Title
Essays on Structural Microeconometrics

Permalink
https://escholarship.org/uc/item/3b18j5m5

Author
Su, Jiun-Hua

Publication Date
2018

Peer reviewed|Thesis/dissertation
Essays on Structural Microeconometrics

by

Jiun-Hua Su

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

Doctor of Philosophy

in

Economics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor James Powell, Chair
Professor Bryan Graham
Professor Demian Pouzo
Professor Aditya Guntuboyina

Spring 2018
Abstract

Essays on Structural Microeconometrics

by

Jiun-Hua Su

Doctor of Philosophy in Economics

University of California, Berkeley

Professor James Powell, Chair

This dissertation consists of three chapters studying microeconometric methods. The first two chapters focus on models with unobserved heterogeneity, and topics include testing shape restrictions imposed by economic theory and estimating counterfactual policy effects in duration analysis. In the last chapter, predictive methods in machine learning are adapted to study model selection within the framework of utility-maximizing binary decision-making. These proposed methods are described in greater detail below.

Causal inference on the individual treatment effect is fundamental in econometric analysis. In Chapter 1, I develop the concept of structural monotonicity, that is, monotonicity of a structural function in a treatment given any observable covariates and unobserved heterogeneity. Different from regression monotonicity, in which heterogeneous factors average out, structural monotonicity emphasizes the sign of ceteris paribus individual treatment effect. Since economic theory may neither detail enough potential heterogeneous factors nor elaborate on parametric structures, I consider a two-period panel data model with nonseparable time-invariant heterogeneity, and avoid imposing restrictions on the dimensionality of heterogeneity and functional form of the structural function. Structural monotonicity in this setup implies shape constraints on the joint cumulative distribution function (CDF) of outcome variables conditional on the observable treatments and covariates over some regions. These regions are parameterized by a nuisance parameter, which can be consistently estimated. According to the shape constraints on the conditional joint CDF over the estimated regions, I propose a test for structural monotonicity and validate the empirical bootstrap method. Some Monte Carlo experiments show that the proposed test can detect departures from structural monotonicity, which are not revealed by some existing tests for regression monotonicity.
The presence of unobserved heterogeneity is also essential for policy effects especially in duration analysis. In Chapter 2, I propose a counterfactual Kaplan-Meier estimator that incorporates time-invariant exogenous covariates and nonseparable heterogeneity in duration models with random censoring. The over-parameterization in traditional duration analysis can be avoided because distributional features of unobserved heterogeneity are unspecified. I establish the joint weak convergence of the proposed counterfactual Kaplan-Meier estimator and the traditional Kaplan-Meier estimator under some regularity conditions. Therefore, by comparing the estimated counterfactual and original unconditional distribution of the duration variable, we can evaluate the policy effects, for example the change of duration dependence in response to an exogenous manipulation of covariates.

In addition to counterfactual analysis in policy research, a better prediction may improve policy-making. In Chapter 3, I show that in a model of binary decision-making based on the prediction of a binary outcome variable, the semiparametric maximum utility estimation can be viewed as cost-sensitive binary classification. Its in-sample overfitting issue is thus similar to that of perceptron learning in the machine learning literature. To alleviate the in-sample overfitting, I apply techniques in structural risk minimization to construct a utility-maximizing prediction rule. This proposed prediction rule, in comparison to the common machine learning Lasso-logit predictor, has larger relative expected utility in some simulation results when the conditional probability of the binary outcome is misspecified. The results show that a better prediction arising from the combination of machine learning techniques and economic theory can improve policy-making.
To Chin-Ming Su, Mei-Yueh Wang,
and Chiao-Yin Chuang
## Contents

1 Testing Monotonicity in a Model with Nonseparable Time-Invariant Heterogeneity  
   1.1 Introduction ................................................................. 1  
   1.2 Testable implications of structural monotonicity .................... 7  
   1.3 Testing structural monotonicity ...................................... 12  
      1.3.1 The test statistic ................................................... 12  
      1.3.2 Validity of the empirical bootstrap ............................... 15  
      1.3.3 Estimation of \[m\] .................................................. 23  
   1.4 Monte Carlo experiments ............................................... 25  
   1.5 Conclusion ................................................................. 30  

Appendix 1.A Technical proofs ............................................... 33  
   1.A.1 Proof of Proposition 1.1 ............................................. 33  
   1.A.2 Proof of Corollary 1.1 ............................................... 34  
   1.A.3 Proof of Proposition 1.2 ............................................. 34  
   1.A.4 Proof of Proposition 1.3 ............................................. 36  
   1.A.5 Proof of Theorem 1.1 ................................................ 36  
   1.A.6 Proof of Theorem 1.2 ................................................ 38  
   1.A.7 Proof of Corollary 1.2 ............................................... 40  
   1.A.8 Proof of Proposition 1.4 ............................................. 40  

Appendix 1.B Smoothed kernel CDF estimation ................................. 41  
   1.B.1 Notation ................................................................. 42  
   1.B.2 Asymptotic theory .................................................... 46  
   1.B.3 Auxiliary lemmas ..................................................... 57  

Appendix 1.C Robustness checks ............................................... 63  

2 Counterfactual Inference in Duration Models with Random Censoring  
   2.1 Introduction ................................................................. 69  
   2.2 Model and estimation ................................................... 72
2.2.1 Setup and objects of interest ........................................ 72
2.2.2 Nonparametric identification and estimation ................... 74
2.3 Asymptotic theory ..................................................... 77
   2.3.1 Representations .................................................. 77
   2.3.2 Asymptotic properties .......................................... 80
2.4 Monte Carlo simulation ............................................. 83
2.5 Conclusion .......................................................... 85
Appendix 2.A Technical proofs ........................................ 86
   2.A.1 Proof of Proposition 2.1 ...................................... 86
   2.A.2 Proof of Proposition 2.2 ...................................... 87
   2.A.3 Proof of Proposition 2.3 ...................................... 87
   2.A.4 Proof of Theorem 2.1 .......................................... 88
   2.A.5 Proof of Theorem 2.2 .......................................... 89
   2.A.6 Proof of Corollary 2.1 ........................................ 89
   2.A.7 Auxiliary lemmas ................................................. 89

3 Model Selection in Utility-Maximizing Binary Prediction ........... 99
   3.1 Introduction ....................................................... 99
   3.2 Maximum utility estimation ....................................... 101
      3.2.1 Model .......................................................... 101
      3.2.2 Nature of the overfitting in MU estimation ................ 103
   3.3 Model selection ................................................... 104
   3.4 Simulation ........................................................ 109
   3.5 Conclusion ........................................................ 111
Appendix 3.A Technical proofs ........................................ 112
   3.A.1 Proof of Proposition 3.1 ...................................... 112
   3.A.2 Proof of Theorem 3.1 .......................................... 113
   3.A.3 Proof of Corollary 3.1 ........................................ 114
   3.A.4 Auxiliary lemma ................................................ 115

Bibliography .............................................................. 117
## List of Figures

1.1 Brackets for Regions ........................................... 17
1.2 Histogram of $\sqrt{n}(\hat{m}_n - m)$ for 500 simulation replications of DGP1 and DGP2 ........................................... 29
1.3 Couplings ......................................................... 47
1.4 Histogram of $\sqrt{n}(\hat{m}_n - m)$ for 500 simulation replications of DGP3 and DGP4 ........................................... 64
List of Tables

1.1 Rejection Rates for DGP1 with $\bar{a} = 0$ ........................................... 27
1.2 Rejection Rates for DGP1 with $\bar{a} \in \{0, 1, 2\}$ ............................... 28
1.3 Rejection Rates for DGP2 with $m = 0.2$ .............................................. 31
1.4 Rejection Rates for DGP2 with $m \in \{0.2, 0.45, 0.7\}$ ......................... 32
1.5 Rejection Rates for DGP3 with $\bar{a} = 0$ .............................................. 65
1.6 Rejection Rates for DGP3 with $\bar{a} \in \{0, 1, 2\}$ ............................... 66
1.7 Rejection Rates for DGP4 with $m = 0.2$ .............................................. 67
1.8 Rejection Rates for DGP4 with $m \in \{0.2, 0.45, 0.7\}$ ......................... 68

2.1 Estimation of the CDFs ........................................................................... 85
2.2 Estimation of the Cumulative Hazard Functions ....................................... 86

3.1 Relative Expected Utility of ML-logit, MU, and UMPR ......................... 111
3.2 Relative Expected Utility of Lasso-logit and UMPR ............................... 112
I am greatly indebted to my advisor James Powell for valuable guidance. I am also grateful to Peter Bartlett, Bryan Graham, Aditya Guntuboyina, Michael Jansson, Kyoo il Kim, Patrick Kline, Matthew Masten, Christopher Paciorek, Demian Pouzo, Michael Reich, Andrés Rodríguez-Clare, Pedro Sant’anna, Andres Santos, Jeffrey Wooldridge, and anonymous referees for useful suggestions and discussions. I would also like to thank seminar participants at Academia Sinica, Erasmus University Rotterdam, National Chengchi University, National Taiwan University, National Tsing Hua University, UC Berkeley, and the 2017 California Econometrics Conference for helpful comments. This work was supported by the Taiwan Top University Strategic Alliance Graduate Fellowship and the Doctoral Completion Fellowship at UC Berkeley.
Chapter 1

Testing Monotonicity in a Model with Nonseparable Time-Invariant Heterogeneity

Abstract

This chapter develops a test for structural monotonicity, that is, monotonicity of a structural function in an explanatory variable given any observable covariates and non-separable time-invariant unobserved heterogeneity. We show that in a two-period panel data model, under some conditions, structural monotonicity implies shape constraints on the joint cumulative distribution function (CDF) of outcome variables conditional on the explanatory variables and covariates over specific regions. These regions are parameterized by a nuisance parameter, which can be consistently estimated. We propose a test for structural monotonicity according to the shape constraints on the conditional joint CDF over the estimated regions, and validate the empirical bootstrap method under some high-level conditions. Some Monte Carlo experiments show that the proposed test can detect departures from structural monotonicity, which are not revealed by some tests for regression monotonicity, for example tests proposed by Ghosal, Sen, and van der Vaart (2000) and Chetverikov (2017).

1.1 Introduction

Economic theory usually implies shape restrictions on the functions of interest. Among shape restrictions, monotonicity in an observable explanatory variable is commonly imposed. For example, a cost function and a demand function are assumed to be monotone (Matzkin (1994) and Manski (1997)), and liquidity preference hypothesis implies an upward sloping yield curve (Richardson, Richardson, and Smith (1992)).
Imposing monotonicity in econometric models has benefits of identification. For instance, if monotonicity of the first-stage reduced form in both an instrument and unobserved heterogeneity holds, the distribution of potential outcomes in triangular systems with nonseparable heterogeneity can be identified by a reweighting approach (See Kasy (2014) and Hoderlein, Holzmann, Kasy, and Meister (2017)). In addition to point identification of structural objects, the identifying power of assuming monotonicity in an observable variable is also prominent in partial identification. For example, Manski (1997) and Manski and Pepper (2000, 2009) establish sharp bounds on some functionals of the distribution of heterogeneous, monotone response functions.

Given the benefits arising from monotonicity, testing monotonicity is of first concern. In this chapter, we aim to test whether the monotonic relationship between an explanatory variable and an outcome variable of interest holds for any covariates and any unobserved heterogeneity. Since additively separable heterogeneity, while convenient in statistical inference, is often difficult to justify by economic theory, we consider nonseparable heterogeneity. Moreover, we do not restrict the dimensionality of heterogeneity because heterogeneity may be ubiquitous but hard to completely model. However, as is suggested by Manski (1997), the aforementioned monotonic relationship may not be refutable if only cross-sectional data are accessible. Therefore, we assume repeated observations are available and consider a two-period panel data model. Specifically, the outcome variable $Y_{it}$ is generated as

$$Y_{it} = s(X_{it}, W_{it}, Z_i, A_i) + \varepsilon_{it}, \quad i = 1, \ldots, n; \quad t = 1, 2,$$  

(1.1)

where $X_{it}$ is a one-dimensional explanatory variable, $W_{it}$ is a $d_W$-dimensional vector of covariates, $Z_i$ is also a $d_Z$-dimensional vector of covariates, $A_i$ is heterogeneity in an arbitrary measurable space of unrestricted dimensionality, $\varepsilon_{it}$ is an error term, both $Z_i$ and $A_i$ are invariant to the subscript $t$, both $A_i$ and $\varepsilon_{it}$ are unobserved to researchers, and $s(\cdot)$ is a structural function that is also unknown to researchers. The partially nonseparable setup in (1.1) is flexible enough to encompass many existing econometric models. We can regard this setup as a combination of a potential outcome

---

1 For example, Imbens (2007) shows that if unobserved heterogeneous factors such as managerial ability and capital are additively separable from the production function, firms facing the same cost shifter would optimally choose identical amount of input, which might be an untenable implication.

2 Browning and Carro (2007) suggest that there is more heterogeneity in empirical studies than is modelled and that inclusion of limited heterogeneity in econometric modelling may cause incorrect inference for objects of interest.

3 If the structural function $s(\cdot)$ is linear in the explanatory variable with a nonrandom but unknown coefficient $\beta$, the setup admits a variant of partially linear models, $Y_{it} = \beta X_{it} + s_1(W_{it}, Z_i, A_i) + \varepsilon_{it}$ for some unknown real function $s_1$ (cf. Honoré and Powell (2005)). If the unobserved heterogeneity is additively separated from the explanatory variable and covariates, say $Y_{it} = s_2(X_{it}, W_{it}, Z_i) + s_3(A_i) + \varepsilon_{it}$ for some unknown real functions $s_2$
\( Y_x(W_{it}, Z_i, A_i) \equiv s(x, W_{it}, Z_i, A_i) \) with treatment \( x \), covariates \((W_{it}, Z_i)\), and unobserved heterogeneity \( A_i \); and an additively separable error term \( \varepsilon_{it} \), which could be viewed as a measurement error in outcome.\(^4\) The null hypothesis of interest is

\[
H_0 : s(\cdot, w, z, a) \text{ is weakly increasing for all } (w, z, a) \text{ in their joint support.} \tag{1.2}
\]

A structural function is *structurally increasing* if it satisfies the hypothesis (1.2). Similarly, a structural function is *structurally decreasing* if it is weakly decreasing in an explanatory variable given any observable covariates and nonseparable time-invariant unobserved heterogeneity. We say that a structural function has *structural monotonicity* if it is either structurally increasing or structurally decreasing.

Testing structural monotonicity can be viewed as a specification test; furthermore, it can be connected to the literature on treatment effects and program evaluation. Policy makers taking heterogeneity into account would be interested in testing the null hypothesis (1.2), which aims to answer whether *every* individual is better off (measured in \( y \)) after an exogenous increase in \( x \). For example, food manufacturers, out of concerns for ethics or potential loss, would embrace the “better safe than sorry” principle to ensure safety of food additives for every customer. Similarly, to grant approval of new drugs, drug authority would also avoid harmful effects on potential patients. Such a precautionary principle indicates the importance of a policy maker’s subjective evaluation in econometrics, as is argued in Heckman and Vytlacil (2007). Consequently, researchers would be interested in the direction of the *ceteris paribus* individual treatment effect of \( X \) on \( Y \). To the best of our knowledge, this question is not fully answered even though various treatment effects are studied in literature.\(^5\)

We avoid imposing strict monotonicity in the nonseparable heterogeneity component because this assumption, while statistically convenient, is sometimes difficult to motivate economically, as is argued in Hoderlein and Mammen (2007).\(^6\) The proposed structure (1.1) thus allows for many heterogeneous factors and heterogeneity that can be treated

---

\(^4\) The partially nonseparable setup is also adopted in Evdokimov (2010) and Su, Hoderlein, and White (2013) where the additive error term is regarded as an idiosyncratic shock in the context of asset pricing.

\(^5\) Although Abrevaya, Hausman, and Khan (2010) investigate the direction of the individual treatment effect, they exclude the possibility that the treatment may have a positive effect for some values of covariates and a negative effect for other values of covariates; in addition, they do not include unobserved heterogeneity except for a scalar error term.

