_0}^{-1}(q),\beta_{o\mathcal{S}}\right)}{f_0\left(F_{X_0}^{-1}(q)\right)}\left(I\left\{X_{0j}\le F_{X_0}^{-1}\left(q\right)\right\}-q\right)I\left\{q\in\widehat{F}_{X_1}\left(\mathcal{S}\right)\right\}\right]\right|\\
&=\sup_q\left|h\left(\widehat{F}_{X_0}^{-1}(q),\beta_{o\mathcal{S}}\right)-h\left(F_{X_0}^{-1}(q),\beta_{o\mathcal{S}}\right)+\frac{h_1\left(F_{X_0}^{-1}(q),\beta_{o\mathcal{S}}\right)}{f_{X_0}\left(F_{X_0}^{-1}(q)\right)}\left(\widehat{F}_{X_0}\left(F_{X_0}^{-1}(q)\right)-q\right)\right|I\left\{q\in\widehat{F}_{X1}\left(\mathcal{S}\right)\right\}\\
&\le\sup_q\left|h\left(\widehat{F}_{X_0}^{-1}(q),\beta_{o\mathcal{S}}\right)-h\left(F_{X_0}^{-1}(q),\beta_{o\mathcal{S}}\right)+\frac{h_1\left(F_{X_0}^{-1}(q),\beta_{o\mathcal{S}}\right)}{f_{X_0}\left(F_{X_0}^{-1}(q)\right)}\left(\widehat{F}_{X_0}\left(F_{X_0}^{-1}(q)\right)-q\right)\right|\\
&=o_p\left(n_0^{-1/2}\right).\qquad\text{(By Lemma 5)}
\end{aligned}
$$

Next, consider (A.21), the absolute value can be bounded by

$$
\begin{aligned}
&\left|\frac{1}{n_1}\frac{1}{n_0}\sum_{i=1}^{n_1}\sum_{j=1}^{n_0}\frac{h_1\left(F_{X_0}^{-1}\left(\widehat{F}_{X_1}\left(X_{1i}\right)\right),\beta_{o\mathcal{S}}\right)}{f_{X_0}\left(F_{X_0}^{-1}\left(\widehat{F}_{X_1}\left(X_{1i}\right)\right)\right)}\left(I\left\{X_{0j}\le F_{X_0}^{-1}\left(\widehat{F}_{X_1}\left(X_{1i}\right)\right)\right\}-\widehat{F}_{X_1}\left(X_{1i}\right)\right)I\left\{X_{1i}\in\mathcal{S}\right\}\right.\\
&\qquad\left.-\frac{1}{n_1}\frac{1}{n_0}\sum_{i=1}^{n_1}\sum_{j=1}^{n_0}\frac{h_1\left(F_{X_0}^{-1}\left(F_{X_1}\left(X_{1i}\right)\right),\beta_{o\mathcal{S}}\right)}{f_{X_0}\left(F_{X_0}^{-1}\left(F_{X_1}\left(X_{1i}\right)\right)\right)}\left(I\left\{X_{0j}\le F_{X_0}^{-1}\left(F_{X_1}\left(X_{1i}\right)\right)\right\}-F_{X_1}\left(X_{1i}\right)\right)I\left\{X_{1i}\in\mathcal{S}\right\}\right|\\
&=\left|\frac{1}{n_1}\sum_{i=1}^{n_1}\frac{h_1\left(F_{X_0}^{-1}\left(\widehat{F}_{X_1}\left(X_{1i}\right)\right),\beta_o\right)}{f_{X_0}\left(F_{X_0}^{-1}\left(\widehat{F}_{X_1}\left(X_{1i}\right)\right)\right)}\left(\widehat{F}_{X_0}\left(F_{X_0}^{-1}\left(\widehat{F}_{X_1}\left(X_{1i}\right)\right)\right)-\widehat{F}_{X_1}\left(X_{1i}\right)\right)I\left\{X_{1i}\in\mathcal{S}\right\}\right.\\
&\qquad\left.-\frac{1}{n_1}\sum_{i=1}^{n_1}\frac{h_1\left(F_{X_0}^{-1}\left(F_{X_1}\left(X_{1i}\right)\right),\beta_o\right)}{f_{X_0}\left(F_{X_0}^{-1}\left(F_{X_1}\left(X_{1i}\right)\right)\right)}\left(\widehat{F}_{X_0}\left(F_{X_0}^{-1}\left(F_{X_1}\left(X_{1i}\right)\right)\right)-F_{X_1}\left(X_{1i}\right)\right)I\left\{X_{1i}\in\mathcal{S}\right\}\right|
\end{aligned}
$$

$$\leq \left| \frac{1}{n_1} \sum_{i=1}^{n_1} \frac{h_1\left(F_{X_0}^{-1}\left(\widehat{F}_{X_1}\left(X_{1i}\right)\right), \beta_o\right)}{f_{X_0}\left(F_{X_0}^{-1}\left(\widehat{F}_{X_1}\left(X_{1i}\right)\right)\right)} \left(\widehat{F}_{X_0}\left(F_{X_0}^{-1}\left(\widehat{F}_{X_1}\left(X_{1i}\right)\right)\right) - \widehat{F}_{X_1}\left(X_{1i}\right)\right) I\left\{X_{1i} \in \mathcal{S}\right\} \right. \tag{A.23}$$

$$\left. - \frac{1}{n_1} \sum_{i=1}^{n_1} \frac{h_1\left(F_{X_0}^{-1}\left(\widehat{F}_{X_1}\left(X_{1i}\right)\right), \beta_o\right)}{f_{X_0}\left(F_{X_0}^{-1}\left(\widehat{F}_{X_1}\left(X_{1i}\right)\right)\right)} \left(\widehat{F}_{X_0}\left(F_{X_0}^{-1}\left(F_{X_1}\left(X_{1i}\right)\right)\right) - F_{X_1}\left(X_{1i}\right)\right) I\left\{X_{1i} \in \mathcal{S}\right\} \right|$$

$$+ \left| \frac{1}{n_1} \sum_{i=1}^{n_1} \frac{h_1\left(F_{X_0}^{-1}\left(\widehat{F}_{X_1}\left(X_{1i}\right)\right), \beta_o\right)}{f_{X_0}\left(F_{X_0}^{-1}\left(\widehat{F}_{X_1}\left(X_{1i}\right)\right)\right)} \left(\widehat{F}_{X_0}\left(F_{X_0}^{-1}\left(F_{X_1}\left(X_{1i}\right)\right)\right) - F_{X_1}\left(X_{1i}\right)\right) I\left\{X_{1i} \in \mathcal{S}\right\} \right. \tag{A.24}$$

$$\left. - \frac{1}{n_1} \sum_{i=1}^{n_1} \frac{h_1\left(F_{X_0}^{-1}\left(F_{X_1}\left(X_{1i}\right)\right), \beta_o\right)}{f_{X_0}\left(F_{X_0}^{-1}\left(F_{X_1}\left(X_{1i}\right)\right)\right)} \left(\widehat{F}_{X_0}\left(F_{X_0}^{-1}\left(F_{X_1}\left(X_{1i}\right)\right)\right) - F_{X_1}\left(X_{1i}\right)\right) I\left\{X_{1i} \in \mathcal{S}\right\} \right|,$$

and (A.23) is bounded by

$$\begin{aligned}
&\left| \frac{1}{n_1} \sum_{i=1}^{n_1} \frac{h_1\left(F_{X_0}^{-1}\left(\widehat{F}_{X_1}\left(X_{1i}\right)\right), \beta_o\right)}{f_{X_0}\left(F_{X_0}^{-1}\left(\widehat{F}_{X_1}\left(X_{1i}\right)\right)\right)} \times \left[\left(\widehat{F}_{X_0}\left(F_{X_0}^{-1}\left(\widehat{F}_{X_1}\left(X_{1i}\right)\right)\right) - \widehat{F}_{X_1}\left(X_{1i}\right)\right) \right.\right. \\
&\quad \left.\left. - \left(\widehat{F}_{X_0}\left(F_{X_0}^{-1}\left(F_{X_1}\left(X_{1i}\right)\right)\right) - F_{X_1}\left(X_{1i}\right)\right)\right] I\left\{X_{1i} \in \mathcal{S}\right\} \right| \\
&\leq \sup_q \left| \frac{h_1\left(F_{X_0}^{-1}\left(q\right), \beta_{o\mathcal{S}}\right)}{f_{X_0}\left(F_{X_0}^{-1}\left(q\right)\right)} \right| \times \sup_x \left| \left(\widehat{F}_{X_0}\left(F_{X_0}^{-1}\left(\widehat{F}_{X_1}\left(x\right)\right)\right) - \widehat{F}_{X_0}\left(F_{X_0}^{-1}\left(F_{X_1}\left(x\right)\right)\right)\right) - \left(\widehat{F}_{X_1}\left(x\right) - F_{X_1}\left(x\right)\right) \right| \\
&= \sup_q \left| \frac{h_1\left(F_{X_0}^{-1}\left(q\right), \beta_o\right)}{f_{X_0}\left(F_{X_0}^{-1}\left(q\right)\right)} \right| \times \sup_x \left| \widehat{F}_{X_0}\left(F_{X_0}^{-1}\left(F_{X_1}\left(x\right)\right) + F_{X_0}^{-1}\left(\widehat{F}_{X_1}\left(x\right)\right) - F_{X_0}^{-1}\left(F_{X_1}\left(x\right)\right)\right) \right. \\
&\quad \left. - \widehat{F}_{X_0}\left(F_{X_0}^{-1}\left(F_{X_1}\left(x\right)\right)\right) - \left(F_{X_0}\left(F_{X_0}^{-1}\left(F_{X_1}\left(x\right)\right) + F_{X_0}^{-1}\left(\widehat{F}_{X_1}\left(x\right)\right) - F_{X_0}^{-1}\left(F_{X_1}\left(x\right)\right)\right) - F_{X_1}\left(x\right)\right) \right| \\
&= o_p\left(n^{-1/2}\right).
\end{aligned}$$

The last equality holds because we can apply Lemma A.5 in Athey and Imbens (2006) and take $\delta = 1/3$ and $\eta = 1/2$. To see if the condition in Lemma A.5 in Athey and Imbens (2006) is satisfied, we have $\sup_x \left| F_{X_0}^{-1}\left(\widehat{F}_{X_1}\left(x\right)\right) - F_{X_0}^{-1}\left(F_{X_1}\left(x\right)\right) \right| \leq \sup_x \left| \left(\widehat{F}_{X_1}(x) - F_{X_1}(x)\right) / \underline{f}_{X_0} \right| = o_p\left(n^{-1/2}\right)$ by mean value theorem and Lemma A.2 in Athey and Imbens (2006), thus the

condition is satisfied and we can apply Lemma A.5 in Athey and Imbens (2006). Together with $F_{X_0}^{-1}(q) \in \mathcal{X}_0$, $h_1(\cdot, \beta_o)$ bounded from infinity and f_{X_0} is bounded from zero, we have (A.23) is $o_p\left(n^{-1/2}\right)$. Equation (A.24) is bounded by

$$\begin{aligned}
&\left| \frac{1}{n_1} \sum_{i=1}^{n_1} \left(\frac{h_1\left(F_{X_0}^{-1}\left(\widehat{F}_{X_1}(X_{1i})\right), \beta_o\right)}{f_{X_0}\left(F_{X_0}^{-1}\left(\widehat{F}_{X_1}(X_{1i})\right)\right)} - \frac{h_1\left(F_{X_0}^{-1}(F_{X_1}(X_{1i})), \beta_{o\mathcal{S}}\right)}{f_0\left(F_{X_0}^{-1}(F_{X_1}(X_{1i}))\right)} \right) \right. \\
&\qquad \left. \times \left(\widehat{F}_{X_0}\left(F_{X_0}^{-1}(F_{X_1}(X_{1i}))\right) - F_{X_1}(X_{1i})\right) \times I\{X_{1i} \in \mathcal{S}\} \right| \\
&\leq \sup_x \left| \frac{h_1\left(F_{X_0}^{-1}\left(\widehat{F}_{X_1}(x)\right), \beta_{o\mathcal{S}}\right)}{f_0\left(F_{X_0}^{-1}\left(\widehat{F}_{X_1}(x)\right)\right)} - \frac{h_1\left(F_{X_0}^{-1}(F_{X_1}(x)), \beta_{o\mathcal{S}}\right)}{f_0\left(F_{X_0}^{-1}(F_{X_1}(x))\right)} \right| \times \sup_x \left| \widehat{F}_{X_0}\left(F_{X_0}^{-1}(F_{X_1}(x))\right) - F_{X_1}(x) \right| \\
&= \sup_x \left| \left(\frac{h_{11}\left(\widetilde{X}_0, \beta_o\right)}{f_{X_0}^2(\widetilde{X}_0)} - \frac{h_1\left(\widetilde{X}_0, \beta_{o\mathcal{S}}\right) f'_{X_0}\left(\widetilde{X}_0\right)}{f_{X_0}^3(\widetilde{X}_0)} \right) \left(\widehat{F}_{X_1}(x) - F_{X_1}(x)\right) \right| \times \sup_x \left| \widehat{F}_{X_0}\left(F_{X_0}^{-1}(F_{X_1}(x))\right) - F_{X_1}(\right.
\end{aligned}$$

for $\widetilde{X}_0 \in \left(F_{X_0}^{-1}(F_{X_1}(x)), F_{X_0}^{-1}\left(\widehat{F}_{X_1}(x)\right)\right)$. Given that $f'_{X_0}\left(\widetilde{X}_0\right) < \infty$ (f_{X_0} is continuously differentiable) and by Lemma A.2 in Athey and Imbens (2006), we have equation (A.24) is $o_p\left(n^{-1/2}\right)$. Hence, (A.17) is $\frac{1}{n_0}\sum_{j=1}^{n_0} q(X_{0j}, \beta_{o\mathcal{S}}) + o_p\left(n^{-1/2}\right)$.

Second, consider (A.18), we have

$$\begin{aligned}
&\frac{1}{n_1}\sum_{i=1}^{n_1} h\left(F_{X_0}^{-1}\left(\widehat{F}_{X_1}(X_{1i})\right), \beta_{o\mathcal{S}}\right) I\{X_{1i} \in \mathcal{S}\} - \frac{1}{n_1}\sum_{i=1}^{n_1} h\left(F_{X_0}^{-1}(F_{X_1}(X_{1i})), \beta_{o\mathcal{S}}\right) I\{X_{1i} \in \mathcal{S}\} \\
&= \frac{1}{n_1}\sum_{i=1}^{n_1} h\left(F_{X_0}^{-1}\left(\widehat{F}_{X_1}(X_{1i})\right), \beta_{o\mathcal{S}}\right) I\{X_{1i} \in \mathcal{S}\} - \frac{1}{n_1}\sum_{i=1}^{n_1} h\left(F_{X_0}^{-1}(F_{X_1}(X_{1i})), \beta_{o\mathcal{S}}\right) I\{X_{1i} \in \mathcal{S}\} \qquad \text{(A.25)} \\
&\quad + \frac{1}{n_1}\frac{1}{n_1}\sum_{j=1}^{n_1}\sum_{i=1}^{n_1} Q\left(T_o(X_{1j}), X_{1i}, \beta_{o\mathcal{S}}\right) \\
&\quad - \frac{1}{n_1}\sum_{i=1}^{n_1} q\left(T_o(X_{1i}), \beta_{o\mathcal{S}}\right) + o_p\left(n^{-1/2}\right). \qquad \text{(A.26)}
\end{aligned}$$

Equation (A.25) can be bounded by

$$
\begin{aligned}
&\left|\frac{1}{n_1}\sum_{i=1}^{n_1}\left(h\left(F_{X_0}^{-1}\left(\widehat{F}_{X_1}(X_{1i})\right),\beta_{o\mathcal{S}}\right)-h\left(F_{X_0}^{-1}\left(F_{X_1}(X_{1i})\right),\beta_o\right)\right.\right.\\
&\qquad\left.\left.-\frac{h_1\left(F_{X_0}^{-1}\left(F_{X_1}(X_1)\right),\beta_{o\mathcal{S}}\right)}{f_{X_0}\left(F_{X_0}^{-1}\left(F_{X_1}(X_{1i})\right)\right)}\times\frac{1}{n_1}\sum_{j=1}^{n_1}\left[I\left\{X_{1j}\le X_{1i}\right\}-F_{X_1}(X_{1i})\right]\right)I\left\{X_{1i}\in\mathcal{S}\right\}\right|\\
&\le\sup_{x\in\mathcal{X}_1}\left|h\left(F_{X_0}^{-1}\left(\widehat{F}_{X_1}(x)\right),\beta_{o\mathcal{S}}\right)-h\left(F_{X_0}^{-1}\left(F_{X_1}(x)\right),\beta_{o\mathcal{S}}\right)\right.\\
&\qquad\left.-\frac{h_1\left(F_{X_0}^{-1}\left(F_{X_1}(x)\right),\beta_{o\mathcal{S}}\right)}{f_{X_0}\left(F_{X_0}^{-1}\left(F_{X_1}(x)\right)\right)}\times\frac{1}{n_1}\sum_{j=1}^{n_1}\left[I\left\{X_{pj}\le x\right\}-F_{X_1}(x)\right]\right|\\
&=\sup_x\left|h\left(F_{X_0}^{-1}\left(\widehat{F}_{X_1}(x)\right),\beta_{o\mathcal{S}}\right)-h\left(F_{X_0}^{-1}\left(F_{X_1}(x)\right),\beta_{o\mathcal{S}}\right)-\frac{h_1\left(F_{X_0}^{-1}\left(F_{X_1}(x)\right),\beta_{o\mathcal{S}}\right)}{f_{X_0}\left(F_{X_0}^{-1}\left(F_{X_1}(x)\right)\right)}\left[\widehat{F}_{X_1}(x)-F_{X_1}(x)\right]\right|\\
&=\sup_x\left|\left[\frac{h_{11}\left(F_{X_0}^{-1}\left(F_{X_1}(x)\right),\beta_{o\mathcal{S}}\right)}{f_{X_0}^2\left(F_{X_0}^{-1}\left(F_{X_1}(x)\right)\right)}-\frac{h_1\left(F_{X_0}^{-1}\left(F_{X_1}(x)\right),\beta_{o\mathcal{S}}\right)\cdot f_{X_0}'\left(F_{X_0}^{-1}\left(F_{X_1}(x)\right)\right)}{f_{X_0}^3\left(F_{X_0}^{-1}\left(F_{X_1}(x)\right)\right)}\right]\left[\widehat{F}_{X_1}(x)-F_{X_1}(x)\right]^2\right.\\
&\qquad\left.+o\left(\left[\widehat{F}_{X_1}(x)-F_{X_1}(x)\right]^2\right)\right|\\
&\le\sup_x\left|\frac{h_{11}\left(F_{X_0}^{-1}\left(F_{X_1}(x)\right),\beta_{o\mathcal{S}}\right)}{f_{X_0}^2\left(F_{X_0}^{-1}\left(F_{X_1}(x)\right)\right)}-\frac{h_1\left(F_{X_0}^{-1}\left(F_{X_1}(x)\right),\beta_{o\mathcal{S}}\right)\cdot f_{X_0}'\left(F_{X_0}^{-1}\left(F_{X_1}(x)\right)\right)}{f_{X_0}^3\left(F_{X_0}^{-1}\left(\widetilde{F}_{X_1}(x)\right)\right)}+o(1)\right|\\
&\qquad\times\sup_x\left|\left[\widehat{F}_{X_1}(x)-F_{X_1}(x)\right]^2\right|,
\end{aligned}
$$

The last step is $o_p\left(n^{-1/2}\right)$ by Lemma A.2 and Lemma A.7 in Athey and Imbens (2006). Hence, equation (A.18) is $\frac{1}{n_1}\sum_{i=1}^{n_1}p\left(X_{1i},\beta_{o\mathcal{S}}\right)+o_p\left(n^{-1/2}\right)$. Note that (A.19) is $o_p\left(n^{-1/2}\right)$. This overall completes the proof. □

A.3 Proofs in Section 1.5.2

First we establish asymptotic linear representation of $\widehat{A}-A$ by the following lemma. Denote the eigendecomposition $\Sigma_1\Sigma_0=UDU^{-1}$, λ_d as the d-th eigenvalue of diagonal matrix D, $[L]_{d,s}=\frac{1}{\sqrt{\lambda_d}+\sqrt{\lambda_s}}$ and $[K]_{d,s}=\frac{\sqrt{\lambda_d\lambda_s}}{\sqrt{\lambda_d}+\sqrt{\lambda_s}}$ as the d-th and s-th element of matrix L and K. Denote $A\circ B$ the Hadamard product between matrices.

Lemma 6. Under Condition 6, the linear representation of $\widehat{A} - A$ is shown as below

$$\begin{aligned}\widehat{A} - A = &-\frac{1}{n_1}\sum_{i=1}^{n_1}\Sigma_0 U\left[K \circ \left(U^{-1}\Sigma_0^{-1}\Sigma_1^{-1}\left(\Sigma_{1i} - \Sigma_1\right)\Sigma_1^{-1}U\right)\right]U^{-1} \\ &+ \frac{1}{n_0}\sum_{j=1}^{n_0}\Sigma_1^{-1}U\left[L \circ \left(U^{-1}\Sigma_1\left(\Sigma_{0j} - \Sigma_0\right)U\right)\right]U^{-1} + o_p\left(n_1^{-1/2}\right),\end{aligned}$$

where $\Sigma_{1i} = \left(X_{1i} - \mu_1\right)\left(X_{1i} - \mu_1\right)'$, $\Sigma_{0j} = \left(X_{0j} - \mu_0\right)\left(X_{0j} - \mu_0\right)'$.

Proof of Lemma 6.

Noting that $\widehat{A}$ is the geometric mean of $\widehat{\Sigma}_1^{-1}$ and $\widehat{\Sigma}_0$ and A is the geometric mean of Σ_1^{-1} and Σ_0, it has the following alternative expressions:

$$A = \Sigma_1^{-1/2}(\Sigma_1^{1/2}\Sigma_0\Sigma_1^{1/2})^{1/2}\Sigma_1^{-1/2} = \Sigma_1^{-1}\left(\Sigma_1\Sigma_0\right)^{1/2} = \Sigma_0\left(\Sigma_0^{-1}\Sigma_1^{-1}\right)^{1/2}.$$

Using the Frechet derivative of the matrix square root function, we obtain

$$\widehat{A} = A + \left(\nabla_{\Sigma_1}A\right)\left(\widehat{\Sigma}_1 - \Sigma_1\right) + \left(\nabla_{\Sigma_0}A\right)\left(\widehat{\Sigma}_0 - \Sigma_0\right) + o\left(\left\|\left(\widehat{\Sigma}_1 - \Sigma_1, \widehat{\Sigma}_0 - \Sigma_0\right)\right\|\right),$$

where $\nabla_{\Sigma_1}A$ and $\nabla_{\Sigma_0}A$ are non-singular matrices which perturb Σ_1 and Σ_0 along an arbitrary direction of $\widehat{\Sigma}_1 - \Sigma_1$ and $\widehat{\Sigma}_0 - \Sigma_0$. Here $\|\cdot\|$ can be any unitary invariant matrix norm, i.e., frobenius norm. Using Chain rule and theorem A.1 in Yang and Tabak (2020), we have

$$\begin{aligned}\left(\nabla_{\Sigma_1}A\right)\left(\widehat{\Sigma}_1 - \Sigma_1\right) &= -\Sigma_0\left(\nabla\left(\Sigma_0^{-1}\Sigma_1^{-1}\right)^{1/2}\right)\left(\Sigma_0^{-1}\Sigma_1^{-1}\left(\widehat{\Sigma}_1 - \Sigma_1\right)\Sigma_1^{-1}\right) \\ &= -\Sigma_0\int_0^\infty e^{-\left(\Sigma_0^{-1}\Sigma_1^{-1}\right)^{1/2}t}\left(\Sigma_0^{-1}\Sigma_1^{-1}\left(\widehat{\Sigma}_1 - \Sigma_1\right)\Sigma_1^{-1}\right)e^{-\left(\Sigma_0^{-1}\Sigma_1^{-1}\right)^{1/2}t}dt.\end{aligned}$$

Using the eigendecomposition $\Sigma_1\Sigma_0 = UDU^{-1}$, $\Sigma_0^{-1}\Sigma_1^{-1} = UD^{-1}U^{-1}$. Denote $[L]_{i,j} = \frac{1}{\sqrt{\lambda_i}+\sqrt{\lambda_j}}$ and $[K]_{i,j} = \frac{\sqrt{\lambda_i\lambda_j}}{\sqrt{\lambda_i}+\sqrt{\lambda_j}}$ as the ith jth element of L and K, where λ_i is the ith eigenvalue of diagonal matrix D. We have

$$\left(\nabla_{\Sigma_1}A\right)\left(\widehat{\Sigma}_1 - \Sigma_1\right) = -\Sigma_0 U\left[K \circ U^{-1}\left(\Sigma_0^{-1}\Sigma_1^{-1}\left(\widehat{\Sigma}_1 - \Sigma_1\right)\Sigma_1^{-1}\right)U\right]U^{-1},$$

where $\circ$ is Hadamard product. The last equation is because for the eigendecomposition $\Sigma = UDU^{-1}$, we have

$$\int_0^\infty e^{-\left(UDU^{-1}\right)^{\frac{1}{2}}t} S e^{-\left(UDU^{-1}\right)^{\frac{1}{2}}t} dt = U\left(\int_0^\infty e^{-D^{\frac{1}{2}}t} U^{-1} S U e^{-D^{\frac{1}{2}}t} dt\right) U^{-1} = U\left(L \circ U^{-1} S U\right) U^{-1}.$$

Similarly, we obtain

$$\begin{aligned}
\left(\nabla_{\Sigma_0} A\right)\left(\widehat{\Sigma}_0 - \Sigma_0\right) &= \Sigma_1^{-1}\left(\nabla\left(\Sigma_1\Sigma_0\right)^{1/2}\right)\left(\Sigma_1\left(\widehat{\Sigma}_0 - \Sigma_0\right)\right) \\
&= \Sigma_1^{-1}\int_0^\infty e^{-\left(\Sigma_1\Sigma_0\right)^{1/2}t}\left(\Sigma_1\left(\widehat{\Sigma}_0 - \Sigma_0\right)\right) e^{-\left(\Sigma_1\Sigma_0\right)^{1/2}t} dt \\
&= \Sigma_1^{-1} U\left[L \circ U^{-1}\left(\Sigma_1\left(\widehat{\Sigma}_0 - \Sigma_0\right)\right) U\right] U^{-1}.
\end{aligned}$$

Because for $j = 0, 1$,

$$\begin{aligned}
\widehat{\Sigma}_j - \Sigma_j &= \frac{1}{n_j}\sum_{i=1}^{n_j}\left\{\left(X_{ji} - \widehat{\mu}_j\right)\left(X_{ji} - \widehat{\mu}_j\right)' - E\left[\left(X_{ji} - \mu_j\right)\left(X_{ji} - \mu_j\right)'\right]\right\} \\
&= \frac{1}{n_j}\sum_{i=1}^{n_j}\left\{\left(X_{ji} - \mu_j\right)\left(X_{ji} - \mu_j\right)' - E\left[\left(X_{ji} - \mu_j\right)\left(X_{ji} - \mu_j\right)'\right]\right\} \\
&\quad - \left(\widehat{\mu}_j - \mu_j\right)\left(\widehat{\mu}_j - \mu_j\right)' \\
&= \frac{1}{n_j}\sum_{i=1}^{n_j}\left\{\left(X_{ji} - \mu_j\right)\left(X_{ji} - \mu_j\right)' - E\left[\left(X_{ji} - \mu_j\right)\left(X_{ji} - \mu_j\right)'\right]\right\} - o_p\left(n_j^{-1/2}\right) \\
&\coloneqq \frac{1}{n_j}\sum_{i=1}^{n_j}\left\{\Sigma_{ji} - \Sigma_j\right\} - o_p\left(n_j^{-1/2}\right),
\end{aligned}$$

we can write

$$\begin{aligned}
\left(\nabla_{\Sigma_1} A\right)\left(\widehat{\Sigma}_1 - \Sigma_1\right) &= -\Sigma_0 U\left[K \circ U^{-1}\left(\Sigma_0^{-1}\Sigma_1^{-1}\left(\frac{1}{n_1}\sum_{i=1}^{n_1}\left\{\Sigma_{1i} - \Sigma_1\right\} - o_p\left(n_1^{-1/2}\right)\right)\Sigma_1^{-1}\right) U\right] U^{-1} \\
&= -\frac{1}{n_1}\sum_{i=1}^{n_1}\Sigma_0 U\left[K \circ \left(U^{-1}\Sigma_0^{-1}\Sigma_1^{-1}\left(\Sigma_{1i} - \Sigma_1\right)\Sigma_1^{-1} U\right)\right] U^{-1} + o_p\left(n_1^{-1/2}\right),
\end{aligned}$$

$$(\nabla_{\Sigma_0} A)\left(\widehat{\Sigma}_0 - \Sigma_0\right) = \frac{1}{n_0}\sum_{j=1}^{n_0} \Sigma_1^{-1} U\left[L \circ \left(U^{-1}\Sigma_1\left(\Sigma_{0j} - \Sigma_0\right)U\right)\right] U^{-1} - o_p\left(n_0^{-1/2}\right),$$

and

$$\begin{aligned}
\widehat{A} - A = & -\frac{1}{n_1}\sum_{i=1}^{n_1} \Sigma_0 U\left[K \circ \left(U^{-1}\Sigma_0^{-1}\Sigma_1^{-1}\left(\Sigma_{1i} - \Sigma_1\right)\Sigma_1^{-1}U\right)\right] U^{-1} \\
& + \frac{1}{n_0}\sum_{j=1}^{n_0} \Sigma_1^{-1} U\left[L \circ \left(U^{-1}\Sigma_1\left(\Sigma_{0j} - \Sigma_0\right)U\right)\right] U^{-1} \\
& + o_p\left(n_1^{-1/2}\right) + o\left(\left\|\left(\frac{1}{n_1}\sum_{i=1}^{n_1}\Sigma_{1i} - \Sigma_1, \frac{1}{n_0}\sum_{j=1}^{n_0}\Sigma_{0j} - \Sigma_0\right)\right\|\right) \\
= & -\frac{1}{n_1}\sum_{i=1}^{n_1} \Sigma_0 U\left[K \circ \left(U^{-1}\Sigma_0^{-1}\Sigma_1^{-1}\left(\Sigma_{1i} - \Sigma_1\right)\Sigma_1^{-1}U\right)\right] U^{-1} \\
& + \frac{1}{n_0}\sum_{j=1}^{n_0} \Sigma_1^{-1} U\left[L \circ \left(U^{-1}\Sigma_1\left(\Sigma_{0j} - \Sigma_0\right)U\right)\right] U^{-1} + o_p\left(n_1^{-1/2}\right).
\end{aligned}$$

□

Proof of Proposition 3.

By Assumption 11,

$$\begin{aligned}
& \left\|h\left(\widehat{T}\left(x\right), \beta\right) - h\left(T_o\left(x\right), \beta\right)\right\|_{\infty,\omega} \\
& = \left\|h_1\left(T_o\left(x\right), \beta\right)'\left(\widehat{T}\left(x\right) - T_o\left(x\right)\right) + o\left(h_1\left(T_o\left(x\right), \beta\right)'\left(\widehat{T}\left(x\right) - T_o\left(x\right)\right)\right)\right\|_{\infty,\omega} \\
& \leq \left\|h_1\left(T_o\left(x\right), \beta\right)'\left(\left(\widehat{A} - A\right)\left(\mu_1 - \widehat{\mu}_1\right) + A\left(\mu_1 - \widehat{\mu}_1\right) + \left(\widehat{A} - A\right)\left(x - \mu_1\right) + \left(\widehat{\mu}_0 - \mu_0\right)\right)\right\|_{\infty,\omega} \\
& \quad + \left\|o\left(h_1\left(T_o\left(x\right), \beta\right)'\left(\widehat{T}\left(x\right) - T_o\left(x\right)\right)\right)\right\|_{\infty,\omega} \\
& \leq \left\|h_1\left(T_o\left(x\right), \beta\right)'\left(\widehat{A} - A\right)\left(\mu_1 - \widehat{\mu}_1\right)\right\|_{\infty,\omega} + \left\|h_1\left(T_o\left(x\right), \beta\right)^T A\left(\mu_1 - \widehat{\mu}_1\right)\right\|_{\infty,\omega} \qquad \text{(A.27)} \\
& \quad + \left\|h_1\left(T_o\left(x\right), \beta\right)'\left(\widehat{A} - A\right)\left(x - \mu_1\right)\right\|_{\infty,\omega} \qquad \text{(A.28)} \\
& \quad + \left\|h_1\left(T_o\left(x\right), \beta\right)'\left(\widehat{\mu}_0 - \mu_0\right)\right\|_{\infty,\omega} \qquad \text{(A.29)} \\
& \quad + \left\|o\left(h_1\left(T_o\left(x\right), \beta\right)'\left(\widehat{T}\left(x\right) - T_o\left(x\right)\right)\right)\right\|_{\infty,\omega}.
\end{aligned}$$

By Lemma 6, and the fact that $\frac{1}{n_j}\sum_{i=1}^{n_j}\Sigma_{ji}-\Sigma_j=o_p(1)$ for $j=0,1$, we have $\widehat{A}-A=o_p(1)$. Together with the fact that $\mu_j-\widehat{\mu}_j=o_p(1)$ for $j=0,1$, and Assumption 4.1, we have (A.27) and (A.29) is $o_p(1)$. Notice that for any $\omega>\omega_1$, denote ith element of $T_o(x)$, we have

$$\sup_x\left|T_o(x)_i\left(1+T_o(x)'T_o(x)\right)^{-(\omega-\omega_1)/2}\right|<\infty.$$

Write it in a matrix form, we have

$$\sup_x\left|\left(A(x-\mu_1)+\mu_0\right)\left(1+T_o(x)'T_o(x)\right)^{-(\omega-\omega_1)/2}\right|=\sup_x\left|T_o(x)\left(1+T_o(x)'T_o(x)\right)^{-(\omega-\omega_1)/2}\right|<\infty.$$

Thus,

$$\sup_x\left|x\left(1+T_o(x)'T_o(x)\right)^{-(\omega-\omega_1)/2}\right|<\infty,$$

and hence (A.28) is $o_p(1)$ and $\left\|h\left(\widehat{T}(x),\beta\right)-h\left(T_o(x),\beta\right)\right\|_{\infty,\omega}=o_p(1)$.

Now let's prove the second part of the proposition. Let $T(x)=(T_1(x),\ldots,T_d(x)^{d_x})'$ for $d=1,\ldots,d_x$,

$$\begin{aligned}
&\left\|\widehat{T}(x)-T_o(x)\right\|_{2,1}^2=\sum_{d=1}^{d_X}\int\left|\widehat{T}_d(x)-T_{o,d}(x)\right|^2dF_{X_1}=E_{X_1}\left[\left\|\widehat{T}(x)-T_o(x)\right\|_2^2\right]\\
&=E_{X_1}\left[\left\|\left(\widehat{A}-A\right)(\mu_1-\widehat{\mu}_1)+A(\mu_1-\widehat{\mu}_1)+\left(\widehat{A}-A\right)(x-\mu_1)+(\widehat{\mu}_0-\mu_0)\right\|_{L2}^2\right]\\
&\leq\left\|\left(\widehat{A}-A\right)(\mu_1-\widehat{\mu}_1)\right\|_2^2+\|A(\mu_1-\widehat{\mu}_1)\|_2^2+E\left[\left\|\left(\widehat{A}-A\right)(x-\mu_1)\right\|_2^2\right]+\|\widehat{\mu}_0-\mu_0\|_2^2\\
&=\left\|\left(\widehat{A}-A\right)(\mu_1-\widehat{\mu}_1)\right\|_2^2+\|A(\mu_1-\widehat{\mu}_1)\|_2^2+\|\widehat{\mu}_0-\mu_0\|_2^2 \qquad \text{(A.30)}\\
&\quad+E\left[(x-\mu_1)'\left(\widehat{A}-A\right)'\left(\widehat{A}-A\right)(x-\mu_1)\right] \qquad \text{(A.31)}\\
&=o_p\left(n^{-1/2}\right).
\end{aligned}$$

The last equality holds because by Lemma 6, $\widehat{A}-A$ is $O_p\left(n^{-1/2}\right)$ together with the fact that $\widehat{\mu}_j-\mu_j$ is $O_p\left(n^{-1/2}\right)$ for $j=0,1$ give us (A.30) is $o_p\left(n^{-1/2}\right)$. Because x has finite second moment, (A.31) is $o_p\left(n^{-1/2}\right)$.

To get the convergence rate for $\left\|\widehat{h}_1\left(T_o(\cdot),\beta_{o\mathcal{S}}\right)-h_1\left(T_o(\cdot),\beta_{o\mathcal{S}}\right)\right\|_{2,0}^2$ let $h_1(x,\beta)=\nabla h(x,\beta)=\begin{bmatrix}\frac{\partial}{\partial x_1}h(x,\beta)\\ \vdots \\ \frac{\partial}{\partial x_d}h(x,\beta)\end{bmatrix}:=\begin{bmatrix}h_1^1(x,\beta)\\ \vdots \\ h_1^{d_X}(x,\beta)\end{bmatrix}$ is the gradient of $h(x,\beta)$ with respect to $x\in\mathbb{R}^d$. By Assumption 4.1, we have $h_1^d(\cdot,\beta)\in\Lambda_c^{\gamma-1}(\mathcal{X}_0,\omega_1)$ for $d=1,\ldots,d_x$. We then can obtain the convergence rate of $\left\|\widehat{h}_1^d(\cdot,\beta_{o\mathcal{S}})-h_1^d(\cdot,\beta_{o\mathcal{S}})\right\|_{2,0}$ by applying Theorem 1 in Chen and Shen (1998), we have

$$\left\|\widehat{h}_1^d(\cdot,\beta_{o\mathcal{S}})-h_1^d(\cdot,\beta_{o\mathcal{S}})\right\|_{2,0}=O_p\left(\sqrt{k_{n_0}/n_0}+(k_{n_0})^{-(\gamma-1)/d_X}\right).$$

By choosing $k_{n_0}=(n_0)^{\frac{d_X}{2\gamma+d_X}}$, one have

$$\left\|\widehat{h}_1^d(\cdot,\beta_{o\mathcal{S}})-h_1^d(\cdot,\beta_{o\mathcal{S}})\right\|_{2,0}=O_p\left(n_0^{-\frac{\gamma-1}{2\gamma+d_X}}\right),$$

and thus

$$\left\|\widehat{h}_1\left(T_o(\cdot),\beta_{o\mathcal{S}}\right)-h_1\left(T_o(\cdot),\beta_{o\mathcal{S}}\right)\right\|_{2,0}^2=O_p\left(n_0^{-\frac{2(\gamma-1)}{2\gamma+d_X}}\right). \tag{A.32}$$

Proposition 3 holds for any $\gamma>0$. □

Proof of Proposition 4.

We have

$$\begin{aligned}
&\frac{1}{n_1}\sum_{i=1}^{n_1}\left(h\left(\widehat{T}\left(X_{1i}\right),\beta_{o\mathcal{S}}\right)I\left\{X_{1i}\in\mathcal{S}\right\}-E\left[h\left(T_o\left(X_1\right),\beta_{o\mathcal{S}}\right)I\left\{X_1\in\mathcal{S}\right\}\right]\right)\\
&=E\left[h\left(\widehat{T}\left(X_1\right),\beta_o\right)I\left\{X_1\in\mathcal{S}\right\}-h\left(T_o\left(X_1\right),\beta_{o\mathcal{S}}\right)I\left\{X_1\in\mathcal{S}\right\}\right] \qquad \text{(A.33)}\\
&\quad+\frac{1}{n_1}\sum_{i=1}^{n_1}\left(h\left(T_o\left(X_{1i}\right),\beta_{o\mathcal{S}}\right)I\left\{X_{1i}\in\mathcal{S}\right\}-E\left[h\left(T_o\left(X_1\right),\beta_{o\mathcal{S}}\right)I\left\{X_1\in\mathcal{S}\right\}\right]\right)\\
&\quad+\frac{1}{n_1}\sum_{i=1}^{n_1}\left(\left[h\left(\widehat{T}\left(X_{1i}\right),\beta_{o\mathcal{S}}\right)-h\left(T_o\left(X_{1i}\right),\beta_{o\mathcal{S}}\right)\right]I\left\{X_{1i}\in\mathcal{S}\right\}\right. \qquad \text{(A.34)}\\
&\quad\left.-E\left[\left(h\left(\widehat{T}\left(X_1\right),\beta_o\right)-h\left(T_o\left(X_1\right),\beta_{o\mathcal{S}}\right)\right)I\left\{X_1\in\mathcal{S}\right\}\right]\right)
\end{aligned}$$

$$= E\left[\left(h\left(\widehat{T}\left(X_1\right),\beta_o\right) - h\left(T_o\left(X_1\right),\beta_{o\mathcal{S}}\right)\right) I\left\{X_1 \in \mathcal{S}\right\}\right] \tag{A.35}$$
$$+ \frac{1}{n_1}\sum_{i=1}^{n_1}\left(h\left(T_o\left(X_{1i}\right),\beta_{o\mathcal{S}}\right) I\left\{X_{1i} \in \mathcal{S}\right\} - E\left[h\left(T_o\left(X_1\right),\beta_{o\mathcal{S}}\right) I\left\{X_1 \in \mathcal{S}\right\}\right]\right) + o_p\left(n_1^{-1/2}\right)$$
$$= E\left[h_1\left(T_o\left(X_1\right),\beta_{o\mathcal{S}}\right)^T\left\{\widehat{T}\left(X_1\right) - T_o\left(X_1\right)\right\} I\left\{X_1 \in \mathcal{S}\right\} + O\left(\left\|\widehat{T}\left(X_1\right) - T_o\left(X_1\right)\right\|_2^2\right) I\left\{X_1 \in \mathcal{S}\right\}\right]$$
$$+ \frac{1}{n_1}\sum_{i=1}^{n_1}\left(h\left(T_o\left(X_{1i}\right),\beta_{o\mathcal{S}}\right) I\left\{X_{1i} \in \mathcal{S}\right\} - E\left[h\left(T_o\left(X_1\right),\beta_{o\mathcal{S}}\right) I\left\{X_1 \in \mathcal{S}\right\}\right]\right) + o_p\left(n_1^{-1/2}\right)$$
$$= E\left[h_1\left(T_o\left(X_1\right),\beta_{o\mathcal{S}}\right)^T\left\{\widehat{\mu}_0 - \mu_0 + \left(\widehat{A} - A\right)\left(X_1 - \mu_1\right) - \left(\widehat{A} - A\right)\left(\widehat{\mu}_1 - \mu_1\right) - A\left(\widehat{\mu}_1 - \mu_1\right)\right\} I\left\{X_1 \in \mathcal{S}\right\}\right]$$
$$+ \frac{1}{n_1}\sum_{i=1}^{n_1}\left(h\left(T_o\left(X_{1i}\right),\beta_{o\mathcal{S}}\right) I\left\{X_{1i} \in \mathcal{S}\right\} - E\left[h\left(T_o\left(X_1\right),\beta_{o\mathcal{S}}\right) I\left\{X_1 \in \mathcal{S}\right\}\right]\right) + o_p\left(n_1^{-1/2}\right)$$
$$= -E\left[I\left\{X_1 \in \mathcal{S}\right\} h_1\left(T_o\left(X_1\right),\beta_{o\mathcal{S}}\right)^T\right] A\left(\widehat{m}_1 - m_1\right) + E\left[I\left\{X_1 \in \mathcal{S}\right\} h_1\left(T\left(X_1\right),\beta_{o\mathcal{S}}\right)^T\right]\left(\widehat{m}_0 - m_0\right) \tag{A.36}$$
$$+ E\left[h_1\left(T_o\left(X_1\right),\beta_{o\mathcal{S}}\right)^T\left(\widehat{A} - A\right)\left(X_1 - m_1\right) I\left\{X_1 \in \mathcal{S}\right\}\right] \tag{A.37}$$
$$- E\left[I\left\{X_1 \in \mathcal{S}\right\} h_1\left(T\left(X_p\right),\beta_{o\mathcal{S}}\right)^T\right]\left(\widehat{A} - A\right)\left(\widehat{m}_1 - m_1\right) \tag{A.38}$$
$$+ \frac{1}{n_1}\sum_{i=1}^{n_1}\left(h\left(T_o\left(X_{1i}\right),\beta_{o\mathcal{S}}\right) I\left\{X_{1i} \in \mathcal{S}\right\} - E\left[h\left(T_o\left(X_1\right),\beta_{o\mathcal{S}}\right) I\left\{X_1 \in \mathcal{S}\right\}\right]\right) + o_p\left(n_1^{-1/2}\right). \tag{A.39}$$

The last equality holds because $E_{X_1}\left\|\widehat{T}\left(X_1\right) - T_o\left(X_1\right)\right\|_2^2 = o_p\left(n_1^{-1/2}\right)$. By central limit theorem, we have $\frac{1}{n_j}\sum_{i=1}^{n_1}\Sigma_{ji} - \Sigma_j$ is $O_p\left(n_j^{-1/2}\right)$ for $j = 0, 1$. Thus, $\widehat{A} - A$ is $O_p\left(n_1^{-1/2}\right)$ by Lemma 6. Using the fact that $\widehat{\mu}_1 - \mu_1$ is $O_p\left(n_1^{-1/2}\right)$, we have equation (A.38) is $o_p\left(n_1^{-1/2}\right)$. Substitute $\widehat{A} - A$ in equation (A.37) we have

$$E\left[h_1\left(T_o\left(X_1\right),\beta_{o\mathcal{S}}\right)^T\left(\widehat{A} - A\right)\left(X_1 - \mu_1\right) I\left\{X_1 \in \mathcal{S}\right\}\right]$$
$$= E\left[h_1\left(T_o\left(X_1\right),\beta_{o\mathcal{S}}\right)^T\left(-\frac{1}{n_1}\sum_{i=1}^{n_1}\Sigma_0 U\left[K \circ \left(U^{-1}\Sigma_0^{-1}\Sigma_1^{-1}\left(\Sigma_{1i} - \Sigma_1\right)\Sigma_1^{-1}U\right)\right]U^{-1}\right)\left(X_1 - \mu_1\right) I\left\{X_1 \in \mathcal{S}\right\}\right]$$
$$+ E\left[h_1\left(T_o\left(X_1\right),\beta_{o\mathcal{S}}\right)^T\left(\frac{1}{n_0}\sum_{j=1}^{n_0}\Sigma_1^{-1}U\left[L \circ \left(U^{-1}\Sigma_1\left(\Sigma_{0j} - \Sigma_0\right)U\right)\right]U^{-1}\right)\left(X_1 - \mu_1\right) I\left\{X_1 \in \mathcal{S}\right\}\right] + o_p\left(n\right.$$

and

$$
\begin{aligned}
&\frac{1}{n_1}\sum_{i=1}^{n_1}\left(h\left(\widehat{T}\left(X_{1i}\right),\beta_{o\mathcal{S}}\right)I\left\{X_{1i}\in\mathcal{S}\right\}-E\left[h\left(T_o\left(X_1\right),\beta_{o\mathcal{S}}\right)I\left\{X_1\in\mathcal{S}\right\}\right]\right)\\
&=\frac{1}{n_1}\sum_{i=1}^{n_1}\left\{h\left(T_o\left(X_{1i}\right),\beta_{o\mathcal{S}}\right)I\left\{X_{1i}\in\mathcal{S}\right\}-E\left[h\left(T_o\left(X_1\right),\beta_{o\mathcal{S}}\right)I\left\{X_1\in\mathcal{S}\right\}\right]\right.\\
&\quad-E_{X_1}\left[I\left\{X_1\in\mathcal{S}\right\}h_1\left(T_o\left(X_1\right),\beta_{o\mathcal{S}}\right)^T A\left(X_{1i}-\mu_1\right)\right]\\
&\quad\left.-E_{X_1}\left[h_1\left(T_o\left(X_1\right),\beta_o\right)^T\left(\Sigma_0 U\left[K\circ\left(U^{-1}\Sigma_0^{-1}\Sigma_1^{-1}\left(\Sigma_{1i}-\Sigma_1\right)\Sigma_1^{-1}U\right)\right]U^{-1}\right)\left(X_1-\mu_1\right)I\left\{X_1\in\mathcal{S}\right\}\right]\right\}\\
&\quad+\frac{1}{n_0}\sum_{j=1}^{n_0}\left\{E_{X_1}\left[I\left\{X_1\in\mathcal{S}\right\}h_1\left(T_o\left(X_1\right),\beta_o\right)^T\left(X_{0j}-\mu_0\right)\right]\right.\\
&\quad\left.+E_{X_1}\left[h_1\left(T_o\left(X_1\right),\beta_o\right)^T\left(\Sigma_1^{-1}U\left[L\circ\left(U^{-1}\Sigma_1\left(\Sigma_{0j}-\Sigma_0\right)U\right)\right]U^{-1}\right)\right)\left(X_1-\mu_1\right)I\left\{X_1\in\mathcal{S}\right\}\right]\right\}\\
&\quad+o_p\left(n_1^{-1/2}\right).
\end{aligned}
$$

□

A.4 Proofs in Section 1.5.3

Proof of Proposition 5.