\(^6\) Tests for monotonicity in scalar unobserved heterogeneity are recently proposed in Su et al. (2013) and Hoderlein, Su, White, and Yang (2016). Test statistics developed in these studies are based on recovery of the unobserved heterogeneity under certain assumptions.
as a random element taking on values in an arbitrary measurable space (for example, heterogeneous preferences and beliefs). The flexibility of modelling heterogeneity however makes the identification of objects of interest such as \( s(x, w, z, a) \) and \( \partial s(x, w, z, a)/\partial x \) for each \((x, w, z, a)\) in the joint support much more difficult, if not impossible.\(^7\)

Inspired by the traditional wisdom that time-invariant heterogeneity could be controlled within a two-period panel data model, we can circumvent the identification of relevant structural objects and deal with the testing problem. We demonstrate that under some regularity conditions, the null hypothesis (1.2) implies the equivalence between the joint CDF of \((Y_1, Y_2)\) conditional on \((X_1, X_2, W_1, W_2, Z)\), and the marginal CDF of \(Y_t\) conditional on \((X_t, W_t, Z)\) over regions that are known up to a nuisance parameter. This nuisance parameter captures the volatility of errors across two periods, and can be consistently estimated by a matching estimator. We thus propose a Kolmogorov-Smirnov type statistic concerning the discrepancy between the joint and marginal CDFs over these estimated regions, and then provide high-level conditions to validate the empirical bootstrap method so that the critical values for the proposed statistic can be simulated. We also show that under extra mild conditions, the proposed test should have power to detect the departure from the null hypothesis (1.2). Our method based on restrictions on the conditional joint CDF can be easily extended to test “structural convexity” if three-period panel data are available because convexity of a function is equivalent to monotonicity of the slope of its chord.

This chapter complements the literature on testing monotonicity in an observable explanatory variable. Existing studies usually focus on regression monotonicity and stochastic monotonicity.\(^8\) Ghosal et al. (2000) build a test statistic that is the supremum of a scaled U-process based on a locally weighted Kendall’s tau under the assumption that the explanatory variable and error term are independent. Following this idea, Gutknecht (2016) modifies the locally weighted Kendall’s tau by the control function approach when the explanatory variable is endogenous, and Chetverikov (2017) allows for an increasing

---

\(^7\) Regarding identification of the structural function, Matzkin (2003) assumes the absence of \(\varepsilon\) and strict monotonicity in scalar heterogeneity \(A\). Altonji and Matzkin (2005) allow for the nonseparable effect of \((A, \varepsilon)\) but this effect is only aggregated by a single index in which the structural function is strictly monotonic. Evdokimov (2010) assumes availability of two-period panel data and also strict monotonicity in scalar heterogeneity \(A\) of the partially nonseparable setup as in (1.1). To achieve the local identification of structural derivatives with respect to observable explanatory variables, Chesher (2003) assumes strict monotonicity in unobserved heterogeneity. Strict monotonicity in heterogeneity also allows for identification of the quantile treatment effect (Chernozhukov and Hansen (2005)) and the quantile structural function (Imbens and Newey (2009)).

\(^8\) Let \(Y\) be a scalar outcome variable and \(X\) a scalar explanatory variable. We say the regression monotonicity of \(Y\) in \(X\) is present if the regression function \(r(\cdot) = \mathbb{E}(Y|X = \cdot)\) is either weakly increasing or weakly decreasing on the support of \(X\). See Gutknecht (2016) and Chetverikov (2017) for regression monotonicity in the presence of other covariates. The stochastic monotonicity of the distribution of \(Y\) given \(X\) is present if \(F_{Y|X}(y|\cdot)\) is either weakly increasing or weakly decreasing on the support of \(X\) for each \(y\). See Lee, Linton, and Whang (2009) for stochastic monotonicity in a vector of explanatory variables.
set of weighting functions with different location points and values of bandwidth so that ex ante information can be incorporated into this adaptive test to increase its power. Lee et al. (2009), assuming smoothness of the conditional distribution of $Y$ given $X$, generalize the locally weighted Kendall’s tau to test stochastic monotonicity. By applying the adaptive procedure in Chetverikov (2017), Chetverikov and Wilhelm (2017) improve the power of testing stochastic monotonicity. Relaxing the smoothness assumption in Lee et al. (2009), Delgado and Escanciano (2012) characterize stochastic monotonicity by concavity of the copula of $Y$ and $X$ and propose an alternative test based on the departure of the copula from its partial least concave majorant. Seo (2016) further improves the power of the test in Delgado and Escanciano (2012) by using a bootstrap procedure to obtain the data-dependent critical value rather than critical value evaluated at the least favorable case. Recently, Hsu, Liu, and Shi (2018) propose a test for the hypothesis of generalized regression monotonicity in which both regression monotonicity and stochastic monotonicity are nested. Further references are documented in these studies.

None of these tests focuses on structural monotonicity; to be specific, none tests monotonicity in an explanatory variable given any unobserved heterogeneity. Since structural monotonicity implies regression monotonicity and stochastic monotonicity, the null hypothesis of structural monotonicity is violated if we reject regression monotonicity or stochastic monotonicity or both. Thus, existing tests for regression monotonicity and stochastic monotonicity offer insights about structural monotonicity. However, failure to reject regression monotonicity and stochastic monotonicity does not necessarily imply structural monotonicity in the presence of nonseparable heterogeneity. The following two examples illustrate this observation.

**Example 1** (Regression Monotonic but not Structurally Monotonic)

Let $Y = (A - \log 2)X + \varepsilon$, where $X$ is independent of $(A, \varepsilon)$, $A$ is exponentially distributed with mean one, and $\varepsilon$ has mean $c_\varepsilon$. The regression function $r(x) \equiv E(Y|X = x) = c_\varepsilon + (1 - \log 2)x$ is strictly increasing in $x$ but the structural function $s(x, a) \equiv (a - \log 2)x$ is strictly decreasing in $x$ when $a < \log 2$. Thus, the regression monotonicity represents an opposite effect for $P(A < \log 2) = 50\%$ of individuals.

**Example 2** (Stochastically Monotonic but not Structurally Monotonic)

Let $Y = AX + \varepsilon$ where both $A$ and $\varepsilon$ are symmetrically distributed around 0 and $X$.

---

9 Suppose $A$ and $\varepsilon$ are independent of $X$ conditional on $(W, Z)$. Under the null hypothesis (1.2), the mean regression function $r(x, w, z) \equiv E(Y|X = x, W = w, Z = z)$ is weakly increasing in $x$ for all $(w, z)$; in addition, the CDF $F_{Y|XWZ}(y|x, w, z)$ is weakly decreasing in $x$ for all $(y, w, z)$. 
takes on 1 and $-1$ with positive probability $p_0$ and $1 - p_0$, respectively. Suppose $X$, $A$, and $\varepsilon$ are mutually independent. The stochastic monotonicity of $F_{Y|X}$ is present because for all $y \in \mathbb{R}$,

$$F_{Y|X}(y|X = 1) = F_{Y|X}(y|X = -1).$$

In addition, the regression function $r(x) \equiv \mathbb{E}(Y|X = x) = 0$ satisfies regression monotonicity and the regression error is independent of $X$. However, one half of individuals have the structural function $s(x, a)$ strictly decreasing in $x$ because $\mathbb{P}(A < 0) = 50\%$.  

These two examples show that tests for regression monotonicity and stochastic monotonicity may have power equal to size in testing structural monotonicity and hence may be inconsistent for some departures from structural monotonicity. This phenomenon has an empirical implication: Even if results of testing regression monotonicity and stochastic monotonicity fail to reject monotonicity, policy makers should still be cautious because regression monotonicity or stochastic monotonicity may represent an opposite effect for at least one half of individuals. Some Monte Carlo experiments carried out in this study show that the proposed Kolmogorov-Smirnov type test statistic, constructed by smoothed Nadaraya-Watson kernel CDF estimation, can detect departures from structural monotonicity, which are not revealed by some tests for regression monotonicity, for example tests proposed by Ghosal et al. (2000) and Chetverikov (2017).

As is also illustrated by the examples, the existence of nonseparable heterogeneity makes structural monotonicity different from regression monotonicity and stochastic monotonicity. The literature on nonseparable heterogeneity focuses on identification and estimation of average and quantile effects, for example the average partial effect (Blundell and Powell (2003), Wooldridge (2005), Graham and Powell (2012)), the average treatment effect (Florens, Heckman, Meghir, and Vytlacil (2008)), and the quantile treatment effect (Chernozhukov and Hansen (2005)). In contrast, we study the direction of the individual treatment effect.

Throughout this chapter, we assume that the underlying probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is complete; that is, all subsets of $\mathbb{P}$-null sets are $\mathcal{F}$-measurable. We also assume that $\{Y_{i1}, Y_{i2}, X_{i1}, X_{i2}, W_{i1}, W_{i2}, Z_i, A_i, \varepsilon_{i1}, \varepsilon_{i2}\}_{i=1}^n$ are independent and identically distributed across subscript $i$. For notational convenience, we usually suppress the subscript $i$ indexing individuals in the remainder of this chapter. For simplicity, the vector of $(\varepsilon_1, \varepsilon_2)$ is assumed to have the same marginal support of each component, and the similar assumption is also imposed on $(Y_1, Y_2)$, $(X_1, X_2)$, and $(W_1, W_2)$, respectively. We write $1_{[B]}$ for the indicator function, which equals one if event $B$ obtains and equals zero otherwise.
The remainder of this chapter is organized as follows. Section 1.2 describes testable implications of the null hypothesis (1.2). In Section 1.3, we propose the Kolmogorov-Smirnov type statistic and establish its rejection region by the empirical bootstrap method, while in Section 1.4, we discuss the finite-sample performance of the proposed test by Monte Carlo simulation. Section 1.5 concludes the chapter with some future developments. Statements of regularity conditions and technical proofs, and robustness checks are gathered in the Appendix.

1.2 Testable implications of structural monotonicity

The object in this chapter is to develop a test to detect departures from the null hypothesis (1.2) not revealed by existing tests for regression monotonicity. To achieve this goal, we seek conditions so that the null hypothesis (1.2) implies testable constraints on the observable data distribution under these maintained conditions.

Structural monotonicity implies shape constraints on the joint CDF \( F_{Y_1Y_2|X_1X_2W_1W_2Z} \) of \((Y_1, Y_2)\) conditional on \((X_1, X_2, W_1, W_2, Z)\). To clarify the idea, suppose for simplicity that \((W_1, W_2, Z)\) has null dimension; that is, \(Y_t = s(X_t, A) + \varepsilon_t\) for \(t = 1, 2\). Let \(m\) be the supremum of the support of \(|\varepsilon_2 - \varepsilon_1|\); concretely, \(m \equiv \sup\{|\varepsilon_2 - \varepsilon_1| : (\varepsilon_1, \varepsilon_2) \in E^2\}\) and \(E\) is the support of \(\varepsilon_t\). The null hypothesis (1.2) implies that it is impossible to have an observation \((x_1, x_2, y_1, y_2)\) with \(x_2 \geq x_1\) and \(y_1 - y_2 > m\). To see this, note that the difference between \(y_1\) and \(y_2\) can be decomposed into the variation in error terms and the structural difference due to the variation in \(x\):

\[
y_1 - y_2 = [s(x_1, a) - s(x_2, a)] + [\epsilon_1 - \epsilon_2]
\]

for some realized heterogeneity and error terms \((a, \epsilon_1, \epsilon_2)\). The difference between error terms \(\epsilon_1 - \epsilon_2\) is bounded above by \(m\); moreover, if \(x_2 \geq x_1\), then the structural difference \(s(x_1, a) - s(x_2, a)\) is nonpositive under the null hypothesis (1.2). Therefore, given \(x_2 \geq x_1\), the conditional joint CDF \(F_{Y_1Y_2|X_1X_2}\) admits no mass at \((y_1, y_2)\) with \(y_1 - y_2 > m\). It follows that the conditional joint CDF is equal to one conditional marginal CDF; specifically,

\[
F_{Y_1Y_2|X_1X_2}(y_1, y_2|x_1, x_2) = F_{Y_1Y_2|X_1X_2}(\infty, y_2|x_1, x_2) = F_{Y_2|X_1X_2}(y_2|x_1, x_2)
\]

whenever \(x_2 \geq x_1\) and \(y_1 - y_2 \geq m\). By symmetry, we have

\[
F_{Y_1Y_2|X_1X_2}(y_1, y_2|x_1, x_2) = F_{Y_1|X_1X_2}(y_1|x_1, x_2)
\]

whenever \(x_1 \geq x_2\) and \(y_2 - y_1 \geq m\). Loosely speaking, we have the equivalence between the joint CDF and one marginal CDF whenever \(|y_1 - y_2|\) is larger than \(m\). In the
presence of covariates \((W_1, W_2, Z)\), these arguments are valid when we focus on the case \(w_1 = w_2\), which ensures the difference between realized outcomes, \(y_1 - y_2\), is attributed to either the variation in error terms or the structural difference due to the variation in \(x\).

These findings show that the null hypothesis (1.2) implies shape constraints on the conditional joint CDF \(F_{Y_1Y_2|X_1X_2W_1W_2Z}\) over two regions of possible values of \((y_1, y_2, x_1, x_2, w_1, w_2, z)\), namely

\[
\mathcal{R}_1(m) \equiv \{(y_1, y_2, x_1, x_2, w_1, w_2, z) \in \mathcal{Y}^2 \mathcal{X}^2 \mathcal{W}^2 \mathcal{Z} : x_1 \geq x_2, w_1 = w_2, \text{ and } y_2 - y_1 \geq m\}
\]

and

\[
\mathcal{R}_2(m) \equiv \{(y_1, y_2, x_1, x_2, w_1, w_2, z) \in \mathcal{Y}^2 \mathcal{X}^2 \mathcal{W}^2 \mathcal{Z} : x_2 \geq x_1, w_1 = w_2, \text{ and } y_1 - y_2 \geq m\}
\]

where a calligraphy letter is used to denote the support of an associated variable. Specifically, Proposition 1.1 below demonstrates that conditional on \((X_1, X_2, W_1, W_2, Z)\), the joint CDF of \((Y_1, Y_2)\) is equal to the marginal CDF of \(Y_i\) over \(\mathcal{R}_i(m)\) for each \(t\).

**Proposition 1.1.** For each \(t \in \{1, 2\}\), we have

\[
F_{Y_1|X_1X_2W_1W_2Z}(y_1|x_1, x_2, w_1, w_2, z) - F_{Y_1Y_2|X_1X_2W_1W_2Z}(y_1, y_2|x_1, x_2, w_1, w_2, z) \geq 0
\]

for all \((y_1, y_2, x_1, x_2, w_1, w_2, z) \in \mathcal{Y}^2 \mathcal{X}^2 \mathcal{W}^2 \mathcal{Z}\). Furthermore, under the null hypothesis (1.2), the joint distribution of \((Y_1, Y_2)\) given \((X_1, X_2, W_1, W_2, Z)\) satisfies

\[
F_{Y_1Y_2|X_1X_2W_1W_2Z}(y_1, y_2|x_1, x_2, w_1, w_2, z) = \begin{cases} 
F_{Y_1|X_1X_2W_1W_2Z}(y_1|x_1, x_2, w_1, w_2, z), & \text{if } (y_1, y_2, x_1, x_2, w_1, w_2, z) \in \mathcal{R}_1(m) \\
F_{Y_2|X_1X_2W_1W_2Z}(y_2|x_1, x_2, w_1, w_2, z), & \text{if } (y_1, y_2, x_1, x_2, w_1, w_2, z) \in \mathcal{R}_2(m)
\end{cases}
\]

where \(m \equiv \sup\{|\varepsilon_2 - \varepsilon_1| : (\varepsilon_1, \varepsilon_2) \in \mathcal{E}^2\}\) is the supremum of the support of \(|\varepsilon_2 - \varepsilon_1|\). \(\square\)

The “symmetry” of support of \(\varepsilon_2 - \varepsilon_1\) in Proposition 1.1 can be weakened as follows. Let \(\overline{m}\) and \(\underline{m}\) be the supremum and infimum of the support of \(\varepsilon_2 - \varepsilon_1\), respectively. A simple modification of the proof of Proposition 1.1 shows that the null hypothesis (1.2) implies

\[
F_{Y_1Y_2|X_1X_2W_1W_2Z}(y_1, y_2|x_1, x_2, w_1, w_2, z) = \begin{cases} 
F_{Y_1|X_1X_2W_1W_2Z}(y_1|x_1, x_2, w_1, w_2, z), & \text{if } (y_1, y_2, x_1, x_2, w_1, w_2, z) \in \mathcal{R}_1(\overline{m}) \\
F_{Y_2|X_1X_2W_1W_2Z}(y_2|x_1, x_2, w_1, w_2, z), & \text{if } (y_1, y_2, x_1, x_2, w_1, w_2, z) \in \mathcal{R}_2(\underline{m}).
\end{cases}
\]

For ease of exposition, we maintain the “symmetry” of support of \(\varepsilon_2 - \varepsilon_1\) throughout this chapter.
In empirical studies that exogenous manipulation of \((X, W)\) is possible, we may impose the following conditional exogeneity:\(^{10}\)

**D1** Conditional on the control vector \(Z\), the unobserved heterogeneity \(A\) and error terms \((\varepsilon_1, \varepsilon_2)\) are jointly independent of the explanatory variables \((X_1, X_2)\) and covariates \((W_1, W_2)\); specifically, \((X_1, X_2, W_1, W_2) \perp (A, \varepsilon_1, \varepsilon_2) \mid Z\).