Step 1. We first proof the convergence rate of $\left\|\widehat{\psi}-\psi_o\right\|_{2,1}=O_p\left(n^{-(\alpha+1)/(2\alpha+2+d)}\right)$. by applying Theorem 3.2 in Chen (2007). We start from verifying Condition 3.7-3.8 in Chen (2007). For $z\in\mathcal{X}_0$, x^* and x_o^* solves for the convex conjugate ψ^* and ψ_o^*, we have

$$
-\left(\psi-\psi_o\right)\left(x_o^*\right)\leq\psi^*\left(z\right)-\psi_o^*\left(z\right)=\left\langle x^*,z\right\rangle-\left\langle x_o^*,z\right\rangle-\psi\left(x^*\right)+\psi_o\left(x_o^*\right)\leq-\left(\psi-\psi_o\right)\left(x^*\right),
$$

where $z=\nabla\psi\left(x^*\right)$ and $z=\nabla\psi_o\left(x_o^*\right)$. The above inequalities hold because by replacing x^* with x_o^* we can get the lower bound and by replacing x_o^* with x^* we can get the upper bound.

We can then bound

$$\begin{aligned}
\mathrm{Var}\left(\psi^*(X_0)-\psi_o^*(X_0)\right) &\leq E\left[\psi^*(X_0)-\psi_o^*(X_0)\right]^2 \\
&\leq E\left[\max\left\{\left|(\psi-\psi_o)\left(\nabla\psi_o^{-1}(X_0)\right)\right|,\left|(\psi-\psi_o)\left(\nabla\psi^{-1}(X_0)\right)\right|\right\}\right]^2 \\
&= \max\left\{E\left[(\psi-\psi_o)\left(\nabla\psi_o^{-1}(X_0)\right)\right]^2, E\left[(\psi-\psi_o)\left(\nabla\psi^{-1}(X_0)\right)\right]^2\right\} \\
&= \max\left\{\|\psi-\psi_o\|_{2,1}^2, E\left[(\psi-\psi_o)\left(\nabla\psi^{-1}\left(\nabla\psi_o(X_1)\right)\right)\right]^2\right\} \\
&\leq \max\left\{\|\psi-\psi_o\|_{2,1}^2, const.\cdot E\left[(\psi-\psi_o)(X_1)\right]^2\right\} \qquad \text{(A.40)} \\
&\leq const.\cdot\|\psi-\psi_o\|_{2,1}^2. \qquad \text{(A.41)}
\end{aligned}$$

where (A.40) holds by Corollary 2 in Stern (2010). Then we have

$$\begin{aligned}
&\mathrm{Var}\left(\psi(X_1)+\psi^*(X_0)-\psi_o(X_1)-\psi_o^*(X_0)\right) \\
&= \mathrm{Var}\left(\psi(X_1)-\psi_o(X_1)\right)+\mathrm{Var}\left(\psi^*(X_0)-\psi_o^*(X_0)\right) \\
&\leq const.\cdot\|\psi-\psi_o\|_{2,1}^2.
\end{aligned}$$

where the last inequality held by Corollary 2 in Stern (2010). Hence Condition 3.7 in Chen (2007) is satisfied for all $\varepsilon\leq 1$. On the other hand,

$$\begin{aligned}
&\left|\psi(X_1)+\psi^*(X_0)-\psi_o(X_1)-\psi_o^*(X_0)\right| \\
&\leq \left|\psi(X_1)-\psi_o(X_1)\right|+\left|\psi^*(X_0)-\psi_o^*(X_0)\right| \\
&\leq \left|\psi(X_1)-\psi_o(X_1)\right|+\max\left\{\left|(\psi-\psi_o)\left(\nabla\psi_o^{-1}(X_o)\right)\right|,\left|(\psi-\psi_o)\left(\nabla\psi^{-1}(X_o)\right)\right|\right\} \\
&\leq 2\|\psi-\psi_o\|_\infty.
\end{aligned}$$

By Theorem 1 of Gabushin (1967), $\|\psi-\psi_o\|_\infty\leq const.\|\psi-\psi_o\|_{2,1}^{2/3}$, we have

$$\sup_{\left\{\psi\in\Psi_n:\|\psi-\psi_o\|_{2,1}\leq\delta\right\}}\left|\psi(X_1)+\psi^*(X_0)-\psi_o(X_1)-\psi_o^*(X_0)\right|\leq const.\|\psi-\psi_o\|_{2,1}^{2/3}, \qquad \text{(A.42)}$$

and Condition 3.8 in Chen (2007) is satisfied. Denote $\pi_n\psi$ by a projection of ψ to Ψ_n. Denote

$$\mathcal{F}_n = \left\{\psi(X_1) + \psi^*(X_0) - \psi_o(X_1) - \psi_o^*(X_0) : \|\psi - \psi_o\|_{2,1} \leq \delta, \psi \in \Psi_n\right\},$$

and for some constant $b > 0$. Apply Theorem 3.2 in Chen (2007) we have $\|\psi_o - \pi_n\psi_o\|_{2,1} = o_p\left(k_n^{-(\alpha+1)/d}\right)$,

$$\begin{aligned}\frac{1}{\sqrt{n}\delta_n^2}\int_{b\delta_n^2}^{\delta_n}\sqrt{H_{[]}\left(w, \mathcal{F}_n, \|\cdot\|_{2,1}\right)}\mathrm{d}w &\leq \frac{1}{\sqrt{n}\delta_n^2}\int_{b\delta_n^2}^{\delta_n}\sqrt{\log N\left(w^{3/2}, \Psi_n, \|\cdot\|_{2,1}\right)}\mathrm{d}w \\ &\leq \frac{1}{\sqrt{n}\delta_n^2}\sqrt{k_n} \times \delta_n \leq const.\end{aligned}$$

and the solution is $\delta_n = \sqrt{k_n/n}$. Choose $k_n = O\left(n^{d/(2\alpha+2+d)}\right)$, we have $\left\|\widehat{\psi} - \psi_o\right\|_{2,1} = O_p\left(n^{-(\alpha+1)/(2\alpha+2+d)}\right)$. Moreover, we can get convergence rate of $\left\|\nabla\widehat{\psi} - \nabla\psi_o\right\|_{2,1}$ through similar steps. Denote $\nabla\psi\left(x\right) = \left[\frac{\partial}{\partial x_1}\psi(x), \ldots, \frac{\partial}{\partial x_{d_x}}\psi(x)\right]'$ and $\|\nabla\psi(x)\|_{2,1}^2 = \int\|\nabla\psi(x)\|_2^2\, dF_{X_1}(x)$.

We make use of the following inequality.

Lemma 7 (Poincare Wirtinger inequality)**.** Assume Condition 7 holds. Denote $\|v\|_{L2(\mathcal{X})} := \int_{\mathcal{X}}\|v(x)\|_2\, dx$. Then there exists a constant C, depending only on $\mathcal{X}_1$ and p, such that for every function $v = \psi - \psi_o$ with $\int_{\mathcal{X}_1} v(x)dx = 0$ we have

$$\|v\|_{L2(\mathcal{X}_1)} \leq C\|\nabla v\|_{L2(\mathcal{X}_1)}.$$

Using Lemma 7, and the fact that $c_f \leq f_{X_1} \leq C_f$, we have

$$\|\psi - \psi_o\|_{2,1}^2 = \int_{\mathcal{X}_1}\left(\left(\psi - \psi_o\right)(x)\right)^2 f_{X_1}(x)dx \leq C_f \cdot \|\psi - \psi_o\|_{L2(\mathcal{X}_1)}^2 \leq \frac{const.}{c_f}E\left(\|\nabla\left(\psi(x) - \psi_o(x)\right)\|_2^2\right).$$

Condition 3.7-3.8 can then be verified for $\|\nabla\psi - \nabla\psi_o\|_{2,1}$. Because $\frac{\partial}{\partial x_j}\psi(x) \in C^\alpha(\overline{\mathcal{X}_1})$ for

$j = 1, \ldots, d_x$, $\left\| \frac{\partial}{\partial x_j}\psi_o - \pi_n \frac{\partial}{\partial x_j}\psi_o \right\|_{2,1} = O\left((k_n)^{-\alpha/d_X} \right)$. Use $k_n = O\left(n^{d_X/(2\alpha+2+d_X)} \right)$ we have

$$\left\| \nabla\widehat{\psi} - \nabla\psi_o \right\|_{2,1}^2 = \sum_{j=1}^{d_X} \left\| \frac{\partial}{\partial x_j}\widehat{\psi} - \frac{\partial}{\partial x_j}\psi_o \right\|_{2,1}^2 = O_p\left(n^{-2\alpha/(2\alpha+2+d_X)} \right). \tag{A.43}$$

Note that

$$\|\psi - \psi_o\| := E\left[\{\nabla(\psi - \psi_o)(X_1)\}^T \left[\nabla^2\psi_o(X_1)\right]^{-1} \nabla(\psi - \psi_o)(X_1) \right] \asymp \left\| \nabla\widehat{\psi} - \nabla\psi_o \right\|_{2,1},$$

we have $\|\psi - \psi_o\| = O_p\left(n^{-\alpha/(2\alpha+2+d_X)} \right)$.

Step 2. In the next step, we verify conditions in Theorem 1 in Shen (1997). Denote $\pi_n\psi$ by a projection of ψ to $\Psi_n \equiv \left\{ \psi(\cdot) = B_{k_{n_1}}(\cdot)'\gamma \in \Psi : \gamma \in \mathbb{R}^{k_{n_1}} \right\}$, Ψ_n is a closed linear span of Ψ.

Suppose, for all $\psi \in \Psi$ and all $x = (x_1', x_0')'$, let $l(\psi, x) = \psi(x_1) + \psi^*(x_0)$, there exists $l'(\psi_o, x)[\psi - \psi_o]$ such that the remainder in the linear approximation can be written as

$$r[\psi - \psi_o, x] = l(\psi, x) - l(\psi_o, x) - l'(\psi_o, x)[\psi - \psi_o] \tag{A.44}$$

where $l'(\psi_o, x)[\psi - \psi_o]$ is the directional derivitive of l at ψ_o, which defined as

$$l'(\psi_o, x)[\psi - \psi_o] = \lim_{\tau \to 0} \frac{l(\psi_o + \tau(\psi - \psi_o), x) - l(\psi_o, x)}{\tau} = \psi(x_1) - \psi_o(x_1) - \psi\left(T_o^{-1}(x_0)\right) + \psi_o\left(T_o^{-1}(x_0)\right),$$

and we can simplify

$$r[\psi - \psi_o, x] = r[\psi - \psi_o, x_0] = \psi^*(x_0) - \psi_o^*(x_0) + \psi\left(T_o^{-1}(x_0)\right) - \psi_o\left(T_o^{-1}(x_0)\right).$$

Let $K(\psi_o, \psi) = E[\psi_o(X_1) + \psi_o^*(X_0)] - E[\psi(X_1) + \psi^*(X_0)]$ and let $\nu_n(\psi, X) = n^{-1/2}\sum_{i=1}^n (\psi(X_i) - E\psi(X_i))$ be the empirical process induced by ψ. Let the convergence rate of the sieve estimate under $\|\cdot\|$ be $o_p(\delta_n)$ and let $\varepsilon_n = o\left(n^{-1/2}\right)$.

For $\psi \in \{\psi \in \Psi_n : \|\psi - \psi_o\| \leq \delta_n\}$, consider a local alternative value $\psi^*(\psi, \varepsilon_n) =$

$(1-\varepsilon_n)\psi+\varepsilon_n(u^*+\psi_o)$, where $u^*=\pm v^*$.

We start with verifying the following condition

$$\sup_{\{\psi\in\Psi_n:\|\psi-\psi_o\|\le\delta_n\}} n^{-1/2}\nu_n\left(r\left[\psi-\psi_o,X\right]\right)=o_p\left(n^{-1}\right) \tag{A.45}$$

$$\sup_{\{\psi\in\Psi_n:0<\|\psi-\psi_o\|\le\delta_n\}} \left[K\left(\pi_n\psi_*\left(\psi,\varepsilon_n,\psi_o\right)\right)-K\left(\psi,\psi_o\right)\right]-\frac{1}{2}\left[\left\|\psi_*\left(\psi,\varepsilon_n\right)-\psi_o\right\|^2-\left\|\psi-\psi_o\right\|^2\right]=O\left(\varepsilon_n^2\right) \tag{A.46}$$

$$\sup_{\{\psi\in\Psi_n:\|\psi-\psi_o\|\le\delta_n\}} n^{-1/2}\nu_n\left(l'(\psi_o,X)\left[\psi-\psi_o\right]\right)=O_p\left(\varepsilon_n\right) \tag{A.47}$$

$$\sup_{\{\psi\in\Psi_n:0<\|\psi-\psi_o\|\le\delta_n\}} \left\|\psi_*\left(\psi,\varepsilon_n\right)-\pi_n\left(\psi_*\left(\psi,\varepsilon_n\right)\right)\right\|=O\left(\delta_n^{-1}\varepsilon_n^2\right) \tag{A.48}$$

$$\sup_{\{\psi\in\Psi_n:\|\psi-\psi_o\|\le\delta_n\}} n^{-1/2}\nu_n\left(l'(\psi_o,X)\left[\psi_*\left(\psi,\varepsilon_n\right)-\pi_n\left(\psi_*\left(\psi,\varepsilon_n\right)\right)\right]\right)=O_p\left(\varepsilon_n^2\right) \tag{A.49}$$

(A.45) can be verified by applying Theorem 3 in Chen and Shen (1998). Note that by (A.41)

$$\begin{aligned}
&\mathrm{Var}\left(\psi^*\left(X_0\right)-\psi_o^*\left(X_0\right)+\psi\left(T_o^{-1}\left(X_0\right)\right)-\psi_o\left(T_o^{-1}\left(X_0\right)\right)\right)\\
&\le E\left[\psi^*\left(X_0\right)-\psi_o^*\left(X_0\right)\right]^2+E\left[\psi\left(X_1\right)-\psi_o\left(X_1\right)\right]^2\\
&\quad+2E\left[\left(\psi^*\left(X_0\right)-\psi_o^*\left(X_0\right)\right)\times\left(\psi\left(T_o^{-1}\left(X_0\right)\right)-\psi_o\left(T_o^{-1}\left(X_0\right)\right)\right)\right]\\
&\le const.\cdot\left\|\psi-\psi_o\right\|_{2,1}^2+2\left\|\psi\left(T_o^{-1}\left(\cdot\right)\right)-\psi_o\left(T_o^{-1}\left(\cdot\right)\right)\right\|_{2,0}\cdot\left\|\psi^*-\psi_o^*\right\|_{2,0}\\
&\le const.\cdot\left\|\psi-\psi_o\right\|_{2,1}^2\le const.\cdot\left\|\psi-\psi_o\right\|^2,
\end{aligned}$$

and similar to (A.42),

$$\left|\psi^*\left(X_0\right)-\psi_o^*\left(X_0\right)+\psi\left(T_o^{-1}\left(X_0\right)\right)-\psi_o\left(T_o^{-1}\left(X_0\right)\right)\right|\le const.\cdot\left\|\psi-\psi_o\right\|_{2,1}^{2/3}\le const.\cdot\left\|\psi-\psi_o\right\|^{2/3}.$$

All the conditions in Chen and Shen (1998) Theorem 3 are satisfied and the empirical process in (A.45) is of order $O_p\left(n^{-2\alpha/(2\alpha+2+d_X)}\right)$. Hence (A.45) holds for $\alpha>1+d_X/2$.

Similarly, empirical process in (A.47) is of order $O_p\left(n^{-2\alpha/(2\alpha+2+d_X)}\right)$ and (A.47) holds

for any $\alpha \geq 0$. (A.46) is verified by the choice of pseudo-metric $\|\cdot\|$. Note that $\|\psi_*(\psi,\varepsilon_n) - \pi_n(\psi_*(\psi,\varepsilon_n))\| = \varepsilon_n \|(u^* + \psi_o) - \pi_n(u^* + \psi_o)\|$ and $n^{-\frac{\alpha+1}{2(\alpha+1)+d_X}} = O(\delta_n^{-1}\varepsilon_n)$, (A.48) holds if $\alpha > d/2$. (A.49) can be easily checked.

Step 3. We then apply Theorem 1 in Shen (1997) to get our proposition. Note that

$$
\begin{aligned}
&\frac{1}{n_1}\sum_{i=1}^{n_1}\left(h\left(\widehat{T}(X_{1i}),\beta_{o\mathcal{S}}\right) I\{X_{1i}\in\mathcal{S}\} - E\left[h\left(T_o(X_1),\beta_{o\mathcal{S}}\right) I\{X_1\in\mathcal{S}\}\right]\right)\\
&= E\left[h\left(\widehat{T}(X_1),\beta_o\right) I\{X_1\in\mathcal{S}\} - h\left(T_o(X_1),\beta_{o\mathcal{S}}\right) I\{X_1\in\mathcal{S}\}\right]\\
&\quad + \frac{1}{n_1}\sum_{i=1}^{n_1}\left(h\left(T_o(X_{1i}),\beta_{o\mathcal{S}}\right) I\{X_{1i}\in\mathcal{S}\} - E\left[h\left(T_o(X_1),\beta_{o\mathcal{S}}\right) I\{X_1\in\mathcal{S}\}\right]\right)\\
&\quad + \frac{1}{n_1}\sum_{i=1}^{n_1}\left(\left[h\left(\widehat{T}(X_{1i}),\beta_{o\mathcal{S}}\right) - h\left(T_o(X_{1i}),\beta_{o\mathcal{S}}\right)\right] I\{X_{1i}\in\mathcal{S}\}\right.\\
&\qquad \left. - E\left[\left(h\left(\widehat{T}(X_1),\beta_{o\mathcal{S}}\right) - h\left(T_o(X_1),\beta_{o\mathcal{S}}\right)\right) I\{X_1\in\mathcal{S}\}\right]\right)\\
&= E\left[\left(h\left(\widehat{T}(X_1),\beta_{o\mathcal{S}}\right) - h\left(T_o(X_1),\beta_{o\mathcal{S}}\right)\right) I\{X_1\in\mathcal{S}\}\right]\\
&\quad + \frac{1}{n_1}\sum_{i=1}^{n_1}\left(h\left(T_o(X_{1i}),\beta_{o\mathcal{S}}\right) I\{X_{1i}\in\mathcal{S}\} - E\left[h\left(T_o(X_1),\beta_{o\mathcal{S}}\right) I\{X_1\in\mathcal{S}\}\right]\right) + o_p\left(n_1^{-1/2}\right),
\end{aligned}
$$

where we have used the result that

$$
\begin{aligned}
&\frac{1}{n_1}\sum_{i=1}^{n_1}\left[h\left(\widehat{T}(X_{1i}),\beta_{o\mathcal{S}}\right) - h\left(T_o(X_{1i}),\beta_{o\mathcal{S}}\right)\right] I\{X_{1i}\in\mathcal{S}\}\\
&- E\left[\left(h\left(\widehat{T}(X_1),\beta_{o\mathcal{S}}\right) - h\left(T_o(X_1),\beta_{o\mathcal{S}}\right)\right) I\{X_1\in\mathcal{S}\}\right]\\
&= o_p\left(n^{-1/2}\right), \text{ see (A.1) in the Appendix.}
\end{aligned}
$$

It remains to show that

$$
\begin{aligned}
&E\left[\left(h\left(\widehat{T}(X_1),\beta_{o\mathcal{S}}\right) - h\left(T_o(X_1),\beta_{o\mathcal{S}}\right)\right) I\{X_1\in\mathcal{S}\}\right]\\
=\ &-\left[n_1^{-1}\sum_{i=1}^{n_1} v^*(X_{1i}) - n_0^{-1}\sum_{j=1}^{n_0} v^*\left(T_o^{-1}(X_{0j})\right) - E\left[v^*(X_1)\right] + E\left[v^*\left(T_o^{-1}(X_0)\right)\right]\right] + o_p\left(n^{-1/2}\right).
\end{aligned}
$$