With Condition D1, we immediately have the following corollary.

**Corollary 1.1.** Suppose Condition D1 holds. Under the null hypothesis (1.2), the joint distribution of \((Y_1, Y_2)\) given \((X_1, X_2, W_1, W_2, Z)\) satisfies

\[
F_{Y_1 Y_2 | X_1 X_2 W_1 W_2 Z}(y_1, y_2 | x_1, x_2, w_1, w_2, z) = \begin{cases} 
F_{Y_1 | X_1 W_1 Z}(y_1 | x_1, w_1, z), & \text{if } (y_1, y_2, x_1, x_2, w_1, w_2, z) \in \mathcal{R}_1(m) \\
F_{Y_2 | X_2 W_2 Z}(y_2 | x_2, w_2, z), & \text{if } (y_1, y_2, x_1, x_2, w_1, w_2, z) \in \mathcal{R}_2(m) 
\end{cases}
\]

where \(m \equiv \sup\{|\varepsilon_2 - \varepsilon_1| : (\varepsilon_1, \varepsilon_2) \in \mathcal{E}^2\}\) is the supremum of the support of \(|\varepsilon_2 - \varepsilon_1|\). □

Condition D1 is closely related to the assumption that given the observable vector \(Z\) of covariates, the potential outcomes are independent of treatment \((X_1, X_2, W_1, W_2)\), an assumption called *unconfoundedness* in the literature on treatment effects.\(^{11}\) Indeed, the unconfoundedness \((X_1, X_2, W_1, W_2) \perp \{Y_z(w, Z, A), (x, w) \in \mathcal{XW}\} \mid Z\) and the conditional exogeneity \((X_1, X_2, W_1, W_2) \perp A \mid Z\) are equivalent when the structural function \(s(\bar{x}, \bar{w}, z, \cdot)\) is invertible for all \(z \in Z\) given some \((\bar{x}, \bar{w}) \in \mathcal{XW}\).\(^{12}\)

When the control vector \(Z\) has null dimension, Condition D1 reduces to “unconditional” exogeneity and hence restricts the joint distribution of \((A, \varepsilon_1, \varepsilon_2)\) given \((X_1, X_2, W_1, W_2)\). Furthermore, if subscript \(t\) refers to time in panel data models, Condition D1 excludes dynamic models and feedback effects because lagged outcome variables are not permitted in the explanatory variable and time-varying covariates. Although the random effect heterogeneity implied by Condition D1 may be restrictive in some empirical studies, the flexibility of nonseparable structure still allows for a correlated random coefficient model, for instance \(Y_t = \bar{s}_1(W_t, A) + \bar{s}_2(W_t, A)X_t + \varepsilon_t\) for some unknown real functions \(\bar{s}_1(\cdot)\) and \(\bar{s}_2(\cdot)\). The “intercept” \(\bar{s}_1(W_t, A)\) may be time-varying

---

\(^{10}\) Condition D1 is called conditional exogeneity by White and Chalak (2013) and similar variants are adopted by recent research, for example Blundell and Powell (2003), Altonji and Matzkin (2005), Hoderlein and Mammen (2007), and Lu and White (2014).

\(^{11}\) Imbens (2004) and Imbens and Wooldridge (2009) point out that the unconfoundedness and its variants are also referred to selection on observables or conditional independence.

\(^{12}\) Suppose \((X_1, X_2, W_1, W_2) \perp A \mid Z\) holds. By Lemmas 4.1 and 4.2(i) of Dawid (1979), the unconfoundedness \((X_1, X_2, W_1, W_2) \perp \{Y_z(w, Z, A), (x, w) \in \mathcal{XW}\} \mid Z\) holds. Suppose the unconfoundedness holds. By Lemmas 4.1 and 4.2(ii) of Dawid (1979), \((X_1, X_2, W_1, W_2) \perp (s(\bar{x}, \bar{w}, Z, A), Z) \mid Z\). Since \(A = s^{-1}(\bar{x}, \bar{w}, Z, s(\bar{x}, \bar{w}, Z, A))\), applying Lemma 4.2(ii) of Dawid (1979) again shows \((X_1, X_2, W_1, W_2) \perp A \mid Z\).
and correlated with the explanatory variable \(X_t\), which is in sharp contrast to the usual assumption in random effect models that the time-invariant heterogeneity is uncorrelated with the explanatory variable.

According to Corollary 1.1, we can test the null hypothesis (1.2) by testing shape constraints on the joint CDF of \((Y_1, Y_2)\) conditional on explanatory variables and covariates \((X_1, X_2, W_1, W_2, Z)\) over the regions \(R_1(m)\) and \(R_2(m)\). A natural question is to what extent tests based on these shape constraints are consistent and powerful. Since these shape constraints under the alternative hypothesis and Condition D1 are still satisfied if \(m = \infty\), we consider the following condition:

D2 The supremum of the support of \(|\varepsilon_2 - \varepsilon_1|\) is finite and bounded away from zero; that is, \(m \equiv \sup\{ |\varepsilon_2 - \varepsilon_1| : (\varepsilon_1, \varepsilon_2) \in \mathcal{E}^2 \} \in (0, \infty)\).

Condition D2 restricts the volatility of error terms across subscript \(t\) but it is possible that each error term is normally distributed under Condition D2. For example, if \(\varepsilon_1\) is normally distributed with mean zero and variance \(\sigma^2_{\varepsilon}\), then for any \(c > 0\), \(\varepsilon_2 = \varepsilon_1(1 - 2\mathbb{I}_{|\varepsilon_1| \leq \varepsilon})\) has the same distribution as \(\varepsilon_1\) and \(|\varepsilon_2 - \varepsilon_1| \leq 2c\).\(^{13}\)

We introduce an extra regularity condition:

D3 Given any \((x, w, z) \in \mathcal{XWZ}\), the structural function \(s(x, w, z, \cdot)\) is continuous with respect to heterogeneity.

The continuity condition in D3 is weaker than strict monotonicity in heterogeneity because the latter implies differentiability of the structural function with respect to heterogeneity almost everywhere in its support \(\mathcal{A}\).\(^{14}\) Technically speaking, the support \(\mathcal{A}\) may not be equipped with a metric or a total order; instead, only a topology on \(\mathcal{A}\) is enough for Condition D3. Conceptually, it is expected that structural functions for similar individuals may have similar features; therefore, the continuity of a structural function with respect to heterogeneity \(\mathcal{A}\) is plausible.

Since the structural function in (1.1) can include a nonrandom intercept term without loss of generality, any location shift of the support \(\mathcal{E}\) of error terms is innocuous and hence the support of error terms can be normalized to contain zero. In addition, it is in general of no interest that non-monotonicity of a structural function in an explanatory variable only occurs at heterogeneity with probability zero; therefore, we exclude these

\(^{13}\) In this example, although \(\varepsilon_1\) and \(\varepsilon_2\) are dependent for any \(c > 0\), there is a constant \(c_0 > 0\) such that \(\varepsilon_1\) and \(\varepsilon_2 = \varepsilon_1(1 - 2\mathbb{I}_{|\varepsilon_1| \leq c})\) are uncorrelated. Moreover, Condition D4’ below is satisfied; specifically, both \(P(\varepsilon_1 < 0, \varepsilon_2 > 0)\) and \(P(\varepsilon_1 > 0, \varepsilon_2 < 0)\) are greater than or equal to \(0.5 - \Phi(-c/\sigma) > 0\).

\(^{14}\) In fact, to prove Proposition 1.2 below, we can weaken Condition D3 and only require continuity with respect to heterogeneity of the structural function at \((x, w, z, a)\) and \((x', w, z, a)\) with \(x > x'\) and \(s(x', w, z, a) - s(x, w, z, a) > m\)
negligible alternatives and only focus on heterogeneity in the support $A^z$ of $A$ conditional on $Z = z$ for any $z \in Z$.\footnote{Heterogeneity $a$ is in the support of $A$ conditional on $Z = z$ if $P(A \in \mathcal{N}_a \mid Z = z) > 0$ for any neighborhood $\mathcal{N}_a$ of $a$. See Chung (2000).} Furthermore, it should be possible that $\varepsilon_1$ and $\varepsilon_2$ have different signs whenever the errors are not perfectly dependent. Finally, we assume that the support $Y$ of the outcome variable $Y$ is convex. This is a technical condition and is satisfied when $Y$ is continuously distributed on an interval in $\mathbb{R}$. We summarize the discussion regarding the support of error terms and that of outcome variables as follows.

**D4** (i) For any $(x, w, z, a) \in XWA$, the support of $\varepsilon_t$ given $(X_t, W_t, Z, A) = (x, w, z, a)$ contains 0 for each $t \in \{1, 2\}$.  
(ii) For any $z \in Z$ and $a \in A^z$, there is a neighborhood $\mathcal{N}_a$ of $a$ such that $P(\varepsilon_1 < 0, \varepsilon_2 > 0 \mid A = a', Z = z) > 0$ and $P(\varepsilon_1 > 0, \varepsilon_2 < 0 \mid A = a', Z = z) > 0$ for every $a' \in \mathcal{N}_a$.  
(iii) The support $Y$ is convex.

Condition D4 still allows for the possibility that error terms are correlated with explanatory variables, covariates, and heterogeneity. When $(\varepsilon_1, \varepsilon_2)$ is viewed as a vector of measurement errors independent of $(X_1, X_2, W_1, W_2, Z, A)$, Condition D4 reduces to Condition D4’:

**D4’** (i) The support of $\varepsilon_t$ contains 0 for each $t \in \{1, 2\}$. (ii) Both $P(\varepsilon_1 < 0, \varepsilon_2 > 0)$ and $P(\varepsilon_1 > 0, \varepsilon_2 < 0)$ are greater than 0. (iii) The support $Y$ is convex.

We partition the collection of alternatives into two sets $S_m$ and $S_m^c$, where

$$S_m \equiv \left\{ s : XWZA \to \mathbb{R} : s(x', w, z, \bar{a}) - s(x, w, z, \bar{a}) > m \text{ for some } (x, x', w, z, \bar{a}) \in X^2WZA \text{ with } x > x', \text{ and } \bar{a} \in A^z \right\}$$

is the collection of structural functions with “magnitude” of non-monotonicity greater than $m$, and $S_m^c$ is the rest of structural functions in the alternatives. Proposition 1.2 below suggests that tests based on the shape constraints on the conditional joint CDF in Corollary 1.1 would be consistent and powerful against alternatives in $S_m$ if extra conditions such as Conditions D2-D4 hold.

**Proposition 1.2.** Suppose Conditions D1-D4 are satisfied. The shape constraints on the conditional joint CDF $F_{Y_1Y_2|X_1X_2W_1W_2Z}$ in Corollary 1.1 do not hold whenever the underlying structural function is in $S_m$.\quad\square
Although Proposition 1.2 does not guarantee that tests based on the shape constraints on the conditional joint CDF have power against alternatives in $S^c_m$, the Monte Carlo experiments in Section 1.4 show that the Kolmogorov-Smirnov type tests may still detect alternatives in $S^c_m$. Given the flexible setup in (1.1), the collection of alternatives could be well approximated by the set $S_m$ if relevant data are not subject to large measurement errors. When the measurement errors are minor, it is expected that volatility of outcome variables (across subscript $t$) can be accounted for mainly by the structural difference due to changes in the explanatory variable and covariates.

Furthermore, although Corollary 1.1 holds whenever $m$ is finite, it is nonetheless difficult to test the shape constraints in practice when $m$ is large. To see this, notice that under the null hypothesis, the conditional joint CDF is restricted over the regions $R_1(m)$ and $R_2(m)$; however, points in these regions have the component $(y_1, y_2)$ with $|y_1 - y_2| > m$ at the boundary of $Y^2$ if $m$ is relatively large in comparison to the magnitude of structural difference. The availability of few data around the boundary would make estimation of the conditional CDFs evaluated at these boundary points imprecise.

Even when errors across subscript $t$ are highly dependent in the sense that they have the same sign with probability one, tests based on the shape constraints in Corollary 1.1 would still be consistent and powerful against a smaller subset $S_{2m}$ of alternatives if the marginal support $\mathcal{E}$ of $\varepsilon_t$ is bounded and equal to the support of $\varepsilon_t$ conditional on $(X_t, W_t, Z, A)$ for each $t \in \{1, 2\}$. Consequently, we consider Condition $D4''$ and summarize the result in Proposition 1.3 below.

**Proposition 1.3.** Suppose Conditions $D1$-$D3$ and $D4''$ are satisfied. The shape constraints on the conditional joint CDF $F_{Y_1Y_2|X_1X_2W_1W_2Z}$ in Corollary 1.1 do not hold whenever the underlying structural function is in $S_{2m}$.

### 1.3 Testing structural monotonicity

#### 1.3.1 The test statistic

Corollary 1.1 shows that under Condition $D1$, the null hypothesis (1.2) implies the equivalence between $F_{Y_1Y_2|X_1X_2W_1W_2Z}$ and $F_{Y_1|X_1W_1Z}$ on $R_t(m)$ for each $t \in \{1, 2\}$. The implied hypothesis should be rejected if large values of the scaled discrepancy between
Although replacing $m$ where CDFs and the nuisance parameter $m$ asymptotic or bootstrap properties of these discrepancy measures are unknown especially when the conditional excursion probabilities of some weak limit implies that the limiting null distribution of of the underlying process $m$

Suppose that we know the nuisance parameter $m$.

Note that for any $\delta > 0$

$F_{t \mid X_i}$ between $R^*$ is taken over an estimated region $\mathcal{R}_t(m)$; in this case, the equivalence between $F_{Y_1 \mid X_1} \approx F_{Y_1 \mid X_1, X_2 W_1 W_2}$ and $F_{Y_1 \mid X_i W_i}$ due to Corollary 1.1.

Therefore, given some small $\delta_0 > 0$, we define a feasible test statistic as

$$
\hat{T}_n(m + \delta_0) 
\equiv \sup \left\{ \hat{Q}_n(y_1, y_2, x_1, x_2, w_1, w_2, z) : (y_1, y_2, x_1, x_2, w_1, w_2, z) \in \mathcal{R}_t(m + \delta_0) \right\}
$$

for each $t \in \{1, 2\}$.

Statistical inference based on the limiting distribution of $\hat{T}_n(m + \delta_0)$ is complicated. Suppose that we know the nuisance parameter $m$ and can establish the weak convergence of the underlying process $\hat{Q}_m$. In this case, although the continuous mapping theorem implies that the limiting null distribution of $\hat{T}_n(m + \delta_0)$ could be calculated by the excursion probabilities of some weak limit $\mathcal{G}_t$ over the region $\mathcal{R}_t(m + \delta_0)$, namely

$$
P \left( \sup \left\{ \mathcal{G}_t(y_1, y_2, x_1, x_2, w_1, w_2, z) : (y_1, y_2, x_1, x_2, w_1, w_2, z) \in \mathcal{R}_t(m + \delta_0) \right\} \geq u \right)
$$

\footnote{In addition to the Kolmogorov-Smirnov type statistic, other scaled discrepancy measures such as the Cramér-von Mises and Anderson-Darling type statistics may yield tests for structural monotonicity. However, the asymptotic or bootstrap properties of these discrepancy measures are unknown especially when the conditional CDFs and the nuisance parameter $m$ are estimated nonparametrically.}
for each $u \in \mathbb{R}$, the distribution of $G_t$ under the null hypothesis is generally nonpivotal because its covariance function depends on the unknown, though consistently estimable, joint CDF of $(Y_1, Y_2, X_1, X_2, W_1, W_2, Z)$. The possible nonstationarity of $G_t$ further makes the evaluation of its excursion probabilities difficult. The evaluation is even more challenging because the process $\hat{Q}_{tn}$ may not converge weakly and the nuisance parameter $m$ is usually unknown in practice.

To avoid the challenge of calculating the limiting distribution of $\hat{T}_{tn}(\hat{m}_n + \delta_0)$, we adopt an empirical bootstrap method to approximate the distribution of $\hat{T}_{tn}(\hat{m}_n + \delta_0)$. In particular, with the data $\mathcal{D}_n \equiv \{Y_{i1}, Y_{i2}, X_{i1}, X_{i2}, W_{i1}, W_{i2}, Z_i\}_{i=1}^n$ of sample size $n$, we randomly draw the bootstrap data $\mathcal{D}_n^* \equiv \{Y_{i1}^*, Y_{i2}^*, X_{i1}^*, X_{i2}^*, W_{i1}^*, W_{i2}^*, Z_i^*\}_{i=1}^n$ of size $n$ from $\mathcal{D}_n$ with replacement, construct the bootstrap CDF counterparts $\hat{F}_{Y_1|X_1,W_1,Z}^*$ and $\hat{F}_{Y_1,Y_2|X_1,X_2,W_1,W_2,Z}^*$, and finally obtain the bootstrap test statistic

$$
\hat{T}_{tn}(\hat{m}_n + \delta_0) \equiv \sup \left\{ \hat{Q}_{tn}^*(y_1, y_2, x_1, x_2, w_1, w_2, z) : (y_1, y_2, x_1, x_2, w_1, w_2, z) \in \mathcal{R}_t(\hat{m}_n + \delta_0) \right\} \quad (1.5)
$$

where

$$
\hat{Q}_{tn}^*(y_1, y_2, x_1, x_2, w_1, w_2, z) \equiv \varphi_n \cdot \left[ \left( \hat{F}_{Y_1|X_1,W_1,Z}^*(y_t|x_t, w_t, z) - \hat{F}_{Y_1,Y_2|X_1,X_2,W_1,W_2,Z}^*(y_1, y_2|x_1, x_2, w_1, w_2, z) \right) \\
- \left( \hat{F}_{Y_1|X_1,W_1,Z}^*(y_t|x_t, w_t, z) - \hat{F}_{Y_1,Y_2|X_1,X_2,W_1,W_2,Z}(y_1, y_2|x_1, x_2, w_1, w_2, z) \right) \right]
$$

for each $t \in \{1, 2\}$. The bootstrap test statistic (1.5) is based on the supremum of scaled discrepancy $\hat{F}_{Y_1|X_1,W_1,Z}^* - \hat{F}_{Y_1,Y_2|X_1,X_2,W_1,W_2,Z}^*$ centered at $\hat{F}_{Y_1|X_1,W_1,Z} - \hat{F}_{Y_1,Y_2|X_1,X_2,W_1,W_2,Z}$ over the estimated region $\mathcal{R}_t(\hat{m}_n + \delta_0)$ where $\hat{m}_n$ is estimated by the sample $\mathcal{D}_n$; namely, we only apply the bootstrap to the conditional CDF estimators. It is expected that when the null hypothesis (1.2) and Condition D1 are valid, the distribution of $\hat{T}_{tn}(\hat{m}_n + \delta_0)$ can be approximated by the conditional distribution of $\hat{T}_{tn}(\hat{m}_n + \delta_0)$ given the sample $\mathcal{D}_n$ under regularity conditions because $\hat{Q}_{tn}$ is centered at $F_{Y_1|X_1,W_1,Z} - F_{Y_1,Y_2|X_1,X_2,W_1,W_2,Z} = 0$ over the region $\mathcal{R}_t(m + \delta_0)$ and $\hat{m}_n$ is consistent for $m$.

Since there are many estimators of the conditional CDFs in literature and different estimators have different (asymptotically) distributional properties, we propose high-level conditions on the supremum statistics and $\hat{m}_n$ to establish the validity of the empirical bootstrap approach, and describe algorithms to test the null hypothesis (1.2) in next subsection.
1.3.2 Validity of the empirical bootstrap

To establish the validity of the empirical bootstrap, we first introduce some notation and conditions. For each $t \in \{1, 2\}$, let $\hat{T}_{tn}^0$ and $\hat{T}_{tn}^*$ be the centered and empirical bootstrap counterpart of $T_{tn}$, respectively; specifically,

$$\hat{T}_{tn}^0(\cdot) \equiv \sup \left\{ \hat{Q}_{tn}^0(y_1, y_2, x_1, x_2, w_1, w_2, z) : (y_1, y_2, x_1, x_2, w_1, w_2, z) \in R_t(\cdot) \right\}$$

and

$$\hat{T}_{tn}^*(\cdot) \equiv \sup \left\{ \hat{Q}_{tn}^*(y_1, y_2, x_1, x_2, w_1, w_2, z) : (y_1, y_2, x_1, x_2, w_1, w_2, z) \in R_t(\cdot) \right\}$$

where

$$\hat{Q}_{tn}^0(y_1, y_2, x_1, x_2, w_1, w_2, z) \equiv \varphi_n \cdot \left[ \left( \hat{F}_{Y_1|X_1} Y_2 X_1 W Z \left( y_1 \big| x_1, w_1 \right) - \hat{F}_{Y_1|X_1} Y_2 X_1 Z \left( y_1 \big| x_1, w_1 \right) \right) - \left( F_{Y_1|X_1} Y_2 X_1 W Z \left( y_1 \big| x_1, w_1 \right) - F_{Y_1|X_1} Y_2 X_1 W Z \left( y_1 \big| x_1, w_1 \right) \right) \right]$$

and $\hat{Q}_{tn}^*(y_1, y_2, x_1, x_2, w_1, w_2, z)$ is defined after the bootstrap test statistic (1.5). Let $l^\infty (R_t(m))$ be the collection of all bounded functions $\xi : R_t(m) \to \mathbb{R}$, equipped with the uniform norm $\|\xi\|_\infty \equiv \sup_{u \in R_t(m)} |\xi(u)|$.

**Condition [EB]:**

**EB1** There is a consistent estimator $\hat{m}_n$ of $m$, and it converges faster than some rate $v_n$; that is, $|\hat{m}_n - m| = o_p(v_n)$ where $v_n \to 0$ as $n \to \infty$.

**EB2** For each $t \in \{1, 2\}$, there is a sequence $\{G_{tn}\}_{n=1}^\infty$ of tight centered Gaussian processes in $(l^\infty (R_t(m)), \|\cdot\|_\infty)$ such that for each $\ell \in \{-1, 1\}$ and $\delta > 0$, as $n \to \infty$,

(i)

$$\sup_{u \in \mathbb{R}} \left| \mathbb{P} \left( \hat{T}_{tn}^0(m + \ell v_n + \delta) \leq u \right) - \mathbb{P} \left( T_{tn}(m + \ell v_n + \delta) \leq u \right) \right| = o(1)$$

and

(ii)

$$\sup_{u \in \mathbb{R}} \left| \mathbb{P} \left( \hat{T}_{tn}^*(m + \ell v_n + \delta) \leq u \big| D_n \right) - \mathbb{P} \left( T_{tn}(m + \ell v_n + \delta) \leq u \big| D_n \right) \right| = o_p(1)$$

where $T_{tn}(\cdot) \equiv \sup \{G_{tn}(y_1, y_2, x_1, x_2, w_1, w_2, z) : (y_1, y_2, x_1, x_2, w_1, w_2, z) \in R_t(\cdot)\}$. 

15
EB3 For each $t \in \{1, 2\}$ and $\delta > 0$, the sequence $\{(T_{tn}(m + v_n + \delta), T_{tn}(m - v_n + \delta))\}_{n=1}^{\infty}$ of suprema satisfies
\[
\sup_{u \in \mathbb{R}} \left| \mathbb{P}(T_{tn}(m + v_n + \delta) \leq u) - \mathbb{P}(T_{tn}(m - v_n + \delta) \leq u) \right| = o(1)
\]
as $n \to \infty$.

Condition EB1 requires that the nuisance parameter $m$, which is the supremum of the support of $|\varepsilon_2 - \varepsilon_1|$ by definition, should be consistently estimated by an estimator $\hat{m}_n$ with convergence rate faster than $v_n$.\(^{17}\) Since the error terms are additively separable, a natural estimator of $m$ is the maximal difference between $Y_{t2}$ and $Y_{t1}$ among individuals who have the explanatory variables and covariates (across two periods) that are close in the Euclidean distance. Consistency of this matching estimator and its convergence rate will be discussed in next subsection. The knowledge that $\hat{m}_n$ converges faster than $v_n$ in Condition EB1 implies that for any $\delta > 0$, the estimated region $\mathcal{R}_t(\hat{m}_n + \delta)$ is bracketed by nonrandom regions $[\mathcal{R}_t(m + v_n + \delta), \mathcal{R}_t(m - v_n + \delta)]$ with probability approaching one; in particular,
\[
\liminf_{n \to \infty} \mathbb{P}(\mathcal{R}_t(m + v_n + \delta) \subseteq \mathcal{R}_t(\hat{m}_n + \delta) \subseteq \mathcal{R}_t(m - v_n + \delta)) \\
\geq \liminf_{n \to \infty} \mathbb{P}(|\hat{m}_n - m| \leq v_n) = 1.
\]

Figure 1.1 illustrates these regions in the absence of covariates. Since $\hat{T}_{tn}^0(\cdot)$ is the supremum of $\hat{\theta}_{tn}^0$ over the region $\mathcal{R}_t(\cdot)$, we also have $\hat{T}_{tn}^0(m + v_n + \delta) \leq \hat{T}_{tn}^0(\hat{m}_n + \delta) \leq \hat{T}_{tn}^0(m - v_n + \delta)$ with probability approaching one. It follows that $\hat{T}_{tn}^0(\hat{m}_n + \delta)$ is asymptotically bracketed by $[\hat{T}_{tn}^0(m + v_n + \delta), \hat{T}_{tn}^0(m - v_n + \delta)]$. This observation permits to disentangle the estimation of regions parameterized by $m$ from the evaluation of the limiting distribution of $\hat{T}_{tn}^0(\hat{m}_n + \delta)$ if $\hat{T}_{tn}^0(m + v_n + \delta)$ and $\hat{T}_{tn}^0(m - v_n + \delta)$ are 
\textit{asymptotically uniformly close} when the nonrandom sequence $\{v_n\}$ shrinks to zero.\(^{18}\) Instead of calculating the limiting distributions of $\hat{T}_{tn}^0(m + v_n + \delta)$ and $\hat{T}_{tn}^0(m - v_n + \delta)$, we adopt the empirical bootstrap method, which is valid if for each $t \in \{1, 2\}$ and $\ell \in \{-1, 1\}$, the distribution of $\hat{T}_{tn}^0(m + \ell v_n + \delta)$ can be well approximated by the conditional distribution of $\hat{T}_{tn}^0(m + \ell v_n + \delta)$ given the data $\mathcal{G}_n$ when $n$ is large. Condition EB2

\(^{17}\) Although the consistency of $\hat{m}_n$ implies the existence of a sequence $\{u_n\}_{n=1}^{\infty}$ with $\mathbb{P}(|\hat{m}_n - m| > u_n) \leq u_n$ and $\lim_{n \to \infty} u_n = 0$, this sequence $\{u_n\}_{n=1}^{\infty}$ may be unknown to researchers; additionally, Condition EB3 may not hold if $\{v_n\}_{n=1}^{\infty}$ is replaced with $\{u_n\}_{n=1}^{\infty}$.

\(^{18}\) Throughout this chapter, two sequences $\{U_{1n}\}_{n=1}^{\infty}$ and $\{U_{2n}\}_{n=1}^{\infty}$ of random variables are called asymptotically uniformly close if their Kolmogorov distance shrinks to zero as $n$ tends to infinity; specifically,
\[
\limsup_{n \to \infty} \sup_{u \in \mathbb{R}} |\mathbb{P}(U_{1n} \leq u) - \mathbb{P}(U_{2n} \leq u)| = 0.
\]
Figure 1.1: Brackets for Regions

Case 1: $x_1 \geq x_2$

Case 2: $x_2 \geq x_1$

Notes: The dashed region is $R_t(m + v_n + \delta)$ whereas the shaded region is $R_t(m - v_n + \delta)$. If the estimate $\hat{m}_n$ is bracketed by $[m - v_n, m + v_n]$, then $R_t(m + v_n + \delta) \subseteq R_t(\hat{m}_n + \delta) \subseteq R_t(m - v_n + \delta)$.

provides the mechanism of coupling between $\hat{T}_{tn}^0$ and $\hat{T}_{tn}^*$ by introducing an intermediate sequence $\{(T_{tn}(m + v_n + \delta), T_{tn}(m - v_n + \delta))\}_{n=1}^{\infty}$ of couplers; each intermediate coupler $T_{tn}(\cdot)$ is the supremum of some tight centered Gaussian process $\mathcal{G}_{tn}$ over the region $R_t(\cdot)$. The twofold couplings could be constructed by the strong approximation and anti-concentration inequality developed in Chernozhukov, Chetverikov, and Kato (2014a, 2016). We postpone the discussion to Appendix 1.B, in which the statistics $\hat{T}_{tn}^0$ and $\hat{T}_{tn}^*$ are constructed by smoothed kernel CDF estimation.

Finally, Condition EB3 makes the suprema $T_{tn}(m + v_n + \delta)$ and $T_{tn}(m - v_n + \delta)$ asymptotically uniformly close. Consequently, Conditions EB2 and EB3 ensure that the supremum $\hat{T}_{tn}^*$ given the data $\mathcal{D}_n$ is asymptotically uniformly close to the supremum $\hat{T}_{tn}^0$ taken over the shrinking bracket of regions $[R_t(m + v_n + \delta), R_t(m - v_n + \delta)]$. Condition EB3 is generally valid if for each $t \in \{1, 2\}$, $\ell \in \{-1, 1\}$, and $\delta > 0$,

$$|T_{tn}(m + \ell v_n + \delta) - T_{tn}(m + \delta)| = O_p(\gamma_n)$$ (1.6)

for some shrinking sequence $\{\gamma_n\}_{n=1}^{\infty}$ with $\gamma_n \mathbb{E}[T_{tn}(m + \delta)] = o(1)$ as $n \to \infty$. To see this, note that the condition (1.6) implies that for any $\eta > 0$, there are constants $c_0 > 0$ and $n_0 \in \mathbb{N}$ such that $\mathbb{P}(|T_{tn}(m + \ell v_n + \delta) - T_{tn}(m + \delta)| > c_0 \gamma_n) < \eta$ for all $n \geq n_0$. 

17
Applying Lemma 2.1 of Chernozhukov et al. (2016) yields that for all $n \geq n_0$,

$$
\sup_{u \in \mathbb{R}} |\mathbb{P}(T_{tn}(m + \ell v_n + \delta) \leq u) - \mathbb{P}(T_{tn}(m + \delta) \leq u)| \\
\leq \sup_{u \in \mathbb{R}} \mathbb{P}(|T_{tn}(m + \delta) - u| \leq c_0 \gamma_n) + \eta.
$$

Following the anti-concentration inequality in Lemma A.1 of Chernozhukov, Chetverikov, and Kato (2014b), we could find constants $c_1$ and $c_2$ such that the last term is bounded by $c_1 \gamma_n \left\{ \mathbb{E} [T_{tn}(m + \delta)] + \sqrt{1 \vee \log (c_2/\gamma_n)} \right\} + \eta$, which is less than $2\eta$ whenever $n$ is sufficiently large.

Theorem 1.1 below establishes that $\hat{T}_{tn}^0(\hat{m}_n + \delta)$ is asymptotically uniformly close to $\hat{T}_{tn}^*(\hat{m}_n + \delta)$ given the data $\mathcal{D}_n$ if Condition EB is satisfied. The proof of Theorem 1.1 involves several steps of approximations discussed above: (i) Condition EB1 implies that $\hat{T}_{tn}^0(\hat{m}_n + \delta)$ is asymptotically bracketed by \left[ \hat{T}_{tn}^0(m + v_n + \delta), \hat{T}_{tn}^0(m - v_n + \delta) \right]$ and that $\hat{T}_{tn}^*(\hat{m}_n + \delta)$ is asymptotically bracketed by \left[ \hat{T}_{tn}^*(m + v_n + \delta), \hat{T}_{tn}^*(m - v_n + \delta) \right]$; (ii) Condition EB2 ensures the coupling between $\hat{T}_{tn}^0$ and $T_{tn}$, and the coupling between $\hat{T}_{tn}^*$ and $T_{tn}$; (iii) The suprema $T_{tn}(m + v_n + \delta)$ and $T_{tn}(m - v_n + \delta)$ are asymptotically uniformly close by Condition EB3.

**Theorem 1.1.** Suppose Condition EB holds. Then for each $t \in \{1, 2\}$ and $\delta > 0$, as $n \to \infty$,

$$
\sup_{u \in \mathbb{R}} \left| \mathbb{P}(\hat{T}_{tn}^0(\hat{m}_n + \delta) \leq u) - \mathbb{P}(\hat{T}_{tn}^*(\hat{m}_n + \delta) \leq u|\mathcal{D}_n) \right| = o_p(1).
$$

As discussed in Subsection 1.3.1, the unknown nuisance parameter $m$ and possible failure of weak convergence of $\hat{Q}_{tn}^0$ (or equivalently, $\hat{Q}_{tn}$ on $\mathcal{R}_t(m)$ under the null hypothesis) make it complicated to do statistical inference based on $\hat{T}_{tn}(\hat{m}_n + \delta)$. When the nuisance parameter $m$ is known, we can establish Theorem 1.1 by imposing only Condition EB2 with $v_n = 0$ and $\delta = 0$ because the coupling between the suprema $\hat{T}_{tn}^0$ and $\hat{T}_{tn}^*$ evaluated over the nonrandom region $\mathcal{R}_t(m)$ is sufficient to validate the empirical bootstrap method.