Apply Theorem 1 in Shen (1997), our functional of interest is $E\left[h\left(T_o\left(X_1\right),\beta_{o\mathcal{S}}\right)I\left\{X_1\in\mathcal{S}\right\}\right]$. For any $\pi_n\psi_n\in\{\pi_n\psi_n\in\Psi_n:\|\pi_n\psi_n-\psi_o\|\leq\delta_n\}$, by (A.44) we have

$$\begin{aligned}\widehat{Q}\left(\pi_n\psi_n\right)=&\widehat{Q}\left(\psi_o\right)-K\left(\psi_o,\pi_n\psi_n\right)+n_1^{-1/2}\nu_{n_1}\left(\pi_n\psi_n-\psi_o,X_1\right)-n_0^{-1/2}\nu_{n_0}\left(\pi_n\psi_n\circ T_o^{-1}-\psi_o\circ T_o^{-1},X_0\right)\\&+n_0^{-1/2}\nu_{n_0}\left(r\left[\pi_n\psi_n-\psi_o,X_0\right]\right),\end{aligned}\tag{A.50}$$

and substitute $\pi_n\psi_n$ by $\widehat{\psi}$ we have

$$\begin{aligned}\widehat{Q}\left(\widehat{\psi}\right)=&\widehat{Q}\left(\psi_o\right)-K\left(\psi_o,\widehat{\psi}\right)+n_1^{-1/2}\nu_{n_1}\left(\widehat{\psi}-\psi_o,X_1\right)-n_0^{-1/2}\nu_{n_0}\left(\widehat{\psi}\circ T_o^{-1}-\psi_o\circ T_o^{-1},X_0\right)\\&+n_0^{-1/2}\nu_{n_0}\left(r\left[\widehat{\psi}-\psi_o,X_0\right]\right).\end{aligned}\tag{A.51}$$

Take the difference of (A.50) (A.51) and substitute $\pi_n\psi_n$ with $\psi_*\left(\psi,\varepsilon_n\right)$. By (A.45) and (A.46), we have

$$\begin{aligned}\widehat{Q}\left(\pi_n\psi_*\left(\widehat{\psi},\varepsilon_n\right)\right)=&\widehat{Q}\left(\widehat{\psi}\right)+K\left(\psi_o,\widehat{\psi}\right)-K\left(\psi_o,\pi_n\psi_*\left(\widehat{\psi},\varepsilon_n\right)\right)\\&-n_1^{-1/2}\nu_{n_1}\left(\widehat{\psi}-\pi_n\psi_*\left(\widehat{\psi},\varepsilon_n\right),X_1\right)+n_0^{-1/2}\nu_{n_0}\left(\widehat{\psi}\circ T_o^{-1}-\pi_n\psi_*\left(\widehat{\psi},\varepsilon_n\right)\circ T_o^{-1},X_0\right)\\&-n_0^{-1/2}\nu_{n_0}\left(r\left[\widehat{\psi}-\pi_n\psi_*\left(\widehat{\psi},\varepsilon_n\right),X_0\right]\right)\\=&\widehat{Q}\left(\widehat{\psi}\right)+\frac{1}{2}\left[\left\|\widehat{\psi}-\psi_o\right\|^2-\left\|\psi_*\left(\widehat{\psi},\varepsilon_n\right)-\psi_o\right\|^2\right]\\&-n_1^{-1/2}\nu_{n_1}\left(\widehat{\psi}-\pi_n\psi_*\left(\widehat{\psi},\varepsilon_n\right),X_1\right)+n_0^{-1/2}\nu_{n_0}\left(\widehat{\psi}\circ T_o^{-1}-\pi_n\psi_*\left(\widehat{\psi},\varepsilon_n\right)\circ T_o^{-1},X_0\right)\\&+O_p\left(\varepsilon_n^2\right).\end{aligned}$$

By Condition (A.48) and (2.20), we have

$$\begin{aligned}-O_p\left(\varepsilon_n^2\right)\leq&\frac{1}{2}\left[\left\|\widehat{\psi}-\psi_o\right\|^2-\left\|\pi_n\psi_*\left(\widehat{\psi},\varepsilon_n\right)-\psi_o\right\|^2\right]\\&-n_1^{-1/2}\nu_{n_1}\left(\widehat{\psi}-\psi_*\left(\widehat{\psi},\varepsilon_n\right),X_1\right)+n_0^{-1/2}\nu_{n_0}\left(\widehat{\psi}\circ T_o^{-1}-\psi_*\left(\widehat{\psi},\varepsilon_n\right)\circ T_o^{-1},X_0\right)\\&+O_p\left(\varepsilon_n^2\right).\end{aligned}\tag{A.52}$$

By (A.48) and (A.49),

$$\begin{aligned}
\left\|\pi_n\psi_*\left(\widehat{\psi},\varepsilon_n\right)-\psi_o\right\|^2 &= \left\|\pi_n\psi_*\left(\widehat{\psi},\varepsilon_n\right)-\psi_*\left(\widehat{\psi},\varepsilon_n\right)\right\|^2+\left\|\psi_*\left(\widehat{\psi},\varepsilon_n\right)-\psi_o\right\|^2\\
&\quad+2\left\|\pi_n\psi_*\left(\widehat{\psi},\varepsilon_n\right)-\psi_*\left(\widehat{\psi},\varepsilon_n\right)\right\|\left\|\psi_*\left(\widehat{\psi},\varepsilon_n\right)-\psi_o\right\|\\
&= \left\|\pi_n\psi_*\left(\widehat{\psi},\varepsilon_n\right)-\psi_*\left(\widehat{\psi},\varepsilon_n\right)\right\|^2+(1-\varepsilon_n)^2\left\|\widehat{\psi}-\psi_o\right\|^2\\
&\quad+2(1-\varepsilon_n)\left\|\pi_n\psi_*\left(\widehat{\psi},\varepsilon_n\right)-\psi_*\left(\widehat{\psi},\varepsilon_n\right)\right\|\left\|\widehat{\psi}-\psi_o\right\|\\
&\quad+2\varepsilon_n\left\|\pi_n\psi_*\left(\widehat{\psi},\varepsilon_n\right)-\psi_*\left(\widehat{\psi},\varepsilon_n\right)\right\|\|u^*\|+\varepsilon_n^2\|u^*\|^2\\
&\quad+2(1-\varepsilon_n)\langle\widehat{\psi}-\psi_o,\varepsilon_n u^*\rangle\\
&\geq (1-\varepsilon_n)^2\left\|\widehat{\psi}-\psi_o\right\|^2+2(1-\varepsilon_n)\left\|\pi_n\psi_*\left(\widehat{\psi},\varepsilon_n\right)-\psi_*\left(\widehat{\psi},\varepsilon_n\right)\right\|\left\|\widehat{\psi}-\psi_o\right\|\\
&\quad+2(1-\varepsilon_n)\langle\widehat{\psi}-\psi_o,\varepsilon_n u^*\rangle+O\left(\varepsilon_n^2\right). \qquad \text{(A.53)}
\end{aligned}$$

By (A.52), (A.53) and (A.47), (A.48), (A.49), we have

$$\begin{aligned}
-O_p\left(\varepsilon_n^2\right) \leq& \left(\frac{1}{2}-\frac{1}{2}(1-\varepsilon_n)^2\right)\left\|\widehat{\psi}-\psi_o\right\|^2-(1-\varepsilon_n)\langle\widehat{\psi}-\psi_o,\varepsilon_n u^*\rangle\\
&-(1-\varepsilon_n)\left\|\pi_n\psi_*\left(\widehat{\psi},\varepsilon_n\right)-\psi_*\left(\widehat{\psi},\varepsilon_n\right)\right\|\left\|\widehat{\psi}-\psi_o\right\|\\
&-n_1^{-1/2}\nu_{n_1}\left(\widehat{\psi}-\psi_*\left(\widehat{\psi},\varepsilon_n\right),X_1\right)+n_0^{-1/2}\nu_{n_0}\left(\widehat{\psi}\circ T_o^{-1}-\psi_*\left(\widehat{\psi},\varepsilon_n\right)\circ T_o^{-1},X_0\right)\\
&+O_p\left(\varepsilon_n^2\right)\\
\leq& \varepsilon_n\left\|\widehat{\psi}-\psi_o\right\|^2-(1-\varepsilon_n)\langle\widehat{\psi}-\psi_o,\varepsilon_n u^*\rangle-\left\|\pi_n\psi_*\left(\widehat{\psi},\varepsilon_n\right)-\psi_*\left(\widehat{\psi},\varepsilon_n\right)\right\|\left\|\widehat{\psi}-\psi_o\right\|\\
&-n_1^{-1/2}\nu_{n_1}\left(\widehat{\psi}-\psi_*\left(\widehat{\psi},\varepsilon_n\right),X_1\right)+n_0^{-1/2}\nu_{n_0}\left(\widehat{\psi}\circ T_o^{-1}-\psi_*\left(\widehat{\psi},\varepsilon_n\right)\circ T_o^{-1},X_0\right)\\
&+O_p\left(\varepsilon_n^2\right). \qquad \text{(A.54)}
\end{aligned}$$

By definition we have

$$n_1^{-1/2}\nu_{n_1}\left(\widehat{\psi}-\psi_*\left(\widehat{\psi},\varepsilon_n\right),X_1\right)=n_1^{-1/2}\nu_{n_1}\left(\varepsilon_n\left(\widehat{\psi}-\psi_o\right)-\varepsilon_n u^*,X_1\right), \qquad \text{(A.55)}$$

$$n_0^{-1/2}\nu_{n_0}\left(\widehat{\psi}\circ T_o^{-1}-\psi_*\left(\widehat{\psi},\varepsilon_n\right)\circ T_o^{-1},X_0\right)=n_0^{-1/2}\nu_{n_0}\left(\varepsilon_n\left(\widehat{\psi}-\psi_o\right)-\varepsilon_n u^*,T_o^{-1}\left(X_0\right)\right). \tag{A.56}$$

Combine (A.55) (A.56) with (A.54) we have

$$\begin{aligned}-O_p\left(\varepsilon_n^2\right)\leq&\varepsilon_n\left\|\widehat{\psi}-\psi_o\right\|^2-\left(1-\varepsilon_n\right)\langle\widehat{\psi}-\psi_o,\varepsilon_n u^*\rangle-\left\|\pi_n\psi_*\left(\widehat{\psi},\varepsilon_n\right)-\psi_*\left(\widehat{\psi},\varepsilon_n\right)\right\|\left\|\widehat{\psi}-\psi_o\right\|\\&-n_1^{-1/2}\nu_{n_1}\left(\varepsilon_n u^*,X_1\right)+n_0^{-1/2}\nu_{n_0}\left(\varepsilon_n u^*,T_o^{-1}\left(X_0\right)\right)+O_p\left(\varepsilon_n^2\right)\\\leq&-\left(1-\varepsilon_n\right)\langle\widehat{\psi}-\psi_o,\varepsilon_n u^*\rangle-n_1^{-1/2}\nu_{n_1}\left(\varepsilon_n u^*,X_1\right)+n_0^{-1/2}\nu_{n_0}\left(\varepsilon_n u^*,T_o^{-1}\left(X_0\right)\right)+O_p\left(\varepsilon_n^2\right).\end{aligned} \tag{A.57}$$

Hence,

$$-o_p\left(n^{-1/2}\right)=-O_p\left(\varepsilon_n^2\right)\leq-\left(1-\varepsilon_n\right)\langle\widehat{\psi}-\psi_o,v^*\rangle-n_1^{-1/2}\nu_{n_1}\left(v^*,X_1\right)+n_0^{-1/2}\nu_{n_0}\left(v^*,T_o^{-1}\left(X_0\right)\right),$$

and for $u^*=-v^*$ we have

$$o_p\left(n^{-1/2}\right)\geq-\left(1-\varepsilon_n\right)\langle\widehat{\psi}-\psi_o,v^*\rangle-n_1^{-1/2}\nu_{n_1}\left(v^*,X_1\right)+n_0^{-1/2}\nu_{n_0}\left(v^*,T_o^{-1}\left(X_0\right)\right),$$

give us

$$\langle\widehat{\psi}-\psi_o,v^*\rangle=-n_1^{-1/2}\nu_{n_1}\left(v^*,X_1\right)+n_0^{-1/2}\nu_{n_0}\left(v^*,T_o^{-1}\left(X_0\right)\right)+o_p\left(n^{-1/2}\right).$$

By Condition 5.1, we have

$$\begin{aligned}&E\left[h\left(\widehat{T}\left(X_1\right),\beta_{o\mathcal{S}}\right)I\left\{X_1\in\mathcal{S}\right\}\right]-E\left[h\left(T_o\left(X_1\right),\beta_{o\mathcal{S}}\right)I\left\{X_1\in\mathcal{S}\right\}\right]\\&=\Gamma(h,\psi_o)[\widehat{\psi}-\psi_o]+o_p\left(O\left(\left\|\widehat{\psi}-\psi_o\right\|^{\omega}\right)\right)\\&=\langle\widehat{\psi}-\psi_o,v^*\rangle+o_p\left(n^{-1/2}\right)\\&=-n_1^{-1/2}\nu_{n_1}\left(v^*,X_1\right)+n_0^{-1/2}\nu_{n_0}\left(v^*,T_o^{-1}\left(X_0\right)\right)+o_p\left(n^{-1/2}\right).\end{aligned}$$

□

Appendix B

APPENDICES FOR CORRECTING STRATEGIC MISREPORTING BEHAVIOR ON OUTCOME IN ESTIMATING TREATMENT EFFECT

B.1 Smoothness Class For Functions

To account for the effect of estimating $\boldsymbol{\mu_o}, \boldsymbol{\mu_o^*}, \boldsymbol{h}$, we first provide some definitions of the function space they belong to. Let $\boldsymbol{g} : \mathbb{R}^{d_X} \to \mathbb{R}^{d_Y}$, $\widehat{\boldsymbol{g}}$ is the sieve Least Squares estimator $\boldsymbol{g}$, we follow Chen et al. (2005) by imposing smoothness on $\boldsymbol{g}(\cdot)$. A typical smoothness assumption is that a function belongs to a Hölder space. For any $1 \times d_X$ vector $\mathbf{a} = (a_1, \ldots, a_{d_X})$ of non-negative integers, we write $|\mathbf{a}| = \sum_{k=1}^{d_X} a_k$, and for any $x = (x_1, \ldots, x_{d_X})' \in \mathcal{X} \subseteq \mathbb{R}^{d_X}$, we denote the $|\mathbf{a}|$-th derivative of a function $g : \mathbb{R}^{d_X} \to \mathbb{R}$ as

$$\nabla^{\mathbf{a}} g(x) = \frac{\partial^{|\mathbf{a}|}}{\partial x_1^{a_1} \ldots \partial x_d^{a_{d_X}}} g(x).$$

For some $\gamma > 0$, let $\underline{\gamma}$ be the largest integer smaller than γ, and let $\Lambda^{\gamma}(\mathcal{X})$ denote a Hölder space with smoothness γ, i.e., a space of functions $g : \mathcal{X} \to \mathbb{R}$ which have up to $\underline{\gamma}$-th continuous derivatives, and the highest $\underline{\gamma}$-th derivatives are Hölder continuous with the Hölder exponent $\gamma - \underline{\gamma} \in (0, 1]$. The Hölder space becomes a Banach space when endowed with the Hölder norm:

$$\|g\|_{\Lambda^{\gamma}} = \sup_x |g(x)| + \max_{|\mathbf{a}|=\underline{\gamma}} \sup_{x \neq \overline{x}} \frac{|\nabla^{\mathbf{a}} g(x) - \nabla^{\mathbf{a}} g(\overline{x})|}{\sqrt{(x-\overline{x})'(x-\overline{x})}^{\gamma-\underline{\gamma}}} < \infty.$$

Following Chen, Hong, and Tamer (2005), we let $\Lambda^{\gamma}(\mathcal{X}, \omega_1)$ denote a weighted Hölder

space of functions $g : \mathcal{X} \to \mathbb{R}$ such that $g(\cdot)\left[1 + |\cdot|^2\right]^{-\omega_1/2}$ is in $\Lambda^\gamma(\mathcal{X})$. We call

$$\Lambda_c^\gamma(\mathcal{X}, \omega_1) \equiv \left\{ g \in \Lambda^\gamma(\mathcal{X}, \omega_1) : \left\| g(\cdot)\left[1 + |\cdot|^2\right]^{-\omega_1/2} \right\|_{\Lambda^\gamma} \leq c < \infty \right\}$$

a weighted Hölder ball (with radius c). We say a function $g : \mathcal{X} \to \mathbb{R}$ is $H(\gamma, \omega_1)$-smooth if it belongs to a weighted Hölder ball $\Lambda_c^\gamma(\mathcal{X}, \omega_1)$ for some $\gamma > 0$ and $\omega_1 \geq 0$. As discussed in Chen, Hong, and Tamer (2005), the weighted Holder ball with $\omega_1 = 0$ reduces to the standard Hölder ball $\Lambda_c^\gamma(\mathcal{X})$, which is a typical sufficient condition especially when the support $\mathcal{X}$ is a bounded subset of $\mathbb{R}^{d_X}$. However, when $\mathcal{X} = \mathbb{R}^{d_X}$, the standard Hölder ball $\Lambda_c^\gamma(\mathcal{X})$ may exclude some functions such as $g(x) = x'x$. It is clear that the weighted Hölder ball $\Lambda_c^\gamma(\mathcal{X}, \omega_1)$ with $\omega_1 > 0$ is a strictly larger space and $x'x \in \Lambda_c^\gamma\left(\mathbb{R}^{d_X}, \omega_1\right)$ with $\omega_1 = 2$.

For $j = 0, 1$, for any square measurable function $g : \mathcal{X}_j \to \mathbb{R}$, we define a Hilbert norm $\|g\|_{2,P(X_j)} \equiv \sqrt{\int_{\mathcal{X}_j} g(x)^2 f_{X_j}(x) dx} < \infty$, and $\mathcal{L}_2(\mathcal{X}_j) = \left\{ g : \mathcal{X}_j \to \mathbb{R} : \|g\|_{2,P(X_j)} < \infty \right\}$ the corresponding Hilbert space. For $g(x)$, $x \in \mathcal{X}_1$, we denote

$$\|g\|_{\infty,\omega} = \sup_{x \in \mathcal{X}_1} \left| g(x)\left[1 + |x|^2\right]^{-\omega/2} \right|.$$

B.2 Proof for Proposition 7

Proof. Denote $\eta_1 = \left(\boldsymbol{\mu}_{\boldsymbol{o}}', \frac{f_{X_0}}{f_{X_1}}\right)'$. According to the definition of Neyman orthogonal moment described in Chernozhukov et al. (2018), we need to prove

$$\frac{\partial}{\partial \eta_1}\left\{ E\left[\left(Y_1 - \boldsymbol{\mu}_o(X_1)\right) \frac{f_{X_0}(X_1)}{f_{X_1}(X_1)} \right] + E\left[\boldsymbol{\mu}_{\boldsymbol{o}}(X_0) - Y_0\right] \right\} = 0. \tag{B.1}$$

By calculation, we have

$$\begin{aligned} &\frac{\partial}{\partial \tau}\left\{ E\left[\left(Y_1 - \boldsymbol{\mu}_o(X_1) - \tau v(X_1)\right) \frac{f_{X_0}(X_1)}{f_{X_1}(X_1)} \right] + E\left[\boldsymbol{\mu}_{\boldsymbol{o}}(X_0) + \tau v(X_0) - Y_0\right] \right\} \Big|_{\tau=0} \\ &= -E\left[v(X_1) \frac{f_{X_0}(X_1)}{f_{X_1}(X_1)} \right] + E\left[v(X_0)\right] = 0, \end{aligned}$$

and

$$\begin{aligned}
&\frac{\partial}{\partial\tau}\left\{E\left[\left(Y_1-\boldsymbol{\mu}_o\left(X_1\right)\right)\left(\frac{f_{X_0}\left(X_1\right)}{f_{X_1}\left(X_1\right)}+\tau v\left(X_1\right)\right)\right]+E\left[\boldsymbol{\mu_o}\left(X_0\right)+\tau v\left(X_0\right)-Y_0\right]\right\}\Big|_{\tau=0}\\
&=E\left[\left(Y_1-\boldsymbol{\mu}_o\left(X_1\right)\right)v\left(X_1\right)\right]\\
&=E\left[\left(E[Y_1|X_1]-\boldsymbol{\mu}_o\left(X_1\right)\right)v\left(X_1\right)\right]=0.
\end{aligned}$$

□

B.3 Proof for Proposition 8

Proof. Similar to proof of Proposition 7, by calculation we have

$$\begin{aligned}
&\frac{\partial}{\partial\boldsymbol{\mu}_{\boldsymbol{o}}^*}\left\{E\left[\boldsymbol{\mu}_{\boldsymbol{o}}^*(X_0)\right]-E\left[Y_0\right]+E\left[\boldsymbol{h_o}\left(Z_0\right)\right]+E\left\{\nu_1\left(X_1\right)\left(Y_1^*-\boldsymbol{\mu}_{\boldsymbol{o}}^*\left(X_1\right)\right)+\nu_2\left(Z_1\right)\left(W_1-\boldsymbol{h_o}\left(Z_1\right)\right)\right\}\right\}\\
&=\frac{\partial}{\partial\tau}\left\{\begin{array}{c}E\left[\boldsymbol{\mu}_{\boldsymbol{o}}^*(X_0)+\tau v(X_0)\right]-E\left[Y_0\right]+E\left[\boldsymbol{h_o}\left(Z_0\right)\right]+E\left\{\nu_1\left(X_1\right)\left(Y_1^*-\boldsymbol{\mu}_{\boldsymbol{o}}^*\left(X_1\right)-\tau v(X_1)\right)\right\}\\ +E\left\{\nu_2\left(Z_1\right)\left(W_1-\boldsymbol{h_o}\left(Z_1\right)\right)\right\}\end{array}\right\}\Big|_{\tau=0}\\
&=E\left[v(X_0)\right]+E\left\{\nu_1\left(X_1\right)\left(-v(X_1)\right)\right\}\\
&=E\left[v(X_0)\right]+E\left\{\frac{f_{X_0}\left(X_1\right)}{f_{X_1}\left(X_1\right)}\left(-v(X_1)\right)\right\}=0,
\end{aligned}$$

$$\begin{aligned}
&\frac{\partial}{\partial\boldsymbol{h_o}}\left\{E\left[\boldsymbol{\mu}_{\boldsymbol{o}}^*(X_0)\right]-E\left[Y_0\right]+E\left[\boldsymbol{h_o}\left(Z_0\right)\right]+E\left\{\nu_1\left(X_1\right)\left(Y_1^*-\boldsymbol{\mu}_{\boldsymbol{o}}^*\left(X_1\right)\right)+\nu_2\left(Z_1\right)\left(W_1-\boldsymbol{h_o}\left(Z_1\right)\right)\right\}\right\}\\
&=\frac{\partial}{\partial\tau}\left\{\begin{array}{c}E\left[\boldsymbol{\mu}_{\boldsymbol{o}}^*(X_0)\right]-E\left[Y_0\right]+E\left[\boldsymbol{h_o}\left(Z_0\right)+\tau v(Z_0)\right]+E\left\{\nu_1\left(X_1\right)\left(Y_1^*-\boldsymbol{\mu}_{\boldsymbol{o}}^*\left(X_1\right)\right)\right\}\\ +E\left\{\nu_2\left(Z_1\right)\left(W_1-\boldsymbol{h_o}\left(Z_1\right)-\tau v(Z_1)\right)\right\}\end{array}\right\}\Big|_{\tau=0}\\
&=E\left[v\left(Z_0\right)\right]+E\left\{\nu_2\left(Z_1\right)\left(-v\left(Z_1\right)\right)\right\}\\
&=E\left[v\left(Z_0\right)\right]+E\left\{\frac{f_{Z_0}\left(Z_1\right)}{f_{Z_1}\left(Z_1\right)}\left(-v\left(Z_1\right)\right)\right\}=0,
\end{aligned}$$

$$\frac{\partial}{\partial \nu_1}\left\{E\left[\boldsymbol{\mu}_{\boldsymbol{o}}^{*}(X_0)\right]-E\left[Y_0\right]+E\left[\boldsymbol{h_o}\left(Z_0\right)\right]+E\left\{\nu_1\left(X_1\right)\left(Y_1^{*}-\boldsymbol{\mu}_{\boldsymbol{o}}^{*}\left(X_1\right)\right)+\nu_2\left(Z_1\right)\left(W_1-\boldsymbol{h_o}\left(Z_1\right)\right)\right\}\right\}$$
$$=\frac{\partial}{\partial \tau}\left\{E\left[\boldsymbol{\mu}_{\boldsymbol{o}}^{*}(X_0)\right]-E\left[Y_0\right]+E\left[\boldsymbol{h_o}\left(Z_0\right)\right]+E\left\{\left(\nu_1+\tau v\right)\left(X_1\right)\left(Y_1^{*}-\boldsymbol{\mu}_{\boldsymbol{o}}^{*}\left(X_1\right)\right)+\nu_2\left(Z_1\right)\left(W_1-\boldsymbol{h_o}\left(Z_1\right)\right)\right\}\right\}\Big|_{\tau}$$
$$=E\left\{v\left(X_1\right)\left(Y_1^{*}-\boldsymbol{\mu}_{\boldsymbol{o}}^{*}\left(X_1\right)\right)\right\}=0,$$

and

$$\frac{\partial}{\partial \nu_2}\left\{E\left[\boldsymbol{\mu}_{\boldsymbol{o}}^{*}(X_0)\right]-E\left[Y_0\right]+E\left[\boldsymbol{h_o}\left(Z_0\right)\right]+E\left\{\nu_1\left(X_1\right)\left(Y_1^{*}-\boldsymbol{\mu}_{\boldsymbol{o}}^{*}\left(X_1\right)\right)+\nu_2\left(Z_1\right)\left(W_1-\boldsymbol{h_o}\left(Z_1\right)\right)\right\}\right\}$$
$$=\frac{\partial}{\partial \tau}\left\{E\left[\boldsymbol{\mu}_{\boldsymbol{o}}^{*}(X_0)\right]-E\left[Y_0\right]+E\left[\boldsymbol{h_o}\left(Z_0\right)\right]+E\left\{\nu_1\left(X_1\right)\left(Y_1^{*}-\boldsymbol{\mu}_{\boldsymbol{o}}^{*}\left(X_1\right)\right)+\left(\nu_2+\tau v\right)\left(Z_1\right)\left(W_1-\boldsymbol{h_o}\left(Z_1\right)\right)\right\}\right\}\Big|_{\tau}$$
$$=E\left\{v\left(Z_1\right)\left(W_1-\boldsymbol{h_o}\left(Z_1\right)\right)\right\}=0.$$

□

B.4 Proof for Proposition 9

$$\begin{aligned}\tau_{ATU,S3}^{*}=&E\left[\boldsymbol{\mu}_{\boldsymbol{o}}^{*}\left(X_0\right)+\boldsymbol{h_o}\left(X_0,T_o(Y_0)\right)-Y_0+\varphi_0\left(Y_0\right)-E\left(\varphi_0\left(Y_0\right)\right)\right]\\&+E\begin{bmatrix}\frac{f_{X_0}(X_1)}{f_{X_1}(X_1)}\left(Y_1^{*}-\boldsymbol{\mu}_{\boldsymbol{o}}^{*}\left(X_1\right)\right)\\+\frac{f_{X_0,Y_0}\left(X_1,T_o^{-1}(Y_1)\right)}{f_{Y_0}\left(T_o^{-1}(Y_1)\right)}\frac{f_{Y_1}(Y_1)}{f_{X_1,Y_1}(X_1,Y_1)}\left(W_1-\boldsymbol{h_o}\left(Z_1\right)\right)\\+\varphi_1\left(Y_1\right)-E\left(\varphi_1\left(Y_1\right)\right)\end{bmatrix},\end{aligned}$$

where the form of $\varphi_1(\cdot)$ and $\varphi_0(\cdot)$ is given in the next section for different cases.

Proof. Most of the calculation is the same as the proof of Proposition 8. We only need to

show the pathwise derivative with respect to $\boldsymbol{h_o}$ and T_o equals zero.

$$
\begin{aligned}
&\frac{\partial}{\partial \boldsymbol{h_o}}\left\{\begin{array}{c} E\left[\boldsymbol{\mu_o^*}\left(X_0\right)+\boldsymbol{h_o}\left(X_0, T_o(Y_0)\right)-Y_0+\varphi_0\left(Y_0\right)-E\left(\varphi_0\left(Y_0\right)\right)\right] \\ +E\left[\frac{f_{X_0}\left(X_1\right)}{f_{X_1}\left(X_1\right)}\left(Y_1^*-\boldsymbol{\mu_o^*}\left(X_1\right)\right)\right] \\ +E\left[\frac{f_{X_0,Y_0}\left(X_1, T_o^{-1}(Y_1)\right)}{f_{Y_0}\left(T_o^{-1}(Y_1)\right)} \frac{f_{Y_1}\left(Y_1\right)}{f_{X_1,Y_1}\left(X_1,Y_1\right)}\left(W_1-\boldsymbol{h_o}\left(Z_1\right)\right)\right] \\ +E\left[\varphi_1\left(Y_1\right)-E\left(\varphi_1\left(Y_1\right)\right)\right] \end{array}\right\} \\
&=\frac{\partial}{\partial \tau}\left\{\begin{array}{c} E\left[\boldsymbol{\mu_o^*}\left(X_0\right)+\left(\boldsymbol{h_o}+\tau v\right)\left(X_0, T_o(Y_0)\right)-Y_0+\varphi_0\left(Y_0\right)-E\left(\varphi_0\left(Y_0\right)\right)\right] \\ +E\left[\frac{f_{X_0}\left(X_1\right)}{f_{X_1}\left(X_1\right)}\left(Y_1^*-\boldsymbol{\mu_o^*}\left(X_1\right)\right)\right] \\ +E\left[\frac{f_{X_0,Y_0}\left(X_1, T_o^{-1}(Y_1)\right)}{f_{Y_0}\left(T_o^{-1}(Y_1)\right)} \frac{f_{Y_1}\left(Y_1\right)}{f_{X_1,Y_1}\left(X_1,Y_1\right)}\left(W_1-\boldsymbol{h_o}\left(Z_1\right)-\tau v\left(Z_1\right)\right)\right] \\ +E\left[\varphi_1\left(Y_1\right)-E\left(\varphi_1\left(Y_1\right)\right)\right] \end{array}\right\}\Bigg|_{\tau=0} \\
&=E\left[v\left(X_0, T_o(Y_0)\right)\right]+E\left[\frac{f_{X_0,Y_0}\left(X_1, T_o^{-1}\left(Y_1\right)\right)}{f_{Y_0}\left(T_o^{-1}\left(Y_1\right)\right)} \frac{f_{Y_1}\left(Y_1\right)}{f_{X_1,Y_1}\left(X_1, Y_1\right)}\left(-v\left(Z_1\right)\right)\right] \\
&=E\left[v\left(X_0, T_o(Y_0)\right)\right]+\int \int \frac{f_{X_0,Y_0}\left(x, T_o^{-1}\left(y\right)\right)}{f_{Y_0}\left(T_o^{-1}\left(y\right)\right)} \frac{f_{Y_1}\left(y\right)}{f_{X_1,Y_1}\left(x, y\right)}\left(-v\left(x, y\right)\right) f_{X_1,Y_1}\left(x, y\right) dxdy \\
&=E\left[v\left(X_0, T_o(Y_0)\right)\right]+\int \int \frac{f_{X_0,Y_0}\left(x, T_o^{-1}\left(y\right)\right)}{f_{Y_0}\left(T_o^{-1}\left(y\right)\right)} \frac{f_{Y_1}\left(y\right)}{f_{X_1,Y_1}\left(x, y\right)}\left(-v\left(x, y\right)\right) f_{X_1,Y_1}\left(x, y\right) dxdy \\
&=E\left[v\left(X_0, T_o(Y_0)\right)\right]+\int \int \frac{f_{X_0,Y_0}\left(x, T_o^{-1}\left(y\right)\right)}{f_{Y_0}\left(T_o^{-1}\left(y\right)\right)} f_{Y_1}\left(y\right)\left(-v\left(x, y\right)\right) dxdy \\
&=E\left[v\left(X_0, T_o(Y_0)\right)\right]-\int\left(\int f_{X_0,Y_0}\left(x, T_o^{-1}\left(y\right)\right) v\left(x, y\right) dx\right) \frac{1}{f_{Y_0}\left(T_o^{-1}\left(y\right)\right)} f_{Y_1}\left(y\right) dy \\
&=E\left[v\left(X_0, T_o(Y_0)\right)\right]-\int\left(\int f_{X_0,Y_0}\left(x, y\right) v\left(x, T_o\left(y\right)\right) dx\right) \frac{1}{f_{Y_0}\left(y\right)} f_{Y_0}\left(y\right) dy \\
&=E\left[v\left(X_0, T_o(Y_0)\right)\right]-\int \int f_{X_0,Y_0}\left(x, y\right) v\left(x, T_o\left(y\right)\right) dxdy=0,
\end{aligned}
$$

where the equality in the last two lines holds due to Monge–Ampère equation.

Recall that $T_o(Y_0) = \nabla\psi_o(Y_0)$ and $\varphi_0\left(Y_0\right) = -[v^*\left(Y_0\right) - E(v^*\left(Y_0\right))]$, $\varphi_1\left(Y_1\right) =$

$v^*\left(T_o^{-1}\left(Y_1\right)\right) - E\left[v^*\left(T_o^{-1}\left(Y_1\right)\right)\right]$, where $v^*(\cdot)$ is given in (2.29). By calculation we have

$$
\begin{aligned}
&\frac{\partial}{\partial \psi_o}\left\{\begin{array}{c} E\left[\boldsymbol{\mu_o^*}\left(X_0\right)+\boldsymbol{h_o}\left(X_0, \nabla\psi_o(Y_0)\right)-Y_0+\varphi_0\left(Y_0\right)-E\left(\varphi_0\left(Y_0\right)\right)\right] \\ +E\left[\frac{f_{X_0}(X_1)}{f_{X_1}(X_1)}\left(Y_1^*-\boldsymbol{\mu_o^*}\left(X_1\right)\right)\right] \\ +E\left[\frac{f_{X_0,Y_0}\left(X_1,\nabla\psi_o^{-1}(Y_1)\right)}{f_{Y_0}\left(\nabla\psi_o^{-1}(Y_1)\right)}\frac{f_{Y_1}(Y_1)}{f_{X_1,Y_1}(X_1,Y_1)}\left(W_1-\boldsymbol{h_o}\left(Z_1\right)\right)\right] \\ +E\left[\varphi_1\left(Y_1\right)-E\left(\varphi_1\left(Y_1\right)\right)\right] \end{array}\right\} \\
&=\frac{\partial}{\partial \tau}\left\{E\left[\boldsymbol{h_o}\left(X_0, \nabla(\psi_o+\tau v)(Y_0)\right)\right]+E\left[v^*\left(\left(\nabla(\psi_o+\tau v)\right)^{-1}\left(Y_1\right)\right)\right]\right\}|_{\tau=0} \\
&=E\left[\boldsymbol{h_{o,2}}\left(X_0, \nabla\psi_o(Y_0)\right)\nabla v(Y_0)\right]-E\left[\left\{\nabla v^*\left(Y_1\right)\right\}'\left[\nabla^2\psi_o\left(Y_1\right)\right]^{-1}\nabla v\left(Y_1\right)\right] \\
&=E\left[\boldsymbol{h_{o,2}}\left(X_0, \nabla\psi_o(Y_0)\right)\nabla v(Y_0)\right]-E\left[\left\{E_{X_0}\left[\frac{h_{o,2}\left(X_0, T_o(Y_1)\right)f_{X_0,Y_0}(X_0,Y_1)}{f_{Y_0}(Y_1)f_{X_0}(X_0)}\nabla T_o(Y_1)\right]\right\}\left[\nabla^2\psi_o\left(Y_1\right)\right]^{-1}\nabla v\right. \\
&=0
\end{aligned}
$$

□

B.5 *Proof for Theorem 4*

Proof. By (2.5), Assumption, 13.1, 13.3-13.6 and Proposition A1 in Chen et al. (2005), we have $\|\widehat{\boldsymbol{\mu}}-\boldsymbol{\mu_o}\|_{\infty,\omega}=o_p\left(1\right)$, and $\|\widehat{\boldsymbol{\mu}}-\boldsymbol{\mu_o}\|_{2,P(X_1)}=O_p\left(\sqrt{\frac{k_{n_1}}{n_1}}+\left(k_{n_1}\right)^{-\gamma/d_X}\right)$. By choosing $k_{n_1}=n_1^{d_X/(d_X+2\gamma)}$, we have

$$
\|\widehat{\boldsymbol{\mu}}-\boldsymbol{\mu_o}\|_{2,P(X_1)}=O_p\left(n_1^{-\gamma/(d_X+2\gamma)}\right). \tag{B.2}
$$

By Newey (1994) and (2.5), Assumption 14, we have $\widehat{\tau}^*_{ATU,S1}-\tau^*_{ATU,S1}=o_p(1)$. □

B.6 *Proof for Theorem 5*

Proof.

$$
\sqrt{n_1}\left(\widehat{\tau}^*_{ATU,S1}-\tau^*_{ATU,S1}\right)=\sqrt{n_1}\left\{\frac{1}{n_0}\sum_{j=1}^{n_0}\widehat{\boldsymbol{\mu}}\left(X_{0j}\right)-E\left[\boldsymbol{\mu_o}\left(X_0\right)\right]-\left(\frac{1}{n_0}\sum_{j=1}^{n_0}Y_{0j}-E\left[Y_0\right]\right)\right\}
$$

$$= \sqrt{n_1} \int (\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu_o}) \, d\left(\widehat{F}_{X_0} - F_{X_0}\right) + \sqrt{n_1} \int (\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu_o}) \, dF_{X_0}$$
$$+ \sqrt{n_1} \int \boldsymbol{\mu_o} d\left(\widehat{F}_{X_0} - F_{X_0}\right) - \sqrt{n_1} \left(\frac{1}{n_0} \sum_{j=1}^{n_0} Y_{0j} - E\left[Y_0\right] \right).$$

In the following, we establish

$$\sqrt{n_1} \int (\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu_o}) \, d\left(\widehat{F}_{X_0} - F_{X_0}\right) = o_p\left(1\right) \tag{B.3}$$

$$\sqrt{n_1} \int (\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu_o}) \, dF_{X_0} = \frac{1}{\sqrt{n_1}} \sum_{i=1}^{n_1} \left(Y_{1i} - \boldsymbol{\mu}_o\left(X_{1i}\right)\right) \frac{f_{X_0}\left(X_{1i}\right)}{f_{X_1}\left(X_{1i}\right)} + o_p\left(1\right). \tag{B.4}$$

By Assumption 13.1, 13.4 and 14.3, the class $\{\mu\left(\cdot\right) : \mu\left(\cdot\right) \in \Lambda_c^{\gamma}\left(\mathcal{X}, \omega_1'\right)\}$ is a F_{X_0}-Donsker class (see Chen et al. (2005)). Thus we have

$$\sup_{\widetilde{\boldsymbol{\mu}}(\cdot) \in \Lambda_c^{\gamma}\left(\mathcal{X}, \omega_1'\right) : \|\widetilde{\boldsymbol{\mu}}(\cdot) - \boldsymbol{\mu}(\cdot)\|_{2, P(X_0)} = o(1)} \left| \int \left[\widetilde{\boldsymbol{\mu}}\left(x\right) - \boldsymbol{\mu}\left(x\right)\right] d\left[\widehat{F}_{X_0}(x) - F_{X_0}(x)\right] \right| = o_p\left(\frac{1}{\sqrt{n_0}}\right),$$

together with the Assumption 13.2, we establish (B.3).

To establish (B.4), recall $\boldsymbol{\mu_o}\left(x\right) = E\left[Y_1 | x, D = 1\right]$ and $\widehat{\boldsymbol{\mu}}$ solves

$$\widehat{\boldsymbol{\mu}} = \arg\min_{\boldsymbol{\mu} \in \mathcal{M}_n} \frac{1}{2n_1} \sum_{i=1}^{n_1} \left(Y_{1i} - \boldsymbol{\mu}\left(X_{1i}\right)\right)^2$$

where $\mathcal{M}_n$ increases with sample size n_1, and is dense in $\Lambda_c^{\gamma}\left(\mathcal{X}, \omega_1\right)$ as $k_{n_1} \to \infty$. $\boldsymbol{\mu_o}$ solves

$$\boldsymbol{\mu_o} = \arg\min_{\boldsymbol{\mu} \in \mathcal{M}} E\left[\frac{1}{2}\left(Y_1 - \boldsymbol{\mu}\left(X_1\right)\right)^2\right].$$

We first define a weak norm for the perturbation space. Let $\mathcal{M}$ be endowed with a pseudo-metric $\|\cdot\|_{2, P(X_1)}$. Let $\mathcal{V}$ be the closed linear span of $\mathcal{M} - \{\boldsymbol{\mu_o}\}$ under norm $\|\cdot\|$, where will be described later. Since $\boldsymbol{\mu_o}$ is the unique minimizer of $Q(\boldsymbol{\mu})$ over $\mathcal{M}$, within any shrinking $\|\cdot\|_{2, P(X_1)}$-neighborhood, $\mathcal{M}_o$ of μ_o, we can define a local pseudo-metric for

$\nu = \boldsymbol{\mu} - \boldsymbol{\mu_o}$ as

$$\|\nu\| = \left\{ \left[\frac{\partial^2}{\partial \tau^2} Q\left(\boldsymbol{\mu_o} + \tau\nu\right) \right] \Big|_{\tau=0} \right\}^{1/2} = \left\{ E\left[\nu^2\left(X_1\right)\right] \right\}^{1/2} = \left\{ \int \nu^2\left(x\right) f_{X_1}\left(x\right) dx \right\}^{1/2}$$

where $\|\boldsymbol{\mu} - \boldsymbol{\mu_o}\| \leq \text{const.} \times \|\boldsymbol{\mu} - \boldsymbol{\mu_o}\|_{2,P(X_1)}$ for any $\mu \in \mathcal{M}_o$. The inner product induced by the norm is given by

$$\langle \nu, \widetilde{\nu} \rangle = \int \nu\left(x\right) \widetilde{\nu}\left(x\right) f_{X_1}\left(x\right) dx.$$

Assume $E\left[\boldsymbol{\mu}\left(X_0\right)\right]$ is pathwise differentiable at $\boldsymbol{\mu_o} \in \mathcal{M}$ in the direction ν, and the pathwise derivative $\Gamma\left(\boldsymbol{\mu_o}\right)[\nu]$ is given by

$$\Gamma\left(\boldsymbol{\mu_o}\right)[\nu] := \frac{\partial E\left[\boldsymbol{\mu_o}\left(X_0\right) + \tau\nu\left(X_0\right)\right]}{\partial \tau}\Big|_{\tau=0} = E\left[\nu\left(X_0\right)\right] = \int \nu\left(x\right) f_{X_0}\left(x\right) dx.$$

The linear functional $\Gamma\left(\boldsymbol{\mu_o}\right)[\cdot]$ is bounded if and only if

$$\sup_{v \in \mathcal{V}, v \neq 0} \frac{\left|\Gamma\left(\boldsymbol{\mu_o}\right)[\nu]\right|^2}{\|\nu\|^2} = \sup_{v \in \mathcal{V}, v \neq 0} \frac{\left|\int \nu\left(x\right) f_{X_0}\left(x\right) dx\right|^2}{\int \nu^2\left(x\right) f_{X_1}\left(x\right) dx} = \int \frac{f_{X_0}\left(x\right)}{f_{X_1}\left(x\right)} f_{X_0}\left(x\right) dx < \infty. \text{ (By Assumption 14.2)}$$

We can then compute the Riesz representor for $E\left[\mu\left(X_0\right)\right]$, by Riesz representation theorem, the linear functional is bounded if and only if there is a Riesz representer $\nu^* \in \mathcal{V}$ such that

$$\Gamma\left(\boldsymbol{\mu_o}\right)[\nu] = \langle \nu^*, \nu \rangle$$

and

$$\|\nu^*\|^2 = \sup_{v \in \mathcal{V}, v \neq 0} \frac{\left|\Gamma\left(\boldsymbol{\mu_o}\right)[\nu]\right|^2}{\|\nu\|^2} = \int \frac{f_{X_0}\left(x\right)}{f_{X_1}\left(x\right)} f_{X_0}\left(x\right) dx,$$

where

$$\nu^*\left(x\right) = \frac{f_{X_0}\left(x\right)}{f_{X_1}\left(x\right)}.$$

Denote $\{b_i\}_{i=1}^{\infty}$ a complete basis for the infinite dimensional Hilbert space $(\mathcal{V}, \|\cdot\|)$ and let

$B_{k(n)}(\cdot) = \left(b_1(\cdot), \ldots, b_{k(n)}(\cdot)\right)'$. Then $\mathcal{V}_{k(n)} = \left\{v(\cdot) = B_{k(n)}(\cdot)'\gamma : \gamma \in \mathbb{R}^{k(n)}\right\}$ becomes dense in $(\mathcal{V}, \|\cdot\|)$ as $k(n) \to \infty$. Let $\nu_n = \Pi_{2n}\nu$ denote the projection of ν on to $\mathcal{V}_{k(n)}$, and $\boldsymbol{\nu_n} = (\nu_n, \ldots, \nu_n)'$.