When the process $\hat{Q}_{tn}^0$ is weakly convergent in $(l^\infty(\mathcal{R}_t(m)), \|\cdot\|_\infty)$, the coupling between the suprema $\hat{T}_{tn}^0$ and $\hat{T}_{tn}^*$ can be constructed through a unique weak limit $T_t(\cdot)$ over a fixed nonrandom bracket of the region $\mathcal{R}_t(m + \delta)$, rather than through a sequence $\{T_{tn}(\cdot)\}_{n=1}^\infty$ of suprema over the shrinking bracket $[\mathcal{R}_t(m + v_n + \delta), \mathcal{R}_t(m - v_n + \delta)]$ in Condition EB2, if the following condition holds:
EB2' For each \( t \in \{1, 2\} \), there exists a tight centered Gaussian process \( \mathbb{G}_t \) in \((L^\infty(\mathcal{R}_t(m)), \|\cdot\|_\infty)\) such that

(i) \( \hat{Q}_{tn}^0 \) converges weakly to \( \mathbb{G}_t \);

(ii) conditional on the data \( \mathcal{R}_n \), \( \hat{Q}_{tn}^\ast \) converges weakly to \( \mathbb{G}_t \);

(iii) \( \mathbb{G}_t \) has continuous paths with respect to the Euclidean distance \( \|\cdot\| \) with probability one;

(iv) the region \( \mathcal{R}_t(m) \) is bounded.

Condition EB2' consists of properties regarding \( \hat{Q}_{tn}^0, \hat{Q}_{tn}^\ast \), and their weak limit \( \mathbb{G}_t \). In the literature on semi-parametric estimation of conditional CDFs, a process of an estimated conditional CDF can usually be approximated by a weakly convergent empirical process with an asymptotically negligible approximation error.\(^{19}\) In this case, the approximation of \( \hat{Q}_{tn}^0 \) by an empirical process \( G_{tn} \) is plausible and the weak convergence of \( G_{tn} \) can be established so that Condition EB2'(i) holds under some regularity conditions. As \( G_{tn} \) converges weakly to \( \mathbb{G}_t \), the empirical bootstrap counterpart \( G_{tn}^\ast \) also converges weakly to \( \mathbb{G}_t \) in probability by Theorem 3.1 of Giné and Zinn (1990). Condition EB2'(ii) is thus ensured as long as the difference between \( \hat{Q}_{tn}^\ast \) and \( G_{tn}^\ast \) is asymptotically negligible.

By Example 1.5.10 of van der Vaart and Wellner (1996), the tightness of \( \mathbb{G}_t \) implies that almost all paths of \( \mathbb{G}_t \) are uniformly continuous on \( \mathcal{R}_t(m) \) with respect to the canonical distance \( \rho_t(u_1, u_2) \equiv \sqrt{\mathbb{E}[(\mathbb{G}_t(u_1) - \mathbb{G}_t(u_2))^2]} \). It follows that Condition EB2'(iii) holds if \( \rho_t(u_1, u_2) \) is continuous with respect to the Euclidean distance \( \|u_1 - u_2\| \).\(^{20}\)

Under Condition EB2', we can apply Berge’s maximum theorem to show that

\[
\lim_{v \to 0} T_t(m + v + \delta) = \lim_{v \to 0} T_t(m - v + \delta) = T_t(m + \delta)
\]

almost surely. It follows from the anti-concentration inequality in Lemma A.1 of Chernozhukov et al. (2014b) that

\[
\limsup_{v \to 0} \sup_{u \in \mathbb{R}} |\mathbb{P}(T_t(m + v + \delta) \leq u) - \mathbb{P}(T_t(m - v + \delta) \leq u)| = 0.
\]

It is thus possible to couple the suprema \( \hat{T}_{tn}^0 \) and \( \hat{T}_{tn}^\ast \) over a fixed nonrandom bracket \([\mathcal{R}_t(m - v_0 + \delta), \mathcal{R}_t(m + v_0 + \delta)]\) of regions for some small \( v_0 > 0 \). To make \( \hat{T}_{tn}^0(\hat{m}_n + \ell v + \delta) \) and \( \hat{T}_{tn}^\ast(\hat{m}_n + \ell v + \delta) \) given the data \( \mathcal{R}_n \) asymptotically uniformly close to \( T_t(m + \ell v + \delta) \), we impose the following technical condition.\(^{21}\)

\(^{19}\) An example is the distribution regression, which is recently studied by Chernozhukov, Fernández-Val, and Melly (2013) and Peracchi and Leorato (2015): see also references cited therein.

\(^{20}\) If the metric space \( (\mathcal{R}_t(m), \rho_t) \) is bounded, then Condition EB2'(iii) implies the continuity of the canonical distance with respect to the Euclidean distance. See the discussion in Section 1.3 of Adler and Taylor (2007).

\(^{21}\) By Tsirelson’s (1976) theorem, the supremum \( T_t(m + \delta) \) is continuously distributed if it is continuous at the left limit of its support because its underlying Gaussian process \( \mathbb{G}_t \) is separable bounded.
EB3’ For each $t \in \{1, 2\}$ and $\delta > 0$, $\mathbb{T}_t(m + \delta) \equiv \sup \{G_t(y_1, y_2, x_1, x_2, w_1, w_2, z) : (y_1, y_2, x_1, x_2, w_1, w_2, z) \in R_t(m + \delta)\}$ is continuously distributed.

We establish the validity of empirical bootstrap method in Theorem 1.2 below.

**Theorem 1.2.** Suppose Conditions EB1, EB2’, and EB3’ hold. Then for each $t \in \{1, 2\}$ and $\delta > 0$, as $n \to \infty$,

$$
\sup_{u \in \mathbb{R}} \left| \mathbb{P} \left( \hat{T}_{tn}^0(\hat{m}_n + \delta) \leq u \right) - \mathbb{P} \left( \hat{T}_{tn}^*(\hat{m}_n + \delta) \leq u | \mathcal{G}_n \right) \right| = o_p(1).
$$

Although semi-parametric estimation of the conditional CDFs, like the distribution regression, usually leads to weak convergence of their estimators, it may be challenging to check whether a semi-parametric specification of $F_{Y_1|X_1, X_2, W_1, W_2, Z}$ is compatible with that of $F_{Y_1|X, W_1, Z}$. In contrast, nonparametric estimation of the conditional CDFs can avoid misspecification. However, scaled nonparametric estimators of the conditional CDFs may not converge weakly; therefore, neither could the associated $\hat{Q}_{tn}^0$. The nonexistence of a weak limit $G_t$ negates the mechanism of coupling the distribution of $\hat{T}_{tn}^0(m + \delta)$ and the conditional distribution of $\hat{T}_{tn}^*(m + \delta)$ in Theorem 1.2. However, as stated in Theorem 1.1, the mechanism of coupling through a sequence $\{G_{tn}\}_{n=1}^{\infty}$ of Gaussian processes is achievable so that the empirical bootstrap method is validated if we have the approximation to both $\hat{T}_{tn}^0$ and $\hat{T}_{tn}^*$ by the supremum $T_{tn}$ of $G_{tn}$.

Although Theorems 1.1 and 1.2 hold even when the null hypothesis (1.2) is not valid, the test statistic $\hat{T}_{tn}^0(\hat{m}_n + \delta)$ is infeasible because the center $F_{Y_1|X_1, W_1, Z} - F_{Y_1|X_1, X_2, W_1, W_2, Z}$ is unknown. Since the null hypothesis (1.2) and Condition EB1 imply

$$
\sup_{u \in \mathbb{R}} \left| \mathbb{P} \left( \hat{T}_n(\hat{m}_n + \delta) \leq u \right) - \mathbb{P} \left( \hat{T}_n^0(\hat{m}_n + \delta) \leq u \right) \right| = o(1)
$$

for each $t \in \{1, 2\}$ and $\delta > 0$, we immediately have the following corollary.

**Corollary 1.2.** Suppose the null hypothesis (1.2) and Condition D1 are valid. Moreover, suppose Condition EB holds. Then for each $t \in \{1, 2\}$ and $\delta > 0$,

$$
\sup_{u \in \mathbb{R}} \left| \mathbb{P} \left( \hat{T}_{tn}(\hat{m}_n + \delta) \leq u \right) - \mathbb{P} \left( \hat{T}_{tn}^*(\hat{m}_n + \delta) \leq u | \mathcal{G}_n \right) \right| = o_p(1),
$$

as $n \to \infty$. $\square$

Corollary 1.2 demonstrates the validity of empirical bootstrap approach. For any $\alpha \in (0, 1)$, $t \in \{1, 2\}$, and $\delta > 0$, let $c_{tn}(\alpha, \delta)$ be the $(1 - \alpha)$-quantile of $\hat{T}_{tn}^*(\hat{m}_n + \delta)$; that is,

$$
c_{tn}(\alpha, \delta) \equiv \inf \left\{ u : \mathbb{P} \left( \hat{T}_{tn}^*(\hat{m}_n + \delta) \leq u | \mathcal{G}_n \right) \geq 1 - \alpha \right\}.
$$
It follows from Corollary 1.2 that \( \mathbb{P}\left( \widehat{T}_{tn}(\hat{m}_n + \delta) \leq c_{tn}(\alpha, \delta) \right) \rightarrow 1 - \alpha \) as \( n \rightarrow \infty \). Consequently, the null hypothesis (1.2) is rejected at an approximate significance level \( \alpha \) if the test statistic \( \widehat{T}_{tn}(\hat{m}_n + \delta) \) is greater than \( c_{tn}(\alpha, \delta) \). Since the shape of the conditional joint CDF is restricted over two nonoverlapping regions by Corollary 1.1, the empirical bootstrap approach is valid for both \( t = 1 \) and \( t = 2 \). Taking this observation into account, we summarize steps of testing the null hypothesis (1.2) based on the empirical bootstrap method in the following algorithm.

**Algorithm 1**

(i) Compute a consistent estimator \( \hat{m}_n \) by the sample \( \mathcal{D}_n \);

(ii) Construct estimators \( \left( \hat{F}_{Y_1|X_1W_1Z}, \hat{F}_{Y_2|X_2W_2Z}, \hat{F}_{Y_1|X_1X_2W_1W_2Z} \right) \) of the conditional CDFs by the sample \( \mathcal{D}_n \);

(iii) Obtain the test statistics \( \widehat{T}_{1n}(\hat{m}_n + \delta_0) \) and \( \widehat{T}_{2n}(\hat{m}_n + \delta_0) \) via (1.4) for some small \( \delta_0 > 0 \);

(iv) Draw the bootstrap data \( \mathcal{D}^{*,b}_n = \{ Y_{i1}^{*,b}, Y_{i2}^{*,b}, X_{i1}^{*,b}, X_{i2}^{*,b}, W_{i1}^{*,b}, W_{i2}^{*,b}, Z_i^{*,b}, i = 1, \ldots, B \} \) of sample size \( n \) from the sample \( \mathcal{D}_n \) with replacement for \( b = 1, \ldots, B \);

(v) Construct bootstrap estimators \( \left( \hat{F}_{Y_1|X_1W_1Z}^{*,b}, \hat{F}_{Y_2|X_2W_2Z}^{*,b}, \hat{F}_{Y_1|X_1X_2W_1W_2Z}^{*,b} \right) \) of the conditional CDFs based on the bootstrap data \( \mathcal{D}^{*,b}_n \) for \( b = 1, \ldots, B \);

(vi) Obtain the bootstrap test statistics \( \widehat{T}_{1n}^{*,b}(\hat{m}_n + \delta_0) \) and \( \widehat{T}_{2n}^{*,b}(\hat{m}_n + \delta_0) \) via (1.5) for \( b = 1, \ldots, B \);

(vii) Set \( c_{tn}^{*}(\alpha/2, \delta_0) \) be the \( (1 - \alpha/2) \)-quantile of \( \left\{ T_{tn}^{*,b}(\hat{m}_n + \delta_0) : b = 1, \ldots, B \right\} \) for \( t = 1, 2 \);

(viii) Reject the null hypothesis (1.2) if \( \widehat{T}_{1n}(\hat{m}_n + \delta_0) > c_{1n}^{*}(\alpha/2, \delta_0) \) or \( \widehat{T}_{2n}(\hat{m}_n + \delta_0) > c_{2n}^{*}(\alpha/2, \delta_0) \).

Testing the null hypothesis (1.2) at an approximate significance level \( \alpha \) based on Algorithm 1 is conservative because Bonferroni’s inequality

\[
\mathbb{P}\left( \widehat{T}_{1n}^0(\hat{m}_n + \delta_0) > c_{1n}^{*}(\alpha/2, \delta_0) \right) \text{ or } \mathbb{P}\left( \widehat{T}_{2n}^0(\hat{m}_n + \delta_0) > c_{2n}^{*}(\alpha/2, \delta_0) \right) \\
\leq \mathbb{P}\left( \overline{\widehat{T}_{1n}^0(\hat{m}_n + \delta_0) > c_{1n}^{*}(\alpha/2, \delta_0)} \right) + \mathbb{P}\left( \overline{\widehat{T}_{2n}^0(\hat{m}_n + \delta_0) > c_{2n}^{*}(\alpha/2, \delta_0)} \right)
\]

is used in the last step to make the statistical decision. To construct a test that is not conservative, we need to adjust the critical values and evaluate the joint probability

\[
\mathbb{P}\left( \overline{\widehat{T}_{1n}(\hat{m}_n + \delta_0) > c_{1n}^{*}(\cdot, \delta_0)} \text{ and } \overline{\widehat{T}_{2n}(\hat{m}_n + \delta_0) > c_{2n}^{*}(\cdot, \delta_0)} \right).
\]
which is neglected by Bonferroni’s inequality.

One approach to evaluating the joint probability is sample splitting. The sample $\mathcal{D}_n$ is randomly partitioned into two subsamples $\mathcal{D}_{n1}$ and $\mathcal{D}_{n2}$ with sample sizes $n_1 \equiv \lfloor n/2 \rfloor$, the largest integer less than or equal to $n/2$, and $n_2 \equiv n - n_1$. For each $t \in \{1, 2\}$, both $\hat{Q}_{tn}$ and $\hat{Q}_{tn}^*$ are constructed only by the subsample $\mathcal{D}_{nt}$. The random sample splitting allows us to calculate the limiting probability

$$\lim_{n \to \infty} \mathbb{P} \left( \hat{T}_{1n}^0 (\hat{m}_n + \delta_0) > c_{1n}^*(\theta, \delta_0) \text{ or } \hat{T}_{2n}^0 (\hat{m}_n + \delta_0) > c_{2n}^*(\theta, \delta_0) \right) = 2\theta - \theta^2$$

for every $\theta \in (0, 1)$ even though a consistent estimator $\hat{m}_n$ is still obtained by the whole sample $\mathcal{D}_n$. Setting $\theta$ with $\theta^2 - 2\theta + \alpha = 0$ gives a non-conservative test with an approximate significance level $\alpha$. Algorithm 2 below describes these steps in detail.

**Algorithm 2**

(i) Compute a consistent estimator $\hat{m}_n$ by the sample $\mathcal{D}_n$;

(ii) Randomly divide the sample $\mathcal{D}_n$ into two nonoverlapping subsamples $\mathcal{D}_{n1}$ and $\mathcal{D}_{n2}$;

(iii) Construct estimators $(\hat{F}_{Y|X,W,Z}^*, \hat{F}_{Y|X_1X_2W,W,Z}^*)$ of the conditional CDFs based on the subsample $\mathcal{D}_{nt}$ for each $t = 1, 2$;

(iv) Obtain the test statistics $\hat{T}_{1n}(\hat{m}_n + \delta_0)$ and $\hat{T}_{2n}(\hat{m}_n + \delta_0)$ via (1.4) for some small $\delta_0 > 0$;

(v) Draw the bootstrap data $\mathcal{D}_{nt}^{*,b}$ of sample size $n_t$ from the sample $\mathcal{D}_{nt}$ with replacement for $b = 1, \ldots, B$ and $t = 1, 2$;

(vi) Construct bootstrap estimators $(\hat{F}_{Y|X,W,Z}^{*,b}, \hat{F}_{Y|X_1X_2W,W,Z}^{*,b})$ of the conditional CDFs based on the bootstrap data $\mathcal{D}_{nt}^{*,b}$ for $b = 1, \ldots, B$ and $t = 1, 2$;

(vii) Obtain the bootstrap test statistics $\hat{T}_{1n}^{*,b}(\hat{m}_n + \delta_0)$ and $\hat{T}_{2n}^{*,b}(\hat{m}_n + \delta_0)$ via (1.5) for $b = 1, \ldots, B$;

(viii) Set $c_{tn}^*(\theta, \delta_0)$ be the $(1 - \theta)$-quantile of $\left\{ \hat{T}_{tn}^{*,b}(\hat{m}_n + \delta_0) : b = 1, \ldots, B \right\}$ for $t = 1, 2$, where $\theta \in (0, 1)$ satisfies $\theta^2 - 2\theta + \alpha = 0$;

(ix) Reject the null hypothesis (1.2) if $\hat{T}_{1n}(\hat{m}_n + \delta_0) > c_{1n}^*(\theta, \delta_0)$ or $\hat{T}_{2n}(\hat{m}_n + \delta_0) > c_{2n}^*(\theta, \delta_0)$.