For $\varepsilon_n = o_p\left(n^{-1/2}\right)$, $\widehat{\boldsymbol{\mu}}_{\boldsymbol{\nu_n}} = \widehat{\boldsymbol{\mu}} \pm \varepsilon_n \boldsymbol{\nu}_n \in \mathcal{M}_n$, $G_{n_1}\{f\} := \sum_{i=1}^{n_1} f(X_{1i}) - E[X_1]$ denotes the empirical process, we need to establish the followings:

$$E\left[\left(\widehat{\boldsymbol{\mu}}(X_1) - \boldsymbol{\mu_o}(X_1)\right)\left(\boldsymbol{\nu_n}(X_1) - \boldsymbol{\nu}(X_1)\right)\right] = O_p(\varepsilon_n) \tag{B.5}$$

$$E\left[\frac{1}{2}\left(Y_{1i} - \widehat{\boldsymbol{\mu}}(X_{1i})\right)^2 - \frac{1}{2}\left(Y_{1i} - \widehat{\boldsymbol{\mu}}_{\boldsymbol{\nu_n}}(X_{1i})\right)^2\right] = \mp\varepsilon_n E\left[\left(\widehat{\boldsymbol{\mu}}(X_1) - \boldsymbol{\mu_o}(X_1)\right)\boldsymbol{\nu}(X_1)\right] - O_p\left(\varepsilon_n^2\right) \tag{B.6}$$

$$G_{n_1}\left\{\left(\boldsymbol{\mu}_o(X_1) - \widehat{\boldsymbol{\mu}}(X_1)\right)\boldsymbol{\nu_n}(X_1)\right\} = O_p(\varepsilon_n) \tag{B.7}$$

$$G_{n_1}\left\{\boldsymbol{\nu_n}(X_1)^2\right\} = O_p(1) \tag{B.8}$$

$$G_{n_1}\left\{\frac{1}{2}\left(Y_{1i} - \widehat{\boldsymbol{\mu}}(X_{1i})\right)^2 - \frac{1}{2}\left(Y_{1i} - \widehat{\boldsymbol{\mu}}_{\boldsymbol{\nu_n}}(X_{1i})\right)^2 + \left(Y_1 - \boldsymbol{\mu}_o(X_1)\right)\left(\widehat{\boldsymbol{\mu}}(X_1) - \widehat{\boldsymbol{\mu}}_{\boldsymbol{\nu_n}}\right)(X_1)\right\} = O_p\left(\varepsilon_n^2\right) \tag{B.9}$$

$$G_{n_1}\left\{\left(Y_1 - \boldsymbol{\mu}_o(X_1)\right)\left(\boldsymbol{\nu_n}(X_1) - \boldsymbol{\nu}(X_1)\right)\right\} = O_p(\varepsilon_n) \tag{B.10}$$

After establishing (B.5)-(B.10), we have

$$\begin{aligned}
-O_p\left(\varepsilon_n^2\right) &\leq \frac{1}{n_1}\sum_{i=1}^{n_1}\frac{1}{2}\left(Y_{1i} - \widehat{\boldsymbol{\mu}}(X_{1i})\right)^2 - \frac{1}{n_1}\sum_{i=1}^{n_1}\frac{1}{2}\left(Y_{1i} - \widehat{\boldsymbol{\mu}}_{\boldsymbol{\nu_n}}(X_{1i})\right)^2 \\
&= E\left[\frac{1}{2}\left(Y_{1i} - \widehat{\boldsymbol{\mu}}(X_{1i})\right)^2 - \frac{1}{2}\left(Y_{1i} - \widehat{\boldsymbol{\mu}}_{\boldsymbol{\nu_n}}(X_{1i})\right)^2\right] \\
&\quad - G_{n_1}\left\{\left(Y_1 - \boldsymbol{\mu}_o(X_1)\right)\left(\widehat{\boldsymbol{\mu}}(X_1) - \widehat{\boldsymbol{\mu}}_{\boldsymbol{\nu_n}}\right)(X_1)\right\} \\
&\quad + G_{n_1}\left\{\frac{1}{2}\left(Y_{1i} - \widehat{\boldsymbol{\mu}}(X_{1i})\right)^2 - \frac{1}{2}\left(Y_{1i} - \widehat{\boldsymbol{\mu}}_{\boldsymbol{\nu_n}}(X_{1i})\right)^2 + \left(Y_1 - \boldsymbol{\mu}_o(X_1)\right)\left(\widehat{\boldsymbol{\mu}}(X_1) - \widehat{\boldsymbol{\mu}}_{\boldsymbol{\nu_n}}(X_1)\right)\right\} \\
&= \mp\varepsilon_n E\left[\left(\widehat{\boldsymbol{\mu}}(X_1) - \boldsymbol{\mu_o}(X_1)\right)\boldsymbol{\nu}(X_1)\right] - G_{n_1}\left\{\left(Y_1 - \boldsymbol{\mu}_o(X_1)\right)\left(\widehat{\boldsymbol{\mu}}(X_1) - \widehat{\boldsymbol{\mu}}_{\boldsymbol{\nu_n}}(X_1)\right)\right\} + O_p\left(\varepsilon_n^2\right) \\
&= \mp\varepsilon_n E\left[\left(\widehat{\boldsymbol{\mu}}(X_1) - \boldsymbol{\mu_o}(X_1)\right)\boldsymbol{\nu}(X_1)\right] \pm \varepsilon_n G_{n_1}\left\{\left(Y_1 - \boldsymbol{\mu}_o(X_1)\right)\boldsymbol{\nu_n}(X_1)\right\} + O_p\left(\varepsilon_n^2\right) \\
&= \mp\varepsilon_n E\left[\left(\widehat{\boldsymbol{\mu}}(X_1) - \boldsymbol{\mu_o}(X_1)\right)\boldsymbol{\nu}(X_1)\right] \pm \varepsilon_n G_{n_1}\left\{\left(Y_1 - \boldsymbol{\mu}_o(X_1)\right)\boldsymbol{\nu}(X_1)\right\} + O_p\left(\varepsilon_n^2\right).
\end{aligned}$$

Thus we have

$$E\left[\left(\widehat{\boldsymbol{\mu}}\left(X_1\right)-\boldsymbol{\mu_o}\left(X_1\right)\right)\boldsymbol{\nu}\left(X_1\right)\right]=G_{n_1}\left\{\left(Y_1-\boldsymbol{\mu}_o\left(X_1\right)\right)\boldsymbol{\nu}\left(X_1\right)\right\}+O_p\left(\varepsilon_n\right). \tag{B.11}$$

Now we prove (B.5)-(B.10). (B.5) can be established by Assumption 14.5 and (B.2). (B.6) can be established by

$$\begin{aligned}
&E\left[\frac{1}{2}\left(Y_{1i}-\widehat{\boldsymbol{\mu}}\left(X_{1i}\right)\right)^2-\frac{1}{2}\left(Y_{1i}-\widehat{\boldsymbol{\mu}}_{\boldsymbol{\nu_n}}\left(X_{1i}\right)\right)^2\right]\\
&=\mp\varepsilon_n E\left[\left(\widehat{\boldsymbol{\mu}}\left(X_1\right)-\boldsymbol{\mu_o}\left(X_1\right)\right)\boldsymbol{\nu_n}\left(X_1\right)\right]-\frac{1}{2}\varepsilon_n^2 E\left[\nu_n^2\left(X_1\right)\right]\\
&=\mp\varepsilon_n E\left[\left(\widehat{\boldsymbol{\mu}}\left(X_1\right)-\boldsymbol{\mu_o}\left(X_1\right)\right)\boldsymbol{\nu}\left(X_1\right)\right]\mp\varepsilon_n E\left[\left(\widehat{\boldsymbol{\mu}}\left(X_1\right)-\boldsymbol{\mu_o}\left(X_1\right)\right)\left(\boldsymbol{\nu_n}\left(X_1\right)-\boldsymbol{\nu}\left(X_1\right)\right)\right]-O_p\left(\varepsilon_n^2\right)\\
&=\mp\varepsilon_n E\left[\left(\widehat{\boldsymbol{\mu}}\left(X_1\right)-\boldsymbol{\mu_o}\left(X_1\right)\right)\boldsymbol{\nu}\left(X_1\right)\right]-O_p\left(\varepsilon_n^2\right)
\end{aligned}$$

where the last two equality are established by Assumption 14.2 and (B.5), respectively.

Similar to Chen et al. (2005), (B.7) can be established by Assumption 13.1, 14. By applying Theorem 3 in Chen and Shen (1998), we can establish (B.7). (B.8) is implied by Markov inequality. (B.10) is implied by Chebychev inequality.

(B.9) can be established by

$$\begin{aligned}
&G_{n_1}\left\{\frac{1}{2}\left(Y_1-\widehat{\boldsymbol{\mu}}\left(X_1\right)\right)^2-\frac{1}{2}\left(Y_1-\widehat{\boldsymbol{\mu}}_{\boldsymbol{\nu_n}}\left(X_1\right)\right)^2+\left(Y_1-\boldsymbol{\mu}_o\left(X_1\right)\right)\left(\widehat{\boldsymbol{\mu}}\left(X_1\right)-\widehat{\boldsymbol{\mu}}_{\boldsymbol{\nu_n}}\left(X_1\right)\right)\right\}\\
&=G_{n_1}\left\{\frac{1}{2}\left(2Y_1-\widehat{\boldsymbol{\mu}}\left(X_1\right)-\widehat{\boldsymbol{\mu}}_{\boldsymbol{\nu_n}}\left(X_1\right)\right)\left(\widehat{\boldsymbol{\mu}}_{\boldsymbol{\nu_n}}\left(X_1\right)-\widehat{\boldsymbol{\mu}}\left(X_1\right)\right)-\left(Y_1-\boldsymbol{\mu}_o\left(X_1\right)\right)\left(\widehat{\boldsymbol{\mu}}_{\boldsymbol{\nu_n}}\left(X_1\right)-\widehat{\boldsymbol{\mu}}\left(X_1\right)\right)\right\}\\
&=G_{n_1}\left\{\frac{1}{2}\left(2Y_1-\widehat{\boldsymbol{\mu}}\left(X_1\right)-\widehat{\boldsymbol{\mu}}_{\boldsymbol{\nu_n}}\left(X_1\right)\right)\left(\widehat{\boldsymbol{\mu}}_{\boldsymbol{\nu_n}}\left(X_1\right)-\widehat{\boldsymbol{\mu}}\left(X_1\right)\right)-\left(Y_1-\boldsymbol{\mu}_o\left(X_1\right)\right)\left(\widehat{\boldsymbol{\mu}}_{\boldsymbol{\nu_n}}\left(X_1\right)-\widehat{\boldsymbol{\mu}}\left(X_1\right)\right)\right\}\\
&=G_{n_1}\left\{\pm\varepsilon_n\frac{1}{2}\left(2\boldsymbol{\mu}_o\left(X_1\right)-2\widehat{\boldsymbol{\mu}}\left(X_1\right)\mp\varepsilon_n\boldsymbol{\nu_n}\left(X_1\right)\right)\boldsymbol{\nu_n}\left(X_1\right)\right\}\\
&=\pm\varepsilon_n G_{n_1}\left\{\left(\boldsymbol{\mu}_o\left(X_1\right)-\widehat{\boldsymbol{\mu}}\left(X_1\right)\right)\boldsymbol{\nu_n}\left(X_1\right)\right\}-\frac{1}{2}\varepsilon_n^2 G_{n_1}\left\{\boldsymbol{\nu_n}\left(X_1\right)^2\right\}\\
&=O_p\left(\varepsilon_n^2\right).
\end{aligned}$$

Now we have

$$
\begin{aligned}
&\sqrt{n_1}\left(\widehat{\tau}^*_{ATU,S1}-\tau^*_{ATU,S1}\right)\\
&=\sqrt{n_1}\int\left(\widehat{\boldsymbol{\mu}}-\boldsymbol{\mu_o}\right)d\left(\widehat{F}_{X_0}-F_{X_0}\right)+\sqrt{n_1}\int\left(\widehat{\boldsymbol{\mu}}-\boldsymbol{\mu_o}\right)dF_{X_0}\\
&\quad+\sqrt{n_1}\int\boldsymbol{\mu_o}d\left(\widehat{F}_{X_0}-F_{X_0}\right)-\sqrt{n_1}\left(\frac{1}{n_0}\sum_{j=1}^{n_0}Y_{0j}-E\left[Y_0\right]\right)\\
&=\frac{1}{\sqrt{n_1}}\sum_{i=1}^{n_1}\left(Y_{1i}-\boldsymbol{\mu}_o\left(X_{1i}\right)\right)\frac{f_{X_0}\left(X_{1i}\right)}{f_{X_1}\left(X_{1i}\right)}+\sqrt{\frac{n_1}{n_0}}\frac{1}{\sqrt{n_0}}\sum_{j=1}^{n_0}\left\{\boldsymbol{\mu_o}\left(X_{0j}\right)-E\left(\boldsymbol{\mu_o}\left(X_0\right)\right)\right\}\\
&\quad-\sqrt{\frac{n_1}{n_0}}\left(\frac{1}{\sqrt{n_0}}\sum_{j=1}^{n_0}Y_{0j}-E\left[Y_0\right]\right)+o_p\left(1\right)
\end{aligned}
$$

□

B.7 Proof for Theorem 6

Proof. The proof is similar to the proof of Theorem 4. To be self-contained, we provide the main steps here. By (2.8), Assumption 15.1, 15.3-13.8 , we have the following:

$$
\begin{aligned}
&\left\|\widehat{\boldsymbol{\mu}}^*-\boldsymbol{\mu}^*_{\boldsymbol{o}}\right\|_{\infty,\omega}=o_p\left(1\right), &&\left\|\widehat{\boldsymbol{\mu}^*}-\boldsymbol{\mu}^*_{\boldsymbol{o}}\right\|_{2,P(X_1)}=O_p\left(n_1^{-\gamma_1/(d_X+2\gamma_1)}\right),\\
&\left\|\widehat{\boldsymbol{h}}-\boldsymbol{h_o}\right\|_{\infty,\omega}=o_p\left(1\right), &&\left\|\widehat{\boldsymbol{h}}-\boldsymbol{h_o}\right\|_{2,P(Z_1)}=O_p\left(n_1^{-\gamma_2/(d_X+d_Y+2\gamma_2)}\right),
\end{aligned}
$$

for $k_{n_1}=n_1^{d_X/(d_X+2\gamma_1)}$ and $q_{n_1}=n_1^{(d_X+d_Y)/(d_X+d_Y+2\gamma_2)}$. Similarly, by Newey (1994) and (2.8), Assumption 15, 16, we have $\widehat{\tau}^*_{ATU,S2}-\tau^*_{ATU,S2}=o_p(1)$. □

B.8 Proof for Theorem 7

Proof. The proof is similar to proof of Theorem 5. Recall $Z_0 := (X_0, Y_0)$.

$$\begin{aligned}\widehat{\boldsymbol{\tau}}^*_{ATU,S2} - \boldsymbol{\tau}^*_{ATU,S2} = &\left(\frac{1}{n_0}\sum_{j=1}^{n_0}\widehat{\boldsymbol{\mu}^*}\left(X_{0j}\right) - E\left(\boldsymbol{\mu}^*_{\boldsymbol{o}}\left(X_{0j}\right)\right)\right)\\ &- \left(\frac{1}{n_0}\sum_{j=1}^{n_0} Y_{0j} - E\left(Y_0\right)\right)\\ &+ \left(\frac{1}{n_0}\sum_{j=1}^{n_0}\widehat{\boldsymbol{h}}\left(Z_{0j}\right) - E\left(\boldsymbol{h}_{\boldsymbol{o}}\left(Z_0\right)\right)\right).\end{aligned}$$

We complete our proof by establishing the following equation:

$$\begin{aligned}&\frac{1}{n_0}\sum_{j=1}^{n_0}\widehat{\boldsymbol{\mu}^*}\left(X_{0j}\right) - E\left(\boldsymbol{\mu}^*_{\boldsymbol{o}}\left(X_{0j}\right)\right)\\ &= \frac{1}{n_0}\sum_{j=1}^{n_0}\boldsymbol{\mu}^*_{\boldsymbol{o}}\left(X_{0j}\right) - E\left(\boldsymbol{\mu}^*_{\boldsymbol{o}}\left(X_{0j}\right)\right) + \frac{1}{n_1}\sum_{i=1}^{n_1}\nu_1^*\left(X_{1i}\right)\left(Y^*_{1i} - \boldsymbol{\mu}^*_{\boldsymbol{o}}\left(X_{1i}\right)\right) + o_p\left(n_1^{-1/2}\right). \quad \text{(B.12)}\\ &\frac{1}{n_0}\sum_{j=1}^{n_0}\widehat{\boldsymbol{h}}\left(Z_{0j}\right) - E\left(\boldsymbol{h}_{\boldsymbol{o}}\left(Z_{0j}\right)\right)\\ &= \frac{1}{n_0}\sum_{j=1}^{n_0}\boldsymbol{h}_{\boldsymbol{o}}\left(Z_{0j}\right) - E\left(\boldsymbol{h}_{\boldsymbol{o}}\left(Z_{0j}\right)\right) + \frac{1}{n_1}\sum_{i=1}^{n_1}\nu_2^*\left(Z_{1i}\right)\left(W_{1i} - \boldsymbol{h}_{\boldsymbol{o}}\left(Z_{1i}\right)\right) + o_p\left(n_1^{-1/2}\right). \quad \text{(B.13)}\end{aligned}$$

where $\nu_1^*(\cdot) = \frac{f_{X_0}(\cdot)}{f_{X_1}(\cdot)}$, $\nu_2^*(\cdot) = \frac{f_{Z_0}(\cdot)}{f_{Z_1}(\cdot)}$. These can be established by Assumption 15 and 16 through a similar step as establishing (B.3) and (B.4). Hence,

$$\begin{aligned}\widehat{\boldsymbol{\tau}}^*_{ATU,S2} - \boldsymbol{\tau}^*_{ATU,S2} = &\frac{1}{n_0}\sum_{j=1}^{n_0}\left[\boldsymbol{\mu}^*_{\boldsymbol{o}}\left(X_{0j}\right) - E\left(\boldsymbol{\mu}^*_{\boldsymbol{o}}\left(X_{0j}\right)\right) + \boldsymbol{h}_{\boldsymbol{o}}\left(Z_{0j}\right) - E\left(\boldsymbol{h}_{\boldsymbol{o}}\left(Z_0\right)\right) - Y_{0j} + E\left(Y_0\right)\right]\\ &+ \frac{1}{n_1}\sum_{i=1}^{n_1}\left[\nu_1^*\left(X_{1i}\right)\left(Y^*_{1i} - \boldsymbol{\mu}^*_{\boldsymbol{o}}\left(X_{1i}\right)\right) + \nu_2^*\left(Z_{1i}\right)\left(W_{1i} - \boldsymbol{h}_{\boldsymbol{o}}\left(Z_{1i}\right)\right)\right] + o_p\left(n_1^{-1/2}\right).\end{aligned}$$

□

B.9 Proof for Theorem 8

Proof. To prove for Theorem 8, we need to verify $\left\|\widehat{\boldsymbol{h}}\left(\cdot,\widehat{T}\left(\cdot\right)\right)-\boldsymbol{h_o}\left(\cdot,T_o\left(\cdot\right)\right)\right\|_{\infty,\omega}=o_p\left(1\right)$, this is because

$$\begin{aligned}
&\left\|\widehat{\boldsymbol{h}}\left(\cdot,\widehat{T}\left(\cdot\right)\right)-\boldsymbol{h_o}\left(\cdot,T_o\left(\cdot\right)\right)\right\|_{\infty,\omega}\\
&=\left\|\widehat{\boldsymbol{h}}\left(\cdot,\widehat{T}\left(\cdot\right)\right)-\boldsymbol{h_o}\left(\cdot,\widehat{T}\left(\cdot\right)\right)+\boldsymbol{h_o}\left(\cdot,\widehat{T}\left(\cdot\right)\right)-\boldsymbol{h_o}\left(\cdot,T_o\left(\cdot\right)\right)\right\|_{\infty,\omega}\\
&\leq\left\|\widehat{\boldsymbol{h}}\left(\cdot,\widehat{T}\left(\cdot\right)\right)-\boldsymbol{h_o}\left(\cdot,\widehat{T}\left(\cdot\right)\right)\right\|_{\infty,\omega}+\left\|\boldsymbol{h_o}\left(\cdot,\widehat{T}\left(\cdot\right)\right)-\boldsymbol{h_o}\left(\cdot,T_o\left(\cdot\right)\right)\right\|_{\infty,\omega}\\
&=o_p\left(1\right),
\end{aligned}$$

by Assumption 17. Similar to proof of Theorem 4 and 6, we have $\widehat{\tau}^*_{ATU,S3}-\tau^*_{ATU,S3}=o_p(1)$. □

B.10 Proof for Theorem 9

Proof. Similar to Theorem 7, we have

$$\frac{1}{n_0}\sum_{j=1}^{n_0}\widehat{\boldsymbol{\mu}}^*\left(X_{0j}\right)-E\left(\boldsymbol{\mu}^*_{\boldsymbol{o}}(X_0)\right)=\frac{1}{n_0}\sum_{j=1}^{n_0}\boldsymbol{\mu}^*_{\boldsymbol{o}}\left(X_{0j}\right)-E\left(\boldsymbol{\mu}^*_{\boldsymbol{o}}\left(X_0\right)\right)+\frac{1}{n_1}\sum_{i=1}^{n_1}\frac{f_{X_0}\left(X_{1i}\right)}{f_{X_1}\left(X_{1i}\right)}\left(Y^*_{1i}-\boldsymbol{\mu}^*_{\boldsymbol{o}}\left(X_{1i}\right)\right)+o_p\left(n_1^{-1/2}\right.$$

Follow a similar procedure to prove for (B.3) (B.4), we can establish

$$\frac{1}{n_0}\sum_{j=1}^{n_0}\widehat{\boldsymbol{h}}\left(X_{0j},\widehat{T}\left(Y_{0j}\right)\right)-E\left(\boldsymbol{h_o}\left(X_0,T_o\left(Y_0\right)\right)\right) \tag{B.14}$$

$$=\frac{1}{n_0}\sum_{j=1}^{n_0}\left\{\boldsymbol{h_o}\left(X_{0j}\right)-E\left(\boldsymbol{h_o}\left(X_0\right)\right)+\varphi_0\left(Y_{0j}\right)-E\left(\varphi_0\left(Y_0\right)\right)\right\} \tag{B.15}$$

$$+\frac{1}{n_1}\sum_{i=1}^{n_1}\left\{\frac{f_{X_0}(X_{1i})f_{Y_1}(Y_{1i})}{f_{Z_1}\left(Z_{1i}\right)}\left(W_{1i}-\boldsymbol{h_o}\left(Z_{1i}\right)\right)+\varphi_1\left(Y_{1i}\right)-E\left(\varphi_1\left(Y_1\right)\right)\right\}. \tag{B.16}$$

In the following, we establish

$$\int \left\{\widehat{\boldsymbol{h}}\left(X_0, \widehat{T}\left(Y_0\right)\right) - \boldsymbol{h_o}\left(X_0, T_o\left(Y_0\right)\right)\right\} d\left(\widehat{F}_{Z_0} - F_{Z_0}\right) = o_p\left(n_0^{-1/2}\right) = o_p \tag{B.17}$$

$$E\left[\widehat{\boldsymbol{h}}\left(X_0, \widehat{T}\left(Y_0\right)\right) - \boldsymbol{h_o}\left(X_0, \widehat{T}\left(Y_0\right)\right) - \widehat{\boldsymbol{h}}\left(X_0, T_o\left(Y_0\right)\right) + \boldsymbol{h_o}\left(X_0, T_o\left(Y_0\right)\right)\right] = o_p \tag{B.18}$$

$$E\left[\widehat{\boldsymbol{h}}\left(X_0, T_o\left(Y_0\right)\right) - \boldsymbol{h_o}\left(X_0, T_o\left(Y_0\right)\right)\right] = \frac{1}{n_1}\sum_{i=1}^{n_1} \frac{f_{X_0,Y_0}\left(X_{1i}, T_o^{-1}\left(Y_{1i}\right)\right)}{f_{Y_0}\left(T_o^{-1}\left(Y_{1i}\right)\right)} \frac{f_{Y_1}\left(Y_{1i}\right)}{f_{X_1,Y_1}\left(X_{1i}, Y_{1i}\right)}\left(W_{1i} - \boldsymbol{h}_o\left(Z_{1i}\right)\right) + o_p \tag{B.19}$$

(B.17) is established by empirical process.

(B.18) is established by

$$\begin{aligned}
&E\left[\widehat{\boldsymbol{h}}\left(X_0, \widehat{T}\left(Y_0\right)\right) - \boldsymbol{h_o}\left(X_0, \widehat{T}\left(Y_0\right)\right) - \widehat{\boldsymbol{h}}\left(X_0, T_o\left(Y_0\right)\right) + \boldsymbol{h_o}\left(X_0, T_o\left(Y_0\right)\right)\right] \\
&= E\left[\widehat{\boldsymbol{h}}_{\boldsymbol{2}}\left(X_0, T_o\left(Y_0\right)\right)\left(\widehat{T}\left(Y_0\right) - T_o\left(Y_0\right)\right) - \boldsymbol{h_{o,2}}\left(X_0, T_o\left(Y_0\right)\right)\left(\widehat{T}\left(Y_0\right) - T_o\left(Y_0\right)\right)\right] \\
&= E\left[\left(\widehat{\boldsymbol{h}}_{\boldsymbol{2}}\left(X_0, T_o\left(Y_0\right)\right) - \boldsymbol{h_{o,2}}\left(X_0, T_o\left(Y_0\right)\right)\right)\left(\widehat{T}\left(Y_0\right) - T_o\left(Y_0\right)\right)\right] \\
&\leq \left\|\widehat{\boldsymbol{h}}_{\boldsymbol{2}}\left(\cdot, T_o\left(\cdot\right)\right) - \boldsymbol{h_{o,2}}\left(\cdot, T_o\left(\cdot\right)\right)\right\|_{2,P(Z_0)} \left\|\widehat{T} - T_o\right\|_{2,P(Z_0)}
\end{aligned}$$

where $\widehat{\boldsymbol{h}}_{\boldsymbol{2}}\left(x, y\right) = \frac{\partial \widehat{\boldsymbol{h}}(x,y)}{\partial y}$ and $\boldsymbol{h_{o,2}}\left(x, y\right) = \frac{\partial \boldsymbol{h_o}(x,y)}{\partial y}$.

(B.19) is established by the following:

Step 1: Similar to proof of Theorem 5, since $\boldsymbol{h_o}$ is a unique minimizer which solves

$$\boldsymbol{h_o} = \underset{\boldsymbol{h}\in\mathcal{H}}{\arg\min}\, E\left[\frac{1}{2}\left(W_1 - \boldsymbol{h}\left(Z_1\right)\right)^2\right].$$

Let $\mathcal{V}_{\mathcal{H}}$ be the closed linear span of $\mathcal{H} - \{\boldsymbol{h_o}\}$ under norm $\|\cdot\|_{\mathcal{H}}$, which is defined as

$$\|\nu\|_{\mathcal{H}} = \left\{E\left[\nu^2\left(Z_1\right)\right]\right\}^{1/2} = \left\{\int \nu^2\left(z\right) f_{Z_1}\left(z\right) dz\right\}^{1/2}$$

where $\|\boldsymbol{h}-\boldsymbol{h_o}\| \leq$ const. $\times \|\boldsymbol{h}-\boldsymbol{h_o}\|_{2,P(Z_1)}$ for any $\boldsymbol{h} \in \mathcal{H}_o$. The inner product induced by the norm is given by

$$\langle \nu, \widetilde{\nu} \rangle = \int \nu(z)\, \widetilde{\nu}(z)\, f_{Z_1}(z)\, dz.$$

Step 2: Calculate the Riesz representation for $\frac{\partial}{\partial \boldsymbol{h}} E\left[\boldsymbol{h_o}\left(X_0, T_o\left(Y_0\right)\right)\right][\nu_h]$.

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{h}} E\left[\boldsymbol{h_o}\left(X_0, T_o\left(Y_0\right)\right)\right][\nu_h] &= \frac{\partial}{\partial \tau} E\left[\left(\boldsymbol{h_o}+\tau\nu_h\right)\left(X_0, T_o\left(Y_0\right)\right)\right]\Big|_{\tau=0} \\
&= E\left[\nu_h\left(X_0, T_o\left(Y_0\right)\right)\right] \\
&= \int\int \nu_h\left(x, T_o\left(y\right)\right) f_{X_0,Y_0}\left(x,y\right) dxdy \\
&= \int\int \nu_h\left(x, T_o\left(y\right)\right) \frac{f_{X_0,Y_0}\left(x,y\right)}{f_{Y_0}\left(y\right)} dx f_{Y_0}\left(y\right) dy \\
&= \int\int \nu_h\left(x, y\right) \frac{f_{X_0,Y_0}\left(x, T_o^{-1}\left(y\right)\right)}{f_{Y_0}\left(T_o^{-1}\left(y\right)\right)} dx f_{Y_1}\left(y\right) dy \\
&= \int\int \nu_h\left(x, y\right) \frac{f_{X_0,Y_0}\left(x, T_o^{-1}\left(y\right)\right)}{f_{Y_0}\left(T_o^{-1}\left(y\right)\right)} \frac{f_{Y_1}\left(y\right)}{f_{X_1,Y_1}\left(x,y\right)} f_{X_1,Y_1}\left(x,y\right) dxdy \\
&= \langle \nu_h, \nu_h^* \rangle,
\end{aligned}$$

where $\nu_h^*(x,y) := \frac{f_{X_0,Y_0}\left(x, T_o^{-1}(y)\right)}{f_{Y_0}\left(T_o^{-1}(y)\right)} \frac{f_{Y_1}(y)}{f_{X_1,Y_1}(x,y)}$ and fifth equation holds by Monge-Ampère equation. By Riesz representation theorem, $\frac{\partial}{\partial \boldsymbol{h}} E\left[\boldsymbol{h_o}\left(X_0, T_o\left(Y_0\right)\right)\right][\nu_h]$ is bounded if and only if

$$\frac{\partial}{\partial \boldsymbol{h}} E\left[\boldsymbol{h_o}\left(X_0, T_o\left(Y_0\right)\right)\right][\nu_h] = \langle \nu_h, \nu_h^* \rangle \text{ for all } \nu_h \in \mathcal{V}_{\mathcal{H}},$$

and

$$\|\nu_h^*\| = \sup_{\nu_h \in \mathcal{V}_{\mathcal{H}}, \nu_h \neq 0} \frac{\frac{\partial}{\partial \boldsymbol{h}} E\left[\boldsymbol{h_o}\left(X_0, T_o\left(Y_0\right)\right)\right][\nu_h]}{\|\nu_h\|} = E\left[\left(\frac{f_{X_0,Y_0}\left(X_1, T_o^{-1}\left(Y_1\right)\right)}{f_{Y_0}\left(T_o^{-1}\left(Y_1\right)\right)} \frac{f_{Y_1}\left(Y_1\right)}{f_{X_1,Y_1}\left(X_1, Y_1\right)}\right)^2\right] < \infty.$$

Then following a simular procedure to establish (B.11), we have

$$E\left[\widehat{\boldsymbol{h}}\left(X_0, T_o\left(Y_0\right)\right) - \boldsymbol{h_o}\left(X_0, T_o\left(Y_0\right)\right)\right] = G_{n_1}\left\{\frac{f_{X_0,Y_0}\left(X_{1i}, T_o^{-1}\left(Y_1\right)\right)}{f_{Y_0}\left(T_o^{-1}\left(Y_1\right)\right)}\frac{f_{Y_1}\left(Y_1\right)}{f_{X_1,Y_1}\left(X_1, Y_1\right)}\left(W_1 - \boldsymbol{h}_o\left(Z_1\right)\right) + O_p\left(\varepsilon_n\right)\right.$$

and thus (B.19) holds. Lastly we can complete our proof by

$$\begin{aligned}
&\frac{1}{n_0}\sum_{j=1}^{n_0}\widehat{\boldsymbol{h}}\left(X_{0j}, \widehat{T}\left(Y_{0j}\right)\right) - E\left(\boldsymbol{h_o}\left(X_0, T_o\left(Y_0\right)\right)\right) \\
&= E\left[\widehat{\boldsymbol{h}}\left(X_0, \widehat{T}\left(Y_0\right)\right)\right] - E\left(\boldsymbol{h_o}\left(X_0, T_o\left(Y_0\right)\right)\right) + \frac{1}{n_0}\sum_{j=1}^{n_0}\boldsymbol{h_o}\left(X_{0j}, T_o\left(Y_{0j}\right)\right) \\
&\quad + \frac{1}{n_0}\sum_{j=1}^{n_0}\left[\widehat{\boldsymbol{h}}\left(X_{0j}, \widehat{T}\left(Y_{0j}\right)\right) - \boldsymbol{h_o}\left(X_{0j}, T_o\left(Y_{0j}\right)\right)\right] - E\left[\widehat{\boldsymbol{h}}\left(X_0, \widehat{T}\left(Y_0\right)\right) - \boldsymbol{h_o}\left(X_0, T_o\left(Y_0\right)\right)\right] \\
&= E\left[\widehat{\boldsymbol{h}}\left(X_0, \widehat{T}\left(Y_0\right)\right)\right] - E\left(\boldsymbol{h_o}\left(X_0, T_o\left(Y_0\right)\right)\right) + \frac{1}{n_0}\sum_{j=1}^{n_0}\boldsymbol{h_o}\left(X_{0j}, T_o\left(Y_{0j}\right)\right) + o_p\left(n_1^{-1/2}\right) \\
&= E\left[\widehat{\boldsymbol{h}}\left(X_0, T_o\left(Y_0\right)\right) - \boldsymbol{h_o}\left(X_0, T_o\left(Y_0\right)\right)\right] + E\left[\boldsymbol{h_o}\left(X_0, \widehat{T}\left(Y_0\right)\right) - \boldsymbol{h_o}\left(X_0, T_o\left(Y_0\right)\right)\right] \\
&\quad + E\left[\widehat{\boldsymbol{h}}\left(X_0, \widehat{T}\left(Y_0\right)\right) - \boldsymbol{h_o}\left(X_0, \widehat{T}\left(Y_0\right)\right) - \widehat{\boldsymbol{h}}\left(X_0, T_o\left(Y_0\right)\right) + \boldsymbol{h_o}\left(X_0, T_o\left(Y_0\right)\right)\right] \\
&\quad + \frac{1}{n_0}\sum_{j=1}^{n_0}\boldsymbol{h_o}\left(X_{0j}, T_o\left(Y_{0j}\right)\right) + o_p\left(n_1^{-1/2}\right) \\
&= E\left[\widehat{\boldsymbol{h}}\left(X_0, T_o\left(Y_0\right)\right) - \boldsymbol{h_o}\left(X_0, T_o\left(Y_0\right)\right)\right] + E\left[\boldsymbol{h_o}\left(X_0, \widehat{T}\left(Y_0\right)\right) - \boldsymbol{h_o}\left(X_0, T_o\left(Y_0\right)\right)\right] \\
&\quad + \frac{1}{n_0}\sum_{j=1}^{n_0}\boldsymbol{h_o}\left(X_{0j}, T_o\left(Y_{0j}\right)\right) + o_p\left(n_1^{-1/2}\right)
\end{aligned}$$

second equality holds by (B.17), fourth equality holds by (B.18).

□

B.11 Robustness Result For Empirical Example

	Total	Stealing	Marijuana	Gambling	Homeless
ATU	-0.367***	-0.093***	-0.076***	-0.087***	-0.111***
	(0.051)	(0.020)	(0.024)	(0.021)	(0.021)
Corrected ATU	-0.499***	-0.143***	-0.097**	-0.133*	-0.200***
	(0.125)	(0.040)	(0.038)	(0.060)	(0.054)

Table B.1: Therapy and Cash: Robustness for Scenario 2

	Total	Stealing	Marijuana	Gambling	Homeless
ATU	-0.175**	-0.046*	-0.027	-0.089***	-0.013
	(0.056)	(0.021)	(0.022)	(0.020)	(0.021)
Corrected ATU	-0.127	-0.023	-0.030	-0.066	0.007
	(0.122)	(0.051)	(0.028)	(0.047)	(0.051)

Table B.2: Therapy Only: Robustness for Scenario 2

	Total	Stealing	Marijuana	Gambling	Homeless
ATU	-0.082	-0.034	0.000	0.022	-0.070***
	(0.057)	(0.023)	(0.023)	(0.023)	(0.024)
Corrected ATU	0.004	-0.017	-0.003	0.080	-0.067*
	(0.093)	(0.048)	(0.038)	(0.049)	(0.027)

Table B.3: Cash Only: Robustness for Scenario 2

	Total	Stealing	Marijuana	Gambling	Homeless
ATU	-0.499	-0.154	-0.415	0.077	-0.007
	(17.212)	(3.201)	(3.980)	(3.678)	(1.545)
Corrected ATU	-0.631	-0.204	-0.436	0.030	-0.096
	(5.666)	(2.120)	(4.402)	(2.746)	(5.147)

Table B.4: Therapy and Cash: Robustness for Scenario 2 (quadratic spline)

	Total	Stealing	Marijuana	Gambling	Homeless
ATU	-0.186	0.131	-0.065	-0.253	0.000
	(3.347)	(2.351)	(4.031)	(4.590)	(1.856)
Corrected ATU	-0.139	0.154	-0.068	-0.230	0.021
	(0.707)	(0.261)	(0.289)	(0.283)	(0.285)

Table B.5: Therapy: Robustness for Scenario 2 (quadratic spline)

	Total	Stealing	Marijuana	Gambling	Homeless
ATU	-4.347	-1.783	-0.307	-1.520	-0.737
	(3.381)	(9.418)	(5.207)	(3.907)	(2.167)
Corrected ATU	-4.261	-1.766	-0.310	-1.462	-0.734
	(30.673)	(4.884)	(7.124)	(3.343)	(5.678)

Table B.6: Cash: Robustness for Scenario 2 (quadratic spline)

Appendix C

APPENDICES FOR COVID-19, URBAN TRANSPORTATION AND AIR POLLUTION

C.1 List of 36 Cities

Beijing, Shanghai, Guangzhou, Shenzhen, Chengdu, Hangzhou, Nanjing, Ningbo, Xiamen, Zhengzhou, Chongqing, Shenyang, Wuhan, Xi'an, Changsha, Qingdao, Hefei, Dalian, Fuzhou, Guiyang, Harbin, Haikou, Hohhot, Jinan, Kunming, Lhasa, Lanzhou, Nanchang, Nanning, Shijiazhuang, Taiyuan, Tianjin, Urumqi, Xining, Yinchuan, Changchun.

C.2 List of 26 COVID-19-related Keywords

Most of the keywords are in Chinese, except three as labelled. COVID-19-related terms (5): Novel coronavirus, COVID-19 pandemic, COVID-19, COVID (in English), Xin Guan (abbreviated name of COVID-19 in Chinese). Specialized remedy (6): Remdesivir, health code, shelter hospital, join rescue operations immediately, COVID-19 vaccine, anti-pandemic. Generic remedy (10): ECMO (in English), quarantine, nucleic acid testing, nasopharyngeal swab testing, testing kit, face mask, sanitizer, ventilator, artificial lung, Lianhua Qingwen. Generic terms for COVID pandemic and terms for other pandemics (5): Pandemic, pneumonia, SARS, SARS (in English), Ebola.

C.3 Results of Diagnostic Tests for Instrument Variable Models

See results in Table C.1.

C.4 Results of Moderating Effects of New Energy Bus Penetration Rate

See results in Table C.2.

Table C.1: Results of Diagnostic Tests for Instrument Variable Approaches

	(1)	(2)	(3)	(4)	(5)	(6)
	2SLS model for overall analysis		2SLS model for subsector analysis			
	ln $PublicVol_{i,t}$	ln $CongIndex_{i,t}$	ln $BusVol_{i,t}$	ln $RailVol_{i,t}$	ln $TaxiVol_{i,t}$	ln $CongSpeed_{i,t}$
ln $Con_{i,t-1}$	-0.313***	-0.016***	-0.274***	-0.505***	-0.297***	0.021***
	(0.030)	(0.003)	(0.039)	(0.051)	(0.031)	(0.004)
ln $Sus_{i,t-1}$	-0.024	0.003	-0.019	-0.029	-0.035	0.003
	(0.028)	(0.003)	(0.037)	(0.049)	(0.029)	(0.004)
ln $COVIDTerm_{i,t-1}$	0.213***	0.013***	0.162***	0.436***	0.196***	0
	(0.032)	(0.003)	(0.042)	(0.056)	(0.034)	(0.004)
ln $GenericRemedy_{i,t-1}$	0.130*	0.016**	0.134	0.142	0.098	0.001
	(0.057)	(0.006)	(0.076)	(0.099)	(0.060)	(0.008)
ln $GenericTerm_{i,t-1}$	-0.528***	-0.044***	-0.491***	-0.539***	-0.484***	0.040***
	(0.068)	(0.007)	(0.089)	(0.117)	(0.071)	(0.009)
$AirTem_{i,t}$	0	-0.000***	-0.001	0	0	0.000***
	(0.001)	0.000	(0.001)	(0.001)	(0.001)	0.000
$DewPoint_{i,t}$	-0.001	0	-0.001	-0.001	-0.001	0
	(0.001)	0.000	(0.001)	(0.001)	(0.001)	0.000
$Precip_{i,t}$	0.055	0.005	0.078	0.123	0.09	-0.008
	(0.103)	(0.010)	(0.136)	(0.178)	(0.108)	(0.014)
$WindSpeed_{i,t}$	0	0.078***	-0.38	1.423***	0.035	-0.066*
	(0.224)	(0.023)	(0.296)	(0.387)	(0.235)	(0.031)
$Production_{i,t}$	-0.085	0.046	-0.256	-0.148	-0.006	-0.096*
	(0.293)	(0.030)	(0.387)	(0.506)	(0.306)	(0.040)
$Constant$	-0.773	0.073	-0.578	-2.481**	-0.878	-0.002
	(0.467)	(0.047)	(0.617)	(0.807)	(0.489)	(0.064)
Month Effects	YES	YES	YES	YES	YES	YES
City Effects	YES	YES	YES	YES	YES	YES
Diagnostic tests						
F-statistic	113.390	54.500	66.670	26.630	95.620	132.330
Sargan statistic	0.792, p= 0.852		0.021, p = 0.886			
N	576	576	576	576	576	576

Note: Standard errors in parentheses. + $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table C.2: Results for Moderating Effects of New Energy Bus Penetration Rate (DV: ln $Synindex_{i,t}$)

	(1) Fixed-effects RR	(2) Two-stage RR	(3) DML RR
ln $BusVol_{i,t}$	-0.003	0.025***	0.014***
	(0.034)	(0.006)	(0.003)
ln $RailVol_{i,t}$	-0.003	0.003	0.005+
	(0.010)	(0.011)	(0.002)
ln $TaxiVol_{i,t}$	0.050*	0.035***	0.017***
	(0.022)	(0.006)	(0.004)
ln $CongIndex_{i,t}$	0.060	0.194+	0.182***
	(0.178)	(0.107)	(0.041)
ln $BusVol_{i,t}$ * $NewEnergyBus_i$	0.014	-0.014	-0.006+
	(0.057)	(0.019)	(0.004)
$AirTem_{i,t}$	-0.000	0.001**	YES
	(0.000)	(0.000)	
$DewPoint_{i,t}$	-0.001**	-0.002***	YES
	(0.000)	(0.000)	
$Precip_{i,t}$	0.005	0.003	YES
	(0.041)	(0.035)	
$WindSpeed_{i,t}$	-0.005	-0.059	YES
	(0.092)	(0.074)	
$Production_{i,t}$	0.075	0.070	YES
	(0.114)	(0.098)	
$Constant$	1.105***	0.000	YES
	(0.261)	(0.007)	
Month Effects	YES	YES	YES
City Effects	YES	YES	YES
IV		YES	YES
DML			YES
K	0.053	0.104	0.836#
F-statistic	16.080	8.366	4.843
N	576	576	576

Note: DV stands for dependent variable. Standard errors in parentheses.
+ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

C.5 Multicollinearity Detection

The correlation coefficients in Table C.3 indicate that $BusVol_{i,t}$, $RailVol_{i,t}$, $TaxiVol_{i,t}$, and $CongIndex_{i,t}$ are highly correlated (the largest correlation coefficient is $r = 0.825$ between $BusVol_{i,t}$ and $RailVol_{i,t}$). The Farrar-Glauber test ($F - G$ test) is leveraged to detect the multicollineairty (Farrar and Glauber, 1967). First, the value of the Chi-square test statistic for the models containing the two sets of independent variables of interest (the first set contains ln $PublicVol_{i,t}$ and ln $CongIndex_{i,t}$, and the second contains ln $BusVol_{i,t}$, ln $RailVol_{i,t}$, ln $TaxiVol_{i,t}$, and ln $CongIndex_{i,t}$) are 771.055 and 1485.218, respectively, evidencing the presence of multicollinearity in the model specifications. The Farrar-Glauber F-test is further conducted and the results are reported in Table C.4. For the model of total public transportation passenger volume, the F-statistic for the variable "ln $PublicVol_{i,t}$" is high (86.412) followed by the variable "ln $CongIndex_{i,t}$" (F-value of 90.588), "$AirTem_{i,t}$" (F-value of 88.925), and "$DewPoint_{i,t}$" (F-value of 88.925). For the model of the transportation subsectors, the F-statistic for the variable "ln $TaxiVol_{i,t}$" is high (166.362) followed by the variable "ln $BusVol_{i,t}$" (F-value of 116.922), "$AirTem_{i,t}$" (F-value of 66.365), "ln $CongIndex_{i,t}$" (F-value of 65.759), and "$DewPoint_{i,t}$" (F-value of 60.244). The scatter plot in Figure C.1 for cross pairs among four transportation modes further demonstrates large correlation coefficients and high significance levels among ln $BusVol_{i,t}$, ln $TaxiVol_{i,t}$, and ln $CongIndex_{i,t}$, revealing the multicollinearity among independent variables of interest.