Algorithm 2 is suggested only when the sample size is large because data splitting may slightly improve the control in size but reduce the power of the proposed test.
1.3.3 Estimation of $m$

For the two-period panel data model in (1.1), the outcome variable $Y_{it}$ is characterized by the structural function $s(X_{it}, W_{it}, Z_i, A_i)$ and the additively separable error term $\varepsilon_{it}$. Pairwise differencing of $Y_{it}$ across two periods for a given individual yields

$$Y_{i2} - Y_{i1} = [s(X_{i2}, W_{i2}, Z_i, A_i) - s(X_{i1}, W_{i1}, Z_i, A_i)] + [\varepsilon_{i2} - \varepsilon_{i1}]$$

so that $|Y_{i2} - Y_{i1}| = |\varepsilon_{i2} - \varepsilon_{i1}|$ whenever $(X_{i2}, W_{i2}) = (X_{i1}, W_{i1})$. Since $m$ is the supremum of the support of $|\varepsilon_{2} - \varepsilon_{1}|$, a feasible estimator of $m$ is the maximum of $|Y_{i2} - Y_{i1}|$ for individuals who have $\| (X_{i2}, W_{i2}) - (X_{i1}, W_{i1}) \|$ in a shrinking neighborhood of $0 \in \mathbb{R}^{(1+dW)}$. Concretely, a natural matching estimator of $m$ is

$$\hat{m}_n \equiv \max_{i=1, \ldots, n} \left\{ 1_{\| (X_{i2}, W_{i2}) - (X_{i1}, W_{i1}) \| \leq \tau_n} |Y_{i2} - Y_{i1}| \right\}; \tag{1.7}$$

where $\{\tau_n\}_{n=1}^\infty$ is a sequence of matching bandwidth that shrinks to zero as $n$ increases.

To establish the consistency of $\hat{m}_n$ with the rate of convergence $\tau_n$, we impose the following conditions.

**Condition [M]:**

M1 The matching bandwidth satisfies $\tau_n \to 0$ and $n\tau_n^{(2+dW)} \to c_0 \in (0, \infty]$ as $n \to \infty$.

M2 For all $(z, a) \in Za$, $s(\cdot, \cdot, z, a)$ is Lipschitz continuous. In other words, there exists a constant $c_L \geq 0$ such that $|s(x_2, w_2, z, a) - s(x_1, w_1, z, a)| \leq c_L \| (x_2, w_2) - (x_1, w_1) \|$ for all $(x_1, x_2, w_1, w_2, z, a) \in X^2W^2ZA$.

M3 For all $z \in Z$, the conditional density of $|\varepsilon_{2} - \varepsilon_{1}|$ given $Z = z$ is bounded away from 0 on $[0, m]$.

M4 In some neighborhood of 0, $\Delta \equiv (X_2 - X_1, W_2 - W_1) \in \mathbb{R}^{(1+dW)}$ is continuously distributed. In addition, the joint density function $f_\Delta(\cdot)$ is continuous at 0 and satisfies $f_\Delta(0) > 0$.

**Proposition 1.4.** Suppose $(X_1, X_2, W_1, W_2)$ is continuously distributed. Under Conditions D1-D2 and M1-M4, $\hat{m}_n$ is consistent for $m$ and its convergence rate is $\hat{m}_n - m = O_p(\tau_n)$.

**Remarks**

1. When the sequence $\{\tau_n\}_{n=1}^\infty$ of matching bandwidth satisfies $n\tau_n^{(2+dW)} \to c_0 \in (0, \infty)$, Proposition 1.4 implies that the convergence rate of the matching estimator
\( \hat{m}_n \) is \( O_p\left(n^{-1/2}\right) \) in the absence of covariates but is slower than \( O_p\left(n^{-1/2}\right) \) in the presence of time-varying covariates. In the case that consistent estimation of \( m \) is enough, the matching estimator \( \hat{m}_n \) is also consistent if we adopt faster shrinking matching bandwidth \( \tau_n \) with \( \tau_n \to 0 \) and \( n\tau_n^{(1+d_w)} \to \infty \).

2. Since the matching bandwidth \( \tau_n \) is selected by researchers, Condition EB1 is valid for the matching estimator \( \hat{m}_n \) with \( v_n = \tau_n^{(1-\eta)} \) for any \( \eta > 0 \) whenever the matching bandwidth \( \tau_n \) satisfies Condition M1. In addition, as pointed out in previous subsections, the suprema \( \hat{T}_{tn} \) and \( \hat{T}_{tn}^* \) are taken over the region \( \mathcal{R}_t(\hat{m}_n + \delta_0) \) for some small \( \delta_0 > 0 \). With the knowledge of convergence rate of \( \hat{m}_n \), we can replace \( \delta_0 \) with a shrinking sequence \( \{\delta_n\}_{n=1}^{\infty} \), consider another matching estimator \( \bar{m}_n \equiv \hat{m}_n + \delta_n \), and take the suprema \( \bar{T}_{tn} \) and \( \bar{T}_{tn}^* \) over the region \( \mathcal{R}_t(\bar{m}_n) \). If the sequence \( \{\delta_n\}_{n=1}^{\infty} \) satisfies \( \delta_n \geq v_n \) and \( \delta_n \to 0 \), then Proposition 1.4 implies not only that the revised matching estimator \( \bar{m}_n \) is consistent for \( m \) but also that \( \bar{m}_n \) is greater than \( m \) with probability approaching one. In this case, \( \mathcal{R}_t(\bar{m}_n) \subseteq \mathcal{R}_t(m) \) with probability approaching one. Despite this appealing property, the optimal or data driven choice of this auxiliary sequence \( \{\delta_n\}_{n=1}^{\infty} \) is unknown in practice.

3. When \( \triangle = (X_2 - X_1, W_2 - W_1) \) is continuously distributed, the matching mechanism is valid because the structural function is smooth. Condition M2 is, however, not essential for consistency of \( \hat{m}_n \). We can replace Lipschitz continuity with Hölder continuity for \( s(\cdot, \cdot, z, a) \); specifically, there exist constants \( c_H \geq 0 \) and \( \beta \in (0, 1] \) such that \( |s(x_2, w_2, z, a) - s(x_1, w_1, z, a)| \leq c_H \| (x_2, w_2) - (x_1, w_1) \|^\beta \) for all \( (x_1, x_2, w_1, w_2, z, a) \in \mathcal{X} \mathcal{W} \mathcal{Z} \mathcal{A} \). Conditions M1 and M3-M4, together with \( \beta \)-Hölder continuity, imply that \( \hat{m}_n - m = O_p(\tau_n^{\beta}) \).

4. When the vector \( (X_1, X_2, W_1, W_2) \) is discrete and \( \mathbb{P}\left( (X_2, W_2) = (X_1, W_1) \right) > 0 \), the matching estimator \( \hat{m}_n \) could achieve a faster convergence rate and this rate is invariant to the dimension of covariates. Specifically, if Condition M3 holds, then there is a constant \( c_d > 0 \) such that for any \( \eta > 0 \), \( \limsup_{n \to \infty} \mathbb{P}(|\hat{m}_n - m| > \eta \tau_n) \leq \exp\{-c_d\eta\} \) provided \( n\tau_n \) converges to some positive constant as \( n \to \infty \). It follows that \( \hat{m}_n - m = O_p(n^{-1}) \). Condition M2 is relaxed in this case because the matching mechanism can rely on only individuals who have the explanatory variables and covariates that are unchanged across two periods.

5. Proposition 1.4 and the remarks above still hold if the null hypothesis (1.2) is invalid. Indeed, Condition M may hold even when a structural function is in the alternatives. Accordingly, the consistency and convergence rate of the matching
estimator \( \hat{m}_n \) can be of independent interest.

### 1.4 Monte Carlo experiments

To evaluate the finite-sample performance of the supremum test statistics, we report the results of a simulation study in this section. The simulation designs below are chosen to examine the performance of the proposed test and are not meant to mimic any data set in empirical studies; additionally, we only consider simulation designs without covariates \((W_1, W_2, Z)\) for simplicity.

For the simulations that follow, we implement Algorithm 1 with sample sizes \( n = 250, 500, \text{ and } 1000 \), number of bootstrap replications \( B = 600 \), and the number of simulation replications \( S = 500 \). The significance level (i.e., the nominal level) is set at \( \alpha = 10\% \). Moreover, we calculate the matching estimator \( \hat{m}_n \) by (1.7) with the matching bandwidth \( \tau_n = n^{-1/2} \) satisfying the regularity conditions in Subsection 1.3.3 and construct the Nadaraya-Watson smoothed kernel estimators \((\hat{F}_{Y_1|X_1}, \hat{F}_{Y_2|X_2}, \hat{F}_{Y_1Y_2|X_1X_1})\) of the conditional CDFs:

\[
\hat{F}_{Y_1Y_2|X_1X_2}(y_1, y_2|x_1, x_2) = \frac{\sum_{i=1}^{n} \psi \left( \frac{y_1-Y_{i1}}{h_n} \right) \psi \left( \frac{y_2-Y_{i2}}{h_n} \right) k \left( \frac{x_1-x_1}{h_n} \right) k \left( \frac{x_2-x_2}{h_n} \right)}{\sum_{i=1}^{n} k \left( \frac{x_1-x_1}{h_n} \right) k \left( \frac{x_2-x_2}{h_n} \right)}
\]

and for each \( t \in \{1, 2\}, \)

\[
\hat{F}_{Y_t|X_1}(y_t|x_t) = \frac{\sum_{i=1}^{n} \psi \left( \frac{y_t-Y_{i1}}{b_n} \right) k \left( \frac{x_t-x_t}{b_n} \right)}{\sum_{i=1}^{n} k \left( \frac{x_t-x_t}{b_n} \right)}
\]

where \( k(u) = 0.75(1 - u^2)1_{|u|<1} \) is the Epanechnikov kernel, \( \psi(u) = \int_{-\infty}^{u} k(v) \; dv \), and \( \{b_n\}_{n=1}^{\infty} \) and \( \{h_n\}_{n=1}^{\infty} \) are sequences of bandwidth with \( h_n = n^{-1/5} \) and \( b_n = h_n^2 \).

Similarly, their bootstrap counterparts \((\hat{F}_{Y_1|X_1}^*, \hat{F}_{Y_2|X_2}^*, \hat{F}_{Y_1Y_2|X_1X_1}^*)\) are constructed by replacing the sample \( \mathcal{D}_n \) with bootstrap data \( \mathcal{D}_n^* \). Finally, we compute the test statistic \( \hat{T}_{tn}(\hat{m}_n + \delta_0) \) by (1.4) and the bootstrap counterpart \( \hat{T}_{tn}(\hat{m}_n + \delta_0) \) by (1.5) for \( \delta_0 = 0.02 \) and \( \varphi_n = (nh_n^2)^{1/2} \). We avoid boundary effects due to nonparametric kernel estimation by taking suprema, \( \hat{T}_{tn}(\hat{m}_n + \delta_0) \) and \( \hat{T}_{tn}(\hat{m}_n + \delta_0) \), over a subset \( \mathcal{Y}_{trim}^2 \mathcal{X}_{trim}^2 \cap \mathcal{R}_4(\hat{m}_n + \delta_0) \). The trimmed set \( \mathcal{X}_{trim} \) is selected to be \([0.05, 0.95]\) for \( X \) having the uniform distribution on \([0,1]\), and the trimmed set \( \mathcal{Y}_{trim} \) is an interval determined by the 5% and 95% sample quantile of \( Y \). This trimming should not affect the size of the proposed test, but only the

---

22 The choice of \((h_n, b_n)\) satisfies the regularity conditions in Appendix 1.B but is somewhat arbitrary; however, the issue on optimal or data driven bandwidth is out of scope of this chapter.
power. The computation of suprema is based on the SQP and GlobalSearch algorithms in MATLAB®, in which we set the number of start points equal to 500.

First, we examine the ability of the proposed test to detect departures from the null hypothesis (1.2) and consider the following set of data generating processes:

DGP1 $Y_{it} = 0.5(A_i - \bar{a})X_{it} + \varepsilon_{it}$, $i = 1, 2, \ldots, n$, and $t = 1, 2$ where $X_{it} \sim \text{Unif}(0, 1)$, $A_i \sim \text{Beta}(2, 2)$, $\varepsilon_{it} \sim \text{Unif}(0, 0.5)$, $\bar{a} \in \{0, 1, 2\}$, and $(X_{i1}, X_{i2}, A_i, \varepsilon_{i1}, \varepsilon_{i2})$ are mutually independent.

In DGP1, the parameter $\bar{a}$ measures the degree of departures in certain direction. If $\bar{a} = 0$, the structural function $s(x, a) = 0.25 + 0.5ax$ is strictly increasing in $x$ for all $a \in (0, 1)$ and hence in the null hypothesis; if $\bar{a} > 0$, the structural function $s(x, a) = 0.25 + 0.5(a - \bar{a})x$ is strictly decreasing in $x$ for all $a \in (0, \bar{a})$ and represents a departure from the null hypothesis (1.2).

Table 1.1 reports the rejection rates of the proposed test for DGP1 with $\bar{a} = 0$. When the nuisance parameter $m$ is known, the test seems conservative because each rejection rate is less than the nominal level. When the nuisance parameter is estimated by $\hat{m}_n$, the rejection rates are slightly larger than those based on $m$ being known. This phenomenon arises because the matching estimator $\hat{m}_n$ has downward bias for $m$, as illustrated in the left panel of Figure 1.2. In more than a half of simulation replications, the estimate $\hat{m}_n + \delta_0$ is still less than $m$, and the suprema $\hat{T}_{tn}$ and $\hat{T}^*_tn$ are taken over $\mathcal{R}_t(\hat{m}_n + \delta_0)$ containing $\mathcal{R}_t(m)$. Replacing $\delta_0 = 0.02$ with $\delta_0' = 0.06$ makes the rejection rates less than the nominal level. This finding suggests that larger values of $\delta_0$ may lessen the over-rejection due to the downward bias of $\hat{m}_n$ for $m$; however, larger values of $\delta_0$ are expected to decrease the power of the proposed test. Therefore, we still adopt $\delta_0 = 0.02$ in the following simulations.

Table 1.2 shows how the rejection rates are affected by $\bar{a}$. As expected, the rejection rate $ceteris paribus$ increases in $\bar{a}$ because the structural function is farther away from the null hypothesis for a larger value of $\bar{a}$. When $\bar{a} \geq 1$, the structural function $s(x, a) = 0.25 + 0.5(a - \bar{a})x$ is strictly decreasing in $x$ with probability $\mathbb{P}(A < 1) = 1$. Since the structural function with $\bar{a} = 1$ is in $S^c_m$; the power of the proposed test is not guaranteed by Proposition 1.2; however, the simulation results show that each rejection rate is larger than the nominal level for $n = 1000$. When $\bar{a} = 2$, the structural function is in $S_m$ and the rejection rates are almost equal to one. Furthermore, $\hat{m}_n$ in these

---

23 Condition M3 does not hold for the error terms in DGP1. Despite the downward bias of $\hat{m}_n$, the simulation results show that the estimator $\hat{m}_n$ seems consistent and the proposed test with $\hat{m}_n + \delta_0$ under the null hypothesis (1.2) would still perform well as long as $\delta_0$ is large enough.

24 If $\bar{a} = 1$, $s(x', a) - s(x, a) \leq 0.5$ for any $(x, x', a) \in (0, 1)^2 \times (0, \bar{a})$.

25 Taking $(x, x', a) = (0.9, 0.1, 0.6)$ yields $s(x', a) - s(x, a) = 0.56 > 0.5$. 

26
<table>
<thead>
<tr>
<th></th>
<th>First Region</th>
<th>Second Region</th>
<th>Both Regions</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>m is known†</strong></td>
<td>$\hat{T}<em>{1n}(m+\delta_0)&gt;c</em>{1n}^*(5%,\delta_0)$</td>
<td>$\hat{T}<em>{2n}(m+\delta_0)&gt;c</em>{2n}^*(5%,\delta_0)$</td>
<td>$\hat{T}<em>{1n}(m+\delta_0)&gt;c</em>{1n}^*(5%,\delta_0)$</td>
</tr>
<tr>
<td>$m = 0.5^a$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$n=250$</td>
<td>0.020</td>
<td>0.018</td>
<td>0.038</td>
</tr>
<tr>
<td>$n=500$</td>
<td>0.022</td>
<td>0.032</td>
<td>0.054</td>
</tr>
<tr>
<td>$n=1000$</td>
<td>0.032</td>
<td>0.034</td>
<td>0.066</td>
</tr>
<tr>
<td><strong>m is unknown‡</strong></td>
<td>$\hat{T}_{1n}(\hat{m}<em>n+\delta_0)&gt;c</em>{1n}^*(5%,\delta_0)$</td>
<td>$\hat{T}_{2n}(\hat{m}<em>n+\delta_0)&gt;c</em>{2n}^*(5%,\delta_0)$</td>
<td>$\hat{T}_{1n}(\hat{m}<em>n+\delta_0)&gt;c</em>{1n}^*(5%,\delta_0)$</td>
</tr>
<tr>
<td>$m = 0.5^a$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$n=250$</td>
<td>0.048</td>
<td>0.052</td>
<td>0.098</td>
</tr>
<tr>
<td>$n=500$</td>
<td>0.058</td>
<td>0.062</td>
<td>0.116</td>
</tr>
<tr>
<td>$n=1000$</td>
<td>0.062</td>
<td>0.070</td>
<td>0.132</td>
</tr>
<tr>
<td><strong>m is unknown§</strong></td>
<td>$\hat{T}_{1n}(\hat{m}<em>n+\delta_1)&gt;c</em>{1n}^*(5%,\delta'_0)$</td>
<td>$\hat{T}_{2n}(\hat{m}<em>n+\delta_0)&gt;c</em>{2n}^*(5%,\delta'_0)$</td>
<td>$\hat{T}_{1n}(\hat{m}<em>n+\delta_1)&gt;c</em>{1n}^*(5%,\delta'_0)$</td>
</tr>
<tr>
<td>$m = 0.5^a$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$n=250$</td>
<td>0.034</td>
<td>0.036</td>
<td>0.068</td>
</tr>
<tr>
<td>$n=500$</td>
<td>0.038</td>
<td>0.046</td>
<td>0.084</td>
</tr>
<tr>
<td>$n=1000$</td>
<td>0.040</td>
<td>0.050</td>
<td>0.090</td>
</tr>
</tbody>
</table>

† For each $t \in \{1, 2\}$, $c_{tn}^*(5\%,\delta_0)$ is the 95% quantile of $\hat{T}_{tn}^*(m+\delta_0)$.
‡ For each $t \in \{1, 2\}$, $c_{tn}^*(5\%,\delta_0)$ is the 95% quantile of $\hat{T}_{tn}^*(\hat{m}_n+\delta_0)$.
§ For each $t \in \{1, 2\}$, $c_{tn}^*(5\%,\delta'_0)$ is the 95% quantile of $\hat{T}_{tn}^*(\hat{m}_n+\delta'_0)$.
a The trimmed set is $Y_{trim}^2, X_{trim}^2 = [0.10, 0.66]^2 \times [0.05, 0.95]^2$. 

a The trimmed set is $Y_{trim}^2, X_{trim}^2 = [0.10, 0.66]^2 \times [0.05, 0.95]^2$. 

27
Table 1.2: Rejection Rates for DGP1 with $\bar{a} \in \{0, 1, 2\}$

<table>
<thead>
<tr>
<th>$m$ is known†</th>
<th>First Region</th>
<th>Second Region</th>
<th>Both Regions</th>
</tr>
</thead>
<tbody>
<tr>
<td>$m$ is known†</td>
<td>$\hat{T}<em>{1n}(m+\delta_0) &gt; c</em>{1n}^*(5%, \delta_0)$</td>
<td>$\hat{T}<em>{2n}(m+\delta_0) &gt; c</em>{2n}^*(5%, \delta_0)$</td>
<td>$\hat{T}<em>{1n}(m+\delta_0) &gt; c</em>{1n}^<em>(5%, \delta_0)$ or $\hat{T}<em>{2n}(m+\delta_0) &gt; c</em>{2n}^</em>(5%, \delta_0)$</td>
</tr>
<tr>
<td>$\bar{a} = 0^a$</td>
<td>$n=250$</td>
<td>0.020</td>
<td>0.018</td>
</tr>
<tr>
<td>$\bar{a} = 0^a$</td>
<td>$n=500$</td>
<td>0.022</td>
<td>0.032</td>
</tr>
<tr>
<td>$\bar{a} = 0^a$</td>
<td>$n=1000$</td>
<td>0.032</td>
<td>0.034</td>
</tr>
<tr>
<td>$\bar{a} = 1^b$</td>
<td>$n=250$</td>
<td>0.052</td>
<td>0.048</td>
</tr>
<tr>
<td>$\bar{a} = 1^b$</td>
<td>$n=500$</td>
<td>0.074</td>
<td>0.066</td>
</tr>
<tr>
<td>$\bar{a} = 1^b$</td>
<td>$n=1000$</td>
<td>0.108</td>
<td>0.126</td>
</tr>
<tr>
<td>$\bar{a} = 2^c$</td>
<td>$n=250$</td>
<td>0.984</td>
<td>0.992</td>
</tr>
<tr>
<td>$\bar{a} = 2^c$</td>
<td>$n=500$</td>
<td>0.990</td>
<td>1.000</td>
</tr>
<tr>
<td>$\bar{a} = 2^c$</td>
<td>$n=1000$</td>
<td>0.994</td>
<td>0.994</td>
</tr>
</tbody>
</table>

| $m$ is unknown† | $\hat{T}_{1n}(\hat{m}_n+\delta_0) > c_{1n}^*(5\%, \delta_0)$ | $\hat{T}_{2n}(\hat{m}_n+\delta_0) > c_{2n}^*(5\%, \delta_0)$ | $\hat{T}_{1n}(\hat{m}_n+\delta_0) > c_{1n}^*(5\%, \delta_0)$ or $\hat{T}_{2n}(\hat{m}_n+\delta_0) > c_{2n}^*(5\%, \delta_0)$ |
| $\bar{a} = 0^a$ | $n=250$ | 0.048 | 0.052 | 0.098 |
| $\bar{a} = 0^a$ | $n=500$ | 0.058 | 0.062 | 0.116 |
| $\bar{a} = 0^a$ | $n=1000$ | 0.062 | 0.070 | 0.132 |
| $\bar{a} = 1^b$ | $n=250$ | 0.166 | 0.202 | 0.320 |
| $\bar{a} = 1^b$ | $n=500$ | 0.218 | 0.250 | 0.400 |
| $\bar{a} = 1^b$ | $n=1000$ | 0.258 | 0.272 | 0.430 |
| $\bar{a} = 2^c$ | $n=250$ | 0.998 | 0.996 | 1.000 |
| $\bar{a} = 2^c$ | $n=500$ | 0.990 | 0.998 | 1.000 |
| $\bar{a} = 2^c$ | $n=1000$ | 0.996 | 1.000 | 1.000 |

† For each $t \in \{1, 2\}$, $c_{tn}^*(5\%, \delta_0)$ is the 95% quantile of $\hat{T}_{tn}^*(m+\delta_0)$.

†† For each $t \in \{1, 2\}$, $c_{tn}^*(5\%, \delta_0)$ is the 95% quantile of $\hat{T}_{tn}^*(\hat{m}_n+\delta_0)$.

a The trimmed set is $Y_{\text{trim}}^2 X_{\text{trim}}^2 = [0.10, 0.66]^2 \times [0.05, 0.95]^2$.

b The trimmed set is $Y_{\text{trim}}^2 X_{\text{trim}}^2 = [-0.16, 0.40]^2 \times [0.05, 0.95]^2$.

c The trimmed set is $Y_{\text{trim}}^2 X_{\text{trim}}^2 = [-0.58, 0.31]^2 \times [0.05, 0.95]^2$. 

28
(a) DGP1 with $\bar{a} = 0$ and $m = 0.5$

(b) DGP2 with $m = 0.2$

Figure 1.2: Histogram of $\sqrt{n}(\hat{m}_n - m)$ for 500 simulation replications of DGP1 and DGP2 designs generally has downward bias for $m$. In comparison to the rejection rates based on $m$ being known, the rejection rates based on $\hat{m}_n$ tend to be larger.

Next, we investigate whether the proposed test can detect structural functions with regression monotonicity but without structural monotonicity. We consider the following set of data generating processes:

\[
DGP2 \quad Y_{it} = (A_i - \log 2)X_{it} + \varepsilon_{it}, \quad i = 1, 2, \ldots, n, \text{ and } t = 1, 2 \text{ where } X_{it} \sim \text{Unif}(0, 1), A_i \sim \text{Exp}(1), \varepsilon_{it} \sim \text{Unif}(0, m), \quad m \in \{0.2, 0.45, 0.7\}, \text{ and } (X_{i1}, X_{i2}, A_i, \varepsilon_{i1}, \varepsilon_{i2}) \text{ are mutually independent.}
\]

Clearly, this design is Example 1 in the introduction with $X_{it}$ uniformly distributed on $[0, 1]$ and $\varepsilon_{it}$ uniformly distributed on $[0, m]$. The nuisance parameter $m$ represents the maximal difference between errors across two periods; moreover, it captures the volatility of noise relative to that of explanatory variables because $\text{Var}(\varepsilon_{it})/\text{Var}(X_{it}) = m^2$. We compare the power of the proposed test with that of tests for regression monotonicity. In particular, we consider the supremum test statistic of Ghosal et al. (2000) (GSV) and the plug-in adaptive test statistic of Chetverikov (2017) (PI-A). Since two-period panel data are available, the GSV and PI-A statistics are both calculated based on data for each period. Regarding the GSV statistics, we take the supremum over the region $[0.05, 0.95]$ with the aforementioned Epanechnikov kernel and the bandwidth $h_{G}$ = $0.5n^{-1/5}$, as suggested in Ghosal et al. (2000). Following the instructions of Chetverikov (2017), we compute the PI-A statistics by the basic set of weighting functions (with $k = 0$ in his Equation (7)), the aforementioned Epanechnikov kernel, the
bandwidth \( h_{n,a}^{\text{PI-A}} = n^{-1/5} \), and the local Rice estimator based on the recommendations in his Comment 4.1; additionally, the critical values are simulated according to \( B = 600 \) bootstrap replications.

Table 1.3 presents the rejection rates of the proposed test for DGP2 with \( m = 0.2 \). The rejection rates for GSV and PI-A statistics are close to the nominal level for \( n = 1000 \). This result is in accordance with expectation because the regression function \( r(x) = 0.1 + (1 - \log 2)x \) is strictly increasing in \( x \). As discussed in the introduction, the structural function \( s(x,a) = 0.1 + (a - \log 2)x \) is, however, strictly decreasing for a half of individuals. Since the structural function is in \( \mathbb{S}_m \), Proposition 1.2 suggests that the proposed test should have power to detect this departure from the null hypothesis (1.2). Evidence is provided by the simulation results. The rejection rates of the proposed test, whether \( m \) is known or estimated, are reasonably high for \( n = 500 \) and close to one for \( n = 1000 \). Although the right panel of Figure 1.2 still shows the downward bias of \( \hat{m}_n \), the estimation of the nuisance parameter \( m \) in DGP2 has minor effects on the rejection rates.\(^{27} \)

As shown in Table 1.4, whether \( m \) is known or estimated by \( \hat{m}_n \), the rejection rate \( \text{ceteris paribus} \) is smaller for larger \( m \). For \( n = 1000 \), the test has rejection rates close to one for \( m = 0.2 \), above the nominal level for \( m = 0.45 \), and close to the nominal level for \( m = 0.7 \). These results confirm the traditional wisdom: making statistical inference from noisy data would be difficult. Furthermore, since the rejection rates for DGP2 with \( m = 0.7 \) are close to the nominal level, the proposed test may not have power against this alternative in \( \mathbb{S}_m^c \).\(^{28} \)

Finally, we replace the uniform error terms with the two-sided truncated and centered Gaussian error terms, but keep the maximal magnitude of \( |\varepsilon_2 - \varepsilon_1| \) unchanged in each design above. The simulation results in Appendix 1.C suggest that the proposed test is robust to the location shift and choice of distributions of error terms.

### 1.5 Conclusion

We have proposed a test for structural monotonicity in the presence of nonseparable time-invariant heterogeneity, which is possibly infinitely dimensional, when two-period panel data are available. The proposed test can be used as a specification test for structural

---

\(^{26} \) Taking \((x,x',a) = (0.8, 0.2, (\log 2)/3)\) yields \( s(x',a) - s(x,a) \approx 0.2773 > 0.2 \).

\(^{27} \) Condition M3 does not hold for the error terms in DGP1. The structural function in DGP2 neither satisfies Condition M2, which is not valid because the time-invariant heterogeneity \( A \) is exponentially distributed on \((0, \infty)\). However, the estimator \( \hat{m}_n \) seems consistent and the proposed test with \( \hat{m}_n \) would still be able to detect departures from the null hypothesis (1.2), as suggested by the simulation results.

\(^{28} \) For any \((x,x',a) \in (0, 1)^2 \times (0, \log 2)\), \( s(x',a) - s(x,a) \leq \log 2 < 0.7 \).
Table 1.3: Rejection Rates for DGP2 with $m = 0.2$

<table>
<thead>
<tr>
<th></th>
<th>First Period</th>
<th>Second Period</th>
<th>Both Periods</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>GSV$^a$</strong></td>
<td>$\hat{T}<em>{1n}^{GSV} &gt; c</em>{1n}^{GSV}(5%)$</td>
<td>$\hat{T}<em>{2n}^{GSV} &gt; c</em>{2n}^{GSV}(5%)$</td>
<td>$\hat{T}<em>{1n}^{GSV} &gt; c</em>{1n}^{GSV}(5%)$ or $\hat{T}<em>{2n}^{GSV} &gt; c</em>{2n}^{GSV}(5%)$</td>
</tr>
<tr>
<td>$n=250$</td>
<td>0.068</td>
<td>0.064</td>
<td>0.132</td>
</tr>
<tr>
<td>$n=500$</td>
<td>0.070</td>
<td>0.046</td>
<td>0.110</td>
</tr>
<tr>
<td>$n=1000$</td>
<td>0.050</td>
<td>0.060</td>
<td>0.108</td>
</tr>
<tr>
<td><strong>PI-A$^b$</strong></td>
<td>$\hat{T}<em>{1n}^{PI-A} &gt; c</em>{1n}^{PI-A}(5%)$</td>
<td>$\hat{T}<em>{2n}^{PI-A} &gt; c</em>{2n}^{PI-A}(5%)$</td>
<td>$\hat{T}<em>{1n}^{PI-A} &gt; c</em>{1n}^{PI-A}(5%)$ or $\hat{T}<em>{2n}^{PI-A} &gt; c</em>{2n}^{PI-A}(5%)$</td>
</tr>
<tr>
<td>$n=250$</td>
<td>0.112</td>
<td>0.120</td>
<td>0.226</td>
</tr>
<tr>
<td>$n=500$</td>
<td>0.078</td>
<td>0.080</td>
<td>0.152</td>
</tr>
<tr>
<td>$n=1000$</td>
<td>0.052</td>
<td>0.032</td>
<td>0.082</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>First Region</th>
<th>Second Region</th>
<th>Both Regions</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>SM$^c$</strong></td>
<td>$\tilde{T}<em>{1n}(m+\delta_0) &gt; c</em>{1n}^{**}(5%,\delta_0)$</td>
<td>$\tilde{T}<em>{2n}(m+\delta_0) &gt; c</em>{2n}^{**}(5%,\delta_0)$</td>
<td>$\tilde{T}<em>{1n}(m+\delta_0) &gt; c</em>{1n}^{<strong>}(5%,\delta_0)$ or $\tilde{T}<em>{2n}(m+\delta_0) &gt; c</em>{2n}^{</strong>}(5%,\delta_0)$</td>
</tr>
<tr>
<td>$m = 0.2$ is known$^+$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$n=250$</td>
<td>0.442</td>
<td>0.484</td>
<td>0.664</td>
</tr>
<tr>
<td>$n=500$</td>
<td>0.696</td>
<td>0.652</td>
<td>0.858</td>
</tr>
<tr>
<td>$n=1000$</td>
<td>0.808</td>
<td>0.808</td>
<td>0.954</td>
</tr>
<tr>
<td><strong>SM$^c$</strong></td>
<td>$\tilde{T}_{1n}(\hat{m}<em>n+\delta_0) &gt; c</em>{1n}^{**}(5%,\delta_0)$</td>
<td>$\tilde{T}_{2n}(\hat{m}<em>n+\delta_0) &gt; c</em>{2n}^{**}(5%,\delta_0)$</td>
<td>$\tilde{T}<em>{1n}(\hat{m}<em>n+\delta_0) &gt; c</em>{1n}^{**}(5%,\delta_0)$ or $\tilde{T}</em>{2n}(\hat{m}<em>n+\delta_0) &gt; c</em>{2n}^{**}(5%,\delta_0)$</td>
</tr>
<tr>
<td>$m = 0.2$ is unknown$^+$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$n=250$</td>
<td>0.476</td>
<td>0.474</td>
<td>0.676</td>
</tr>
<tr>
<td>$n=500$</td>
<td>0.666</td>
<td>0.690</td>
<td>0.860</td>
</tr>
<tr>
<td>$n=1000$</td>
<td>0.772</td>
<td>0.782</td>
<td>0.940</td>
</tr>
</tbody>
</table>

---

$^a$ The supremum test statistics $\hat{T}_{1n}^{GSV}$ and $\hat{T}_{2n}^{GSV}$ of Ghosal et al. (2000) are calculated based on the first-period and second-period data, respectively.