Table C.3: Correlations

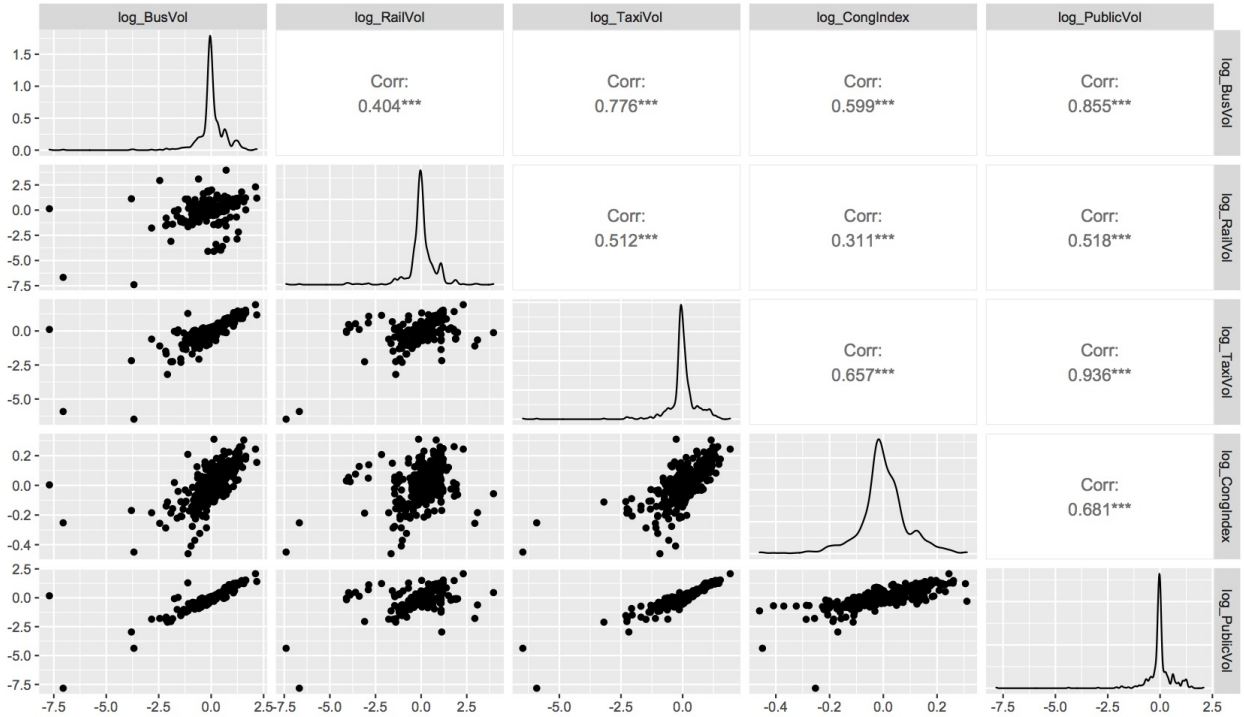
		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(1
(1)	$Synindex_{i,t}$															
(2)	$PublicVol_{i,t}$	0.034														
(3)	$BusVol_{i,t}$	0.074	0.947*													
(4)	$RailVol_{i,t}$	-0.017	0.955*	0.825*												
(5)	$TaxiVol_{i,t}$	0.081	0.695*	0.744*	0.506*											
(6)	$CongIndex_{i,t}$	0.082*	0.506*	0.590*	0.362*	0.583*										
(7)	$Con_{i,t}$	-0.030	-0.065	-0.086*	-0.037	-0.084*	-0.140*									
(8)	$Sus_{i,t}$	-0.019	0.056	0.022	0.085*	-0.010	0.094*	0.006								
(9)	$COVIDTerm_{i,t}$	-0.073	-0.153*	-0.232*	-0.054	-0.253*	-0.112*	0.205*	0.457*							
(10)	$GenericTerm_{i,t}$	-0.010	-0.070	-0.139*	0.016	-0.202*	-0.214*	0.131*	0.350*	0.824*						
(11)	$GenericRemedy_{i,t}$	0.011	0.071	-0.016	0.163*	-0.133*	-0.148*	0.223*	0.301*	0.843*	0.937*					
(12)	$AirTem_{i,t}$	-0.468*	0.214*	0.195*	0.216*	0.122*	0.157*	-0.002	0.004	-0.069	-0.166*	-0.114*				
(13)	$DewPoint_{i,t}$	-0.489*	0.230*	0.213*	0.223*	0.170*	0.155*	0.011	-0.032	-0.084*	-0.131*	-0.067	0.929*			
(14)	$Precip_{i,t}$	-0.018	-0.033	-0.006	-0.050	-0.001	0.076	-0.067	-0.041	-0.064	-0.094*	-0.101*	-0.051	-0.049		
(15)	$WindSpeed_{i,t}$	-0.308*	0.097*	0.052	0.126*	0.060	0.096*	0.011	0.065	0.081	0.054	0.080	0.115*	0.127*	-0.042	
(16)	$Production_{i,t}$	-0.016	0.806*	0.773*	0.796*	0.402*	0.395*	0.019	0.211*	0.151*	0.224*	0.346*	0.205*	0.206*	-0.116*	0.

Note: * $p < 0.05$.

Table C.4: Results of the Farrar-Glauber F-Test for Multicollinearity (DV: ln $Synindex_{i,t}$)

	(1)		(2)	
Variable	F-test	p-value	F-test	p-value
ln $PublicVol_{i,t}$	86.412	0.000		
ln $CongIndex_{i,t}$	90.588	0.000	65.759	0.000
ln $BusVol_{i,t}$			116.922	0.000
ln $RailVol_{i,t}$			28.680	0.000
ln $TaxiVol_{i,t}$			166.362	0.000
$AirTem_{i,t}$	88.925	0.000	66.365	0.000
$DewPoint_{i,t}$	80.535	0.000	60.244	0.000
$Precip_{i,t}$	0.399	0.985	0.354	0.997
$WindSpeed_{i,t}$	9.988	0.002	10.262	0.000
$Production_{i,t}$	3.783	0.033	3.017	0.035

Figure C.1: Multicollinearity Scatter Plot



Note: The data were log-transformed and demeaned.

C.6 Results of Robustness Check

See results in Table C.5, C.6,C.7, C.8, C.9, C.10, C.11, C.12, C.13.

Table C.5: Results of Robustness Check of Using $CongSpeed_{i,t}$ as Measurement of Private Transportation for Overall and Subsector Effects (DV: ln $Synindex_{i,t}$)

	(1) Fixed-effects RR	(2) Two-stage RR	(3) DML RR	(4) Fixed-effects RR	(5) Two-stage RR	(6) DML RR
ln $PublicVol_{i,t}$	0.040*	0.043**	0.031***			
	(0.016)	(0.014)	(0.008)			
ln $BusVol_{i,t}$				-0.002	0.021***	0.017***
				(0.016)	(0.005)	(0.004)
ln $RailVol_{i,t}$				-0.004	0.005	0.005
				(0.010)	(0.011)	(0.006)
ln $TaxiVol_{i,t}$				0.048*	0.031***	0.021***
				(0.021)	(0.006)	(0.005)
ln $CongSpeed_{i,t}$	-0.115	-0.220*	-0.248***	-0.107	-0.159*	-0.191***
	(0.114)	(0.092)	(0.059)	(0.114)	(0.074)	(0.048)
$AirTem_{i,t}$	0.000	0.001**	YES	0.000	0.001**	YES
	(0.000)	(0.000)		(0.000)	(0.000)	
$DewPoint_{i,t}$	-0.001**	-0.002***	YES	-0.001**	-0.002***	YES
	(0.000)	(0.000)		(0.000)	(0.000)	
$Precip_{i,t}$	0.006	0.005	YES	0.005	0.003	YES
	(0.041)	(0.036)		(0.041)	(0.035)	
$WindSpeed_{i,t}$	-0.016	-0.066	YES	-0.012	-0.057	YES
	(0.088)	(0.076)		(0.090)	(0.075)	
$Production_{i,t}$	0.075	0.061	YES	0.073	0.066	YES
	(0.114)	(0.100)		(0.114)	(0.099)	
$Constant$	1.514**	0.000	YES	1.541**	0.000	YES
	(0.495)	(0.007)		(0.492)	(0.007)	
Month Effects	YES	YES	YES	YES	YES	YES
City Effects	YES	YES	YES	YES	YES	YES
IV		YES	YES		YES	YES
DML			YES			YES
K	0.043	0.077	0.207	0.046	0.095	0.271
F-statistic	19.223	12.163	12.104	17.261	9.379	6.004
N	576	576	576	576	576	576

Note: Standard errors in parentheses. + $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table C.6: Results of Robustness Check Using $CongSpeed_{i,t}$ as Measurement of Private Transportation for Effects on Primary Air Pollutants (N = 576)

	(1) ln $CO_{i,t}$	(2) ln $NO2_{i,t}$	(3) ln $O3_{i,t}$	(4) ln $PM25_{i,t}$	(5) ln $PM10_{i,t}$	(6) ln $SO2_{i,t}$
ln $PublicVol_{i,t}$	0.034*	0.063**	-0.070**	0.043*	0.056***	0.011+
	(0.015)	(0.020)	(0.026)	(0.020)	(0.013)	(0.007)
ln $CongSpeed_{i,t}$	-0.372***	-0.394**	-0.143	-0.355**	-0.363***	-0.104*
	(0.101)	(0.125)	(0.165)	(0.135)	(0.087)	(0.052)
Month Effects	YES	YES	YES	YES	YES	YES
City Effects	YES	YES	YES	YES	YES	YES
IV	YES	YES	YES	YES	YES	YES
DML	YES	YES	YES	YES	YES	YES
K	0.097#	0.044	0.047#	0.106#	0.129	0.645#
F-statistic	11.425	28.397	4.319	6.741	18.665	2.042

Note: Standard errors in parentheses. + $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. The K with a "#" is obtained following Hoerl and Kennard (1970a,b), because the procedure following Hoerl and Kennard (1976) fails to converge.

Table C.7: Results of Robustness Check of Instrumenting Production for Overall and Subsector Effects (DV: ln $Synindex_{i,t}$)

	(1) Two-stage RR	(2) DML RR	(3) Two-stage RR	(4) DML RR
ln $PublicVol_{i,t}$	0.051***	0.034***		
	(0.011)	(0.007)		
ln $BusVol_{i,t}$			0.025***	0.015***
			(0.007)	(0.003)
ln $RailVol_{i,t}$			0.002	0.006*
			(0.011)	(0.003)
ln $TaxiVol_{i,t}$			0.035***	0.019***
			(0.007)	(0.004)
ln $CongIndex_{i,t}$	0.313*	0.287***	0.195+	0.177***
	(0.125)	(0.069)	(0.111)	(0.042)
$Production_{i,t}$	0.100	0.081*	0.095	0.048*
	(0.139)	(0.039)	(0.085)	(0.019)
$AirTem_{i,t}$	0.001**	YES	0.001**	YES
	(0.000)		(0.000)	
$DewPoint_{i,t}$	-0.002***	YES	-0.002***	YES
	(0.000)		(0.000)	
$Precip_{i,t}$	0.005	YES	0.003	YES
	(0.036)		(0.035)	
$WindSpeed_{i,t}$	-0.077	YES	-0.057	YES
	(0.076)		(0.075)	
$Constant$	0.000	YES	0.000	YES
	(0.007)		(0.007)	
Month Effects	YES	YES	YES	YES
City Effects	YES	YES	YES	YES
IV	YES	YES	YES	YES
DML		YES		YES
K	0.077	0.341	0.097	0.700
F-statistic	12.016	8.685	9.286	5.194
N	576	576	576	576

Note: Standard errors in parentheses. + $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table C.8: Results of Robustness Check of Instrumenting Production for Effects on Primary Air Pollutants (N = 576)

	(1) ln $CO_{i,t}$	(2) ln $NO2_{i,t}$	(3) ln $O3_{i,t}$	(4) ln $PM25_{i,t}$	(5) ln $PM10_{i,t}$	(6) ln $SO2_{i,t}$
ln $PublicVol_{i,t}$	0.028***	0.073***	-0.017+	0.048***	0.054***	0.021+
	(0.007)	(0.011)	(0.009)	(0.013)	(0.009)	(0.011)
ln $CongIndex_{i,t}$	0.238***	0.605***	-0.128	0.317*	0.498***	0.228+
	(0.064)	(0.154)	(0.094)	(0.141)	(0.098)	(0.118)
$Production_{i,t}$	0.049+	0.124	-0.077	0.071	0.124*	0.037
	(0.029)	(0.145)	(0.055)	(0.093)	(0.063)	(0.081)
Month Effects	YES	YES	YES	YES	YES	YES
City Effects	YES	YES	YES	YES	YES	YES
IV	YES	YES	YES	YES	YES	YES
DML	YES	YES	YES	YES	YES	YES
K	0.641#	0.067	0.307#	0.226#	0.240	0.207#
F-statistic	7.167	22.264	1.236	4.921	37.827	1.576

Note: Standard errors in parentheses. + $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. The K with a "#" is obtained following Hoerl and Kennard (1970a,b), because the procedure following Hoerl and Kennard (1976) fails to converge.

Table C.9: Results of Robustness Check of Adding Dummy of Lunar New Year for Overall and Subsector Effects (DV: ln $Synindex_{i,t}$)

	(1) Fixed-effects RR	(2) Two-stage RR	(3) DML RR	(4) Fixed-effects RR	(5) Two-stage RR	(6) DML RR
ln $PublicVol_{i,t}$	0.047**	0.052***	0.034***			
	(0.016)	(0.010)	(0.007)			
ln $BusVol_{i,t}$				0.004	0.025***	0.029***
				(0.016)	(0.006)	(0.007)
ln $RailVol_{i,t}$				-0.003	0.004	-0.004
				(0.010)	(0.011)	(0.012)
ln $TaxiVol_{i,t}$				0.050*	0.035***	0.034***
				(0.021)	(0.006)	(0.007)
ln $CongIndex_{i,t}$	0.095	0.319**	0.301***	0.071	0.200*	0.283*
	(0.179)	(0.114)	(0.071)	(0.178)	(0.100)	(0.137)
$AirTem_{i,t}$	-0.000	0.001**	YES	-0.000	0.001**	YES
	(0.000)	(0.000)		(0.000)	(0.000)	
$DewPoint_{i,t}$	-0.001**	-0.002***	YES	-0.001**	-0.002***	YES
	(0.000)	(0.000)		(0.000)	(0.000)	
$Precip_{i,t}$	0.007	0.005	YES	0.006	0.003	YES
	(0.041)	(0.035)		(0.041)	(0.035)	
$WindSpeed_{i,t}$	-0.008	-0.069	YES	-0.004	-0.053	YES
	(0.090)	(0.075)		(0.091)	(0.073)	
$Production_{i,t}$	0.070	0.063	YES	0.070	0.069	YES
	(0.115)	(0.099)		(0.115)	(0.097)	
$LunarNewYear_t$	-0.020	-0.018		-0.019	-0.021	
	(0.034)	(0.021)		(0.034)	(0.021)	
$Constant$	0.969***	0.000	YES	1.036***	0.000	YES
	(0.148)	(0.007)		(0.146)	(0.007)	
Month Effects	YES	YES	YES	YES	YES	YES
City Effects	YES	YES	YES	YES	YES	YES
IV		YES	YES		YES	YES
DML			YES			YES
K	0.059	0.093	0.396	0.059	0.117	0.105#
F-statistic	17.479	10.530	11.285	15.936	8.405	6.275
N	576	576	576	576	576	576

Note: Standard errors in parentheses. + $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table C.10: Results of Robustness Check of Dropping Wuhan for Overall and Subsector Effects (DV: ln $Synindex_{i,t}$)

	(1) Fixed-effects RR	(2) Two-stage RR	(3) DML RR	(4) Fixed-effects RR	(5) Two-stage RR	(6) DML RR
ln $PublicVol_{i,t}$	0.056**	0.067***	0.045***			
	(0.021)	(0.013)	(0.010)			
ln $BusVol_{i,t}$				0.000	0.025**	0.020***
				(0.017)	(0.007)	(0.005)
ln $RailVol_{i,t}$				-0.003	0.011	0.005
				(0.011)	(0.013)	(0.005)
ln $TaxiVol_{i,t}$				0.061*	0.045***	0.028***
				(0.024)	(0.010)	(0.007)
ln $CongIndex_{i,t}$	0.023	0.204+	0.299***	0.022	0.118	0.216**
	(0.199)	(0.113)	(0.081)	(0.190)	(0.097)	(0.062)
$AirTem_{i,t}$	-0.000	0.001**	YES	-0.000	0.001**	YES
	(0.000)	(0.000)		(0.000)	(0.000)	
$DewPoint_{i,t}$	-0.001**	-0.002***	YES	-0.001**	-0.002***	YES
	(0.000)	(0.000)		(0.000)	(0.000)	
$Precip_{i,t}$	0.007	0.007	YES	0.005	0.004	YES
	(0.041)	(0.036)		(0.041)	(0.036)	
$WindSpeed_{i,t}$	-0.007	-0.062	YES	-0.005	-0.057	YES
	(0.091)	(0.076)		(0.092)	(0.076)	
$Production_{i,t}$	0.074	0.073	YES	0.073	0.081	YES
	(0.120)	(0.104)		(0.120)	(0.102)	
$Constant$	0.943***	0.000	YES	1.021***	0.000	YES
	(0.159)	(0.007)		(0.152)	(0.007)	
Month Effects	YES	YES	YES	YES	YES	YES
City Effects	YES	YES	YES	YES	YES	YES
IV		YES	YES		YES	YES
DML			YES			YES
K	0.049	0.080	0.243	0.053	0.105	0.356
F-statistic	17.831	10.836	9.641	16.048	8.313	4.809
N	560	560	560	560	560	560

Note: Standard errors in parentheses. + $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table C.11: Results of Robustness Check of Dropping 9 Cities for Overall and Subsector Effects (DV: ln $Synindex_{i,t}$)

	(1) Fixed-effects RR	(2) Two-stage RR	(3) DML RR	(4) Fixed-effects RR	(5) Two-stage RR	(6) DML RR
ln $PublicVol_{i,t}$	0.048+	0.052***	0.032**			
	(0.026)	(0.013)	(0.010)			
ln $BusVol_{i,t}$				-0.004	0.019**	0.016**
				(0.019)	(0.006)	(0.005)
ln $RailVol_{i,t}$				-0.003	0.005	0.002
				(0.012)	(0.008)	(0.006)
ln $TaxiVol_{i,t}$				0.055*	0.039***	0.029**
				(0.028)	(0.010)	(0.009)
ln $CongIndex_{i,t}$	-0.003	0.158	0.230**	0.016	0.103	0.204**
	(0.238)	(0.102)	(0.076)	(0.225)	(0.086)	(0.078)
$AirTem_{i,t}$	-0.000	0.000*	YES	-0.000	0.000+	YES
	(0.000)	(0.000)		(0.000)	(0.000)	
$DewPoint_{i,t}$	-0.001*	-0.001***	YES	-0.001+	-0.001***	YES
	(0.000)	(0.000)		(0.000)	(0.000)	
$Precip_{i,t}$	0.008	0.004	YES	0.006	0.001	YES
	(0.050)	(0.039)		(0.050)	(0.038)	
$WindSpeed_{i,t}$	-0.005	-0.045	YES	-0.004	-0.040	YES
	(0.102)	(0.076)		(0.104)	(0.073)	
$Production_{i,t}$	0.016	0.004	YES	0.013	0.008	YES
	(0.137)	(0.108)		(0.137)	(0.105)	
$Constant$	1.043***	0.000	YES	1.102***	0.000	YES
	(0.188)	(0.009)		(0.181)	(0.009)	
Month Effects	YES	YES	YES	YES	YES	YES
City Effects	YES	YES	YES	YES	YES	YES
IV		YES	YES		YES	YES
DML			YES			YES
K	0.065	0.189	0.456	0.098	0.230	0.335
F-statistic	11.661	6.147	5.366	10.134	4.823	2.811
N	432	432	432	432	432	432

Note: Standard errors in parentheses. + $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table C.12: Results of Robustness Check of sampling with 11 Cities and 10 City Clusters for Overall and Subsector Effects (DV: ln $Synindex_{i,t}$)

	(1) Fixed-effects RR	(2) Two-stage RR	(3) DML RR	(4) Fixed-effects RR	(5) Two-stage RR	(6) DML RR
ln $PublicVol_{i,t}$	0.037	0.045***	0.041*			
	(0.027)	(0.012)	(0.019)			
ln $BusVol_{i,t}$				-0.002	0.015*	0.015+
				(0.019)	(0.006)	(0.008)
ln $RailVol_{i,t}$				-0.004	0.014	-0.001
				(0.016)	(0.012)	(0.016)
ln $TaxiVol_{i,t}$				0.043	0.037***	0.034***
				(0.033)	(0.009)	(0.009)
ln $CongSpeed_{i,t}$	0.060	0.196+	0.396+	0.065	0.116	0.250+
	(0.291)	(0.106)	(0.233)	(0.288)	(0.101)	(0.133)
$AirTem_{i,t}$	0.000	0.001**	YES	-0.000	0.001**	YES
	(0.000)	(0.000)		(0.000)	(0.000)	
$DewPoint_{i,t}$	-0.001	-0.001***	YES	-0.001	-0.001***	YES
	(0.000)	(0.000)		(0.000)	(0.000)	
$Precip_{i,t}$	0.030	0.019	YES	0.028	0.015	YES
	(0.072)	(0.053)		(0.072)	(0.055)	
$WindSpeed_{i,t}$	0.173	0.121	YES	0.161	0.131	YES
	(0.153)	(0.104)		(0.156)	(0.105)	
$Production_{i,t}$	0.117	0.170	YES	0.090	0.157	YES
	(0.234)	(0.172)		(0.236)	(0.176)	
$Constant$	0.991***	0.000	YES	1.044***	0.000	YES
	(0.206)	(0.010)		(0.207)	(0.010)	
Month Effects	YES	YES	YES	YES	YES	YES
City Effects	YES	YES	YES	YES	YES	YES
IV		YES	YES		YES	YES
DML			YES			YES
K	0.098	0.252	0.053	0.143	0.239	0.158#
F-statistic	6.713	5.833	5.272	5.657	4.636	3.534
N	336	336	336	336	336	336

Note: Standard errors in parentheses. + $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.
The K with a "#" is obtained following Hoerl and Kennard (1970a,b), because the procedure following Hoerl and Kennard (1976) fails to converge.

Table C.13: Results of Robustness Check of sampling with 11 Cities and 10 City Clusters for Effects on Primary Air Pollutants (N = 336)

	(1) ln $CO_{i,t}$	(2) ln $NO2_{i,t}$	(3) ln $O3_{i,t}$	(4) ln $PM25_{i,t}$	(5) ln $PM10_{i,t}$	(6) ln $SO2_{i,t}$
ln $PublicVol_{i,t}$	0.043**	0.061***	-0.062	0.043+	0.034***	0.034+
	(0.015)	(0.014)	(0.041)	(0.022)	(0.008)	(0.017)
ln $CongIndex_{i,t}$	0.266+	0.492***	0.682	0.360+	0.306***	0.292+
	(0.149)	(0.136)	(0.544)	(0.216)	(0.075)	(0.176)
Month Effects	YES	YES	YES	YES	YES	YES
City Effects	YES	YES	YES	YES	YES	YES
IV	YES	YES	YES	YES	YES	YES
DML	YES	YES	YES	YES	YES	YES
K	0.209#	0.147	0.024#	0.184#	0.897	0.137#
F-statistic	3.882	11.578	1.170	2.115	8.699	2.314

Note: Standard errors in parentheses. + $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. The K with a "#" is obtained following Hoerl and Kennard (1970a,b), because the procedure following Hoerl and Kennard (1976) fails to converge.