$^b$ The plug-in adaptive test statistics $\hat{T}_{1n}^{PI-A}$ and $\hat{T}_{2n}^{PI-A}$ of Chetverikov (2017) are calculated based on the first-period and second-period data, respectively.

$^c$ The trimmed set is $Y_{trim}^2, X_{trim}^2 = [-0.34, 1.37]^2 \times [0.05, 0.95]^2$.

$^+$ For each $t \in \{1, 2\}$, $c_{tn}^{*}(5\%,\delta_0)$ is the 95% quantile of $\tilde{T}_{tn}(m+\delta_0)$.

$^\dagger$ For each $t \in \{1, 2\}$, $c_{tn}^{**}(5\%,\delta_0)$ is the 95% quantile of $\tilde{T}_{tn}(\hat{m}_n+\delta_0)$. 
Table 1.4: Rejection Rates for DGP2 with \( m \in \{0.2, 0.45, 0.7\} \)

<table>
<thead>
<tr>
<th>( m ) is known</th>
<th>First Region</th>
<th>Second Region</th>
<th>Both Regions</th>
</tr>
</thead>
<tbody>
<tr>
<td>( m = 0.2^a )</td>
<td>( \tilde{T}<em>{1n}(m+\delta_0) &gt; c</em>{1n}^*(5%, \delta_0) )</td>
<td>( \tilde{T}<em>{2n}(m+\delta_0) &gt; c</em>{2n}^*(5%, \delta_0) )</td>
<td>( \tilde{T}<em>{1n}(m+\delta_0) &gt; c</em>{1n}^<em>(5%, \delta_0) ) or ( \tilde{T}<em>{2n}(m+\delta_0) &gt; c</em>{2n}^</em>(5%, \delta_0) )</td>
</tr>
<tr>
<td>( n=250 )</td>
<td>0.442</td>
<td>0.484</td>
<td>0.664</td>
</tr>
<tr>
<td>( n=500 )</td>
<td>0.696</td>
<td>0.652</td>
<td>0.858</td>
</tr>
<tr>
<td>( n=1000 )</td>
<td>0.808</td>
<td>0.808</td>
<td>0.954</td>
</tr>
<tr>
<td>( m = 0.45^b )</td>
<td>( \tilde{T}<em>{1n}(m+\delta_0) &gt; \tilde{T}</em>{2n}(m+\delta_0) )</td>
<td>( \tilde{T}_{1n}(\hat{m}<em>n+\delta_0) &gt; c</em>{1n}^*(5%, \delta_0) )</td>
<td>( \tilde{T}<em>{1n}(m+\delta_0) &gt; c</em>{1n}^<em>(5%, \delta_0) ) or ( \tilde{T}<em>{2n}(m+\delta_0) &gt; c</em>{2n}^</em>(5%, \delta_0) )</td>
</tr>
<tr>
<td>( n=250 )</td>
<td>0.084</td>
<td>0.078</td>
<td>0.148</td>
</tr>
<tr>
<td>( n=500 )</td>
<td>0.092</td>
<td>0.068</td>
<td>0.152</td>
</tr>
<tr>
<td>( n=1000 )</td>
<td>0.112</td>
<td>0.082</td>
<td>0.186</td>
</tr>
<tr>
<td>( m = 0.7^c )</td>
<td>( \tilde{T}<em>{1n}(m+\delta_0) &gt; c</em>{1n}^*(5%, \delta_0) )</td>
<td>( \tilde{T}<em>{2n}(m+\delta_0) &gt; c</em>{2n}^*(5%, \delta_0) )</td>
<td>( \tilde{T}<em>{1n}(m+\delta_0) &gt; c</em>{1n}^<em>(5%, \delta_0) ) or ( \tilde{T}<em>{2n}(m+\delta_0) &gt; c</em>{2n}^</em>(5%, \delta_0) )</td>
</tr>
<tr>
<td>( n=250 )</td>
<td>0.054</td>
<td>0.036</td>
<td>0.088</td>
</tr>
<tr>
<td>( n=500 )</td>
<td>0.082</td>
<td>0.046</td>
<td>0.124</td>
</tr>
<tr>
<td>( n=1000 )</td>
<td>0.070</td>
<td>0.040</td>
<td>0.102</td>
</tr>
<tr>
<td>( m = 0.2^a )</td>
<td>( \tilde{T}_{1n}(\hat{m}<em>n+\delta_0) &gt; c</em>{1n}^*(5%, \delta_0) )</td>
<td>( \tilde{T}_{2n}(\hat{m}<em>n+\delta_0) &gt; c</em>{2n}^*(5%, \delta_0) )</td>
<td>( \tilde{T}<em>{1n}(m+\delta_0) &gt; c</em>{1n}^<em>(5%, \delta_0) ) or ( \tilde{T}<em>{2n}(m+\delta_0) &gt; c</em>{2n}^</em>(5%, \delta_0) )</td>
</tr>
<tr>
<td>( n=250 )</td>
<td>0.476</td>
<td>0.474</td>
<td>0.676</td>
</tr>
<tr>
<td>( n=500 )</td>
<td>0.666</td>
<td>0.690</td>
<td>0.860</td>
</tr>
<tr>
<td>( n=1000 )</td>
<td>0.772</td>
<td>0.782</td>
<td>0.940</td>
</tr>
<tr>
<td>( m = 0.45^b )</td>
<td>( \tilde{T}_{1n}(\hat{m}<em>n+\delta_0) &gt; \tilde{T}</em>{2n}(\hat{m}_n+\delta_0) )</td>
<td>( \tilde{T}_{1n}(\hat{m}<em>n+\delta_0) &gt; c</em>{2n}^*(5%, \delta_0) )</td>
<td>( \tilde{T}<em>{1n}(\hat{m}<em>n+\delta_0) &gt; c</em>{1n}^*(5%, \delta_0) ) or ( \tilde{T}</em>{2n}(m+\delta_0) &gt; c_{2n}^*(5%, \delta_0) )</td>
</tr>
<tr>
<td>( n=250 )</td>
<td>0.104</td>
<td>0.086</td>
<td>0.182</td>
</tr>
<tr>
<td>( n=500 )</td>
<td>0.124</td>
<td>0.096</td>
<td>0.204</td>
</tr>
<tr>
<td>( n=1000 )</td>
<td>0.144</td>
<td>0.112</td>
<td>0.238</td>
</tr>
<tr>
<td>( m = 0.7^c )</td>
<td>( \tilde{T}_{1n}(\hat{m}<em>n+\delta_0) &gt; c</em>{1n}^*(5%, \delta_0) )</td>
<td>( \tilde{T}_{2n}(\hat{m}<em>n+\delta_0) &gt; c</em>{2n}^*(5%, \delta_0) )</td>
<td>( \tilde{T}<em>{1n}(\hat{m}<em>n+\delta_0) &gt; c</em>{1n}^*(5%, \delta_0) ) or ( \tilde{T}</em>{2n}(m+\delta_0) &gt; c_{2n}^*(5%, \delta_0) )</td>
</tr>
<tr>
<td>( n=250 )</td>
<td>0.050</td>
<td>0.060</td>
<td>0.108</td>
</tr>
<tr>
<td>( n=500 )</td>
<td>0.070</td>
<td>0.036</td>
<td>0.106</td>
</tr>
<tr>
<td>( n=1000 )</td>
<td>0.072</td>
<td>0.038</td>
<td>0.110</td>
</tr>
</tbody>
</table>

\(^\dagger\) For each \( t \in \{1, 2\} \), \( c_{tn}^*(5\%, \delta_0) \) is the 95% quantile of \( \tilde{T}_{tn}(m+\delta_0) \).

\(^\ddagger\) For each \( t \in \{1, 2\} \), \( c_{tn}^*(5\%, \delta_0) \) is the 95% quantile of \( \tilde{T}_{tn}(\hat{m}_n+\delta_0) \).

\(a\) The trimmed set is \( Y_{2 \text{ trim}}^2 X_{2 \text{ trim}}^2 = [-0.34, 1.37]^2 \times [0.05, 0.95]^2 \).

\(b\) The trimmed set is \( Y_{2 \text{ trim}}^2 X_{2 \text{ trim}}^2 = [-0.25, 1.51]^2 \times [0.05, 0.95]^2 \).

\(c\) The trimmed set is \( Y_{2 \text{ trim}}^2 X_{2 \text{ trim}}^2 = [-0.18, 1.65]^2 \times [0.05, 0.95]^2 \).
monotonicity implied by economic theory; additionally, it allows us to investigate the sign of individual treatment effect, which is different from summary statistics in the literature on treatment effects. Instead of identifying any structural objects, our approach has been developed based on the implied shape constraints on the conditional joint CDF. As is noted in the introduction, this approach can be easily extended to test the convexity of a structural function in the explanatory variable given any nonseparable time-invariant unobserved heterogeneity if three-period panel data are available.

There are some future directions for this research. First, practitioners may be interested in the optimal or data dependent choices of matching bandwidth and smoothing bandwidth. Second, although our simulation results show the rejection rates increase as structural functions are farther away from the null hypothesis, the local power of the proposed test is still an interesting question. Finally, it would be desirable to develop a test for structural monotonicity when multi-period panel data are available.

Appendix 1.A Technical proofs

1.A.1 Proof of Proposition 1.1

Proof. Fix $t \in \{1, 2\}$. Let $t' \equiv 3 - t \in \{1, 2\}$. For all $(y_1, y_2, x_1, x_2, w_1, w_2, z) \in \mathcal{X}^2 \mathcal{X} \mathcal{W}^2 \mathcal{Z},$

$$F_{y_1|x_1,x_2,w_1,w_2,z}(y_1|x_1,x_2, w_1, w_2, z) - F_{y_1,y_2|x_1,x_2,w_1,w_2,z}(y_1, y_2|x_1, x_2, w_1, w_2, z)$$

$$= \mathbb{P}(s(x_t, w_t, z, A) + \varepsilon_t < y_t, s(x_{t'}, w_{t'}, z, A) + \varepsilon_{t'} > y_{t'} | X_1 = x_1, X_2 = x_2, W_1 = w_1, W_2 = w_2, Z = z).$$

(1.A.1)

It is clear that the last term is greater than or equal to zero.

Suppose $(y_1, y_2, x_1, x_2, w_1, w_2, z) \in \mathcal{R}_t(m)$. The null hypothesis implies that with probability one, $s(x_{t'}, w_{t'}, z, A) \leq s(x_t, w_t, z, A)$ because $x_t \geq x_{t'}, w_t = w_{t'},$ and $y_{t'} - y_t \geq m$. The equivalence between $F_{y_1,y_2|x_1,x_2,w_1,w_2,z}$ and $F_{y_1|x_1,w_1,z}$ over $\mathcal{R}_t(m)$ follows by showing

$$\mathbb{P}(s(x_t, w_t, z, A) + \varepsilon_t \leq y_t, s(x_{t'}, w_{t'}, z, A) + \varepsilon_{t'} > y_{t'} | X_1 = x_1, X_2 = x_2, W_1 = w_1, W_2 = w_2, Z = z)$$

$$\leq \mathbb{P}(y_{t'} - \varepsilon_{t'} < s(x_t, w_t, z, A) \leq y_t - \varepsilon_t | X_1 = x_1, X_2 = x_2, W_1 = w_1, W_2 = w_2, Z = z)$$

$$\leq \mathbb{P}(m \leq y_{t'} - y_t < |\varepsilon_{t'} - \varepsilon_t| | X_1 = x_1, X_2 = x_2, W_1 = w_1, W_2 = w_2, Z = z)$$

$$= 0.$$
where the last inequality holds because \( m \geq |\varepsilon_2 - \varepsilon_1| \) with probability one by definition. \( \square \)

1.A.2 Proof of Corollary 1.1

**Proof.** It suffices to show that \( F_{Y_i|X_i,W_i,Z} = F_{Y_i|X_i,X_2,W_i,W_2} \) under Condition D1. Fix \( t \in\{1, 2\} \). Let \( t' \equiv 3 - t \in\{1, 2\} \). Condition D1 implies \((X_t, W_t) \perp (A, \varepsilon_t) | (X_t, W_t, Z)\) by Lemma 4.2 of Dawid (1979). It follows that

\[
F_{Y_i|X_i,X_2,W_i,W_2}(y_t|x_1, x_2, w_1, w_2, z)
= \mathbb{P}(s(x_t, w_t, z, A) + \varepsilon_t \leq y_t|X_1 = x_1, X_2 = x_2, W_1 = w_1, W_2 = w_2, Z = z)
= \mathbb{P}(s(x_t, w_t, z, A) + \varepsilon_t \leq y_t|X_t = x_t, W_t = w_t, Z = z)
= F_{Y_i|X_i,W_i,Z}(y_t|x_t, w_t, z)
\]

for all \((y_t, x_1, x_2, w_1, w_2, z) \in \mathcal{Y}^2\mathcal{W}^2\mathcal{Z}\). \( \square \)

1.A.3 Proof of Proposition 1.2

**Proof.** Suppose the structural function \( s \) is in \( \mathbb{S}_m \). There exists \((x_1, x_2, w, z, \bar{a}) \in \mathcal{X}^2\mathcal{W}\mathcal{Z}\mathcal{A}\) with \( x_1 > x_2 \) such that \( s(x_2, w, z, \bar{a}) - s(x_1, w, z, \bar{a}) > m \) and \( \mathbb{P}(A \in \mathcal{N}|Z = z) > 0 \) for all neighborhood \( \mathcal{N} \) of \( \bar{a} \). Taking two linearly interpolated points \( y_1 \equiv (1 - \vartheta)s(x_1, w, z, \bar{a}) + \vartheta s(x_2, w, z, \bar{a}) \) and \( y_2 \equiv (1 - \vartheta)s(x_2, w, z, \bar{a}) + \vartheta s(x_1, w, z, \bar{a}) \) where \( \vartheta \equiv \frac{1}{2} \left(1 - \frac{m}{s(x_2, w, z, \bar{a}) - s(x_1, w, z, \bar{a})}\right) \in \left(0, \frac{1}{2}\right) \), we have

\[
y_2 - y_1 = (1 - 2\vartheta) \left[ s(x_2, w, z, \bar{a}) - s(x_1, w, z, \bar{a}) \right] = m.
\]

Condition D4(i) implies that both \( s(x_1, w, z, \bar{a}) \) and \( s(x_2, w, z, \bar{a}) \) are in the support \( \mathcal{Y} \) of \( Y \); hence, \( y_1 \) and \( y_2 \) are in \( \mathcal{Y} \) by the convexity of \( \mathcal{Y} \).

By Condition D3, the continuity of \( s(x_t, w, z, \cdot) \) implies the existence of a neighborhood \( \mathcal{N}_1 \) of \( \bar{a} \) such that

\[
\max_{t=1,2} |s(x_t, w, z, a) - s(x_t, w, z, \bar{a})| < \vartheta \left[ s(x_2, w, z, \bar{a}) - s(x_1, w, z, \bar{a}) \right] \equiv \triangle_s
\]

whenever \( a \in \mathcal{N}_1 \). By Condition D4(ii), there is another neighborhood \( \mathcal{N}_2 \) of \( \bar{a} \) such that \( \mathbb{P}(\varepsilon_1 < 0, \varepsilon_2 > 0 | A = a, Z = z) > 0 \) for every \( a \in \mathcal{N}_2 \). Let \( \mathcal{N}_0 \equiv \mathcal{N}_1 \cap \mathcal{N}_2 \). For
any \( a \in \mathcal{M}_0 \), we have

\[
\Delta_s = -\Delta_s + \theta [s(x_2, w, z, \bar{a}) - s(x_1, w, z, \bar{a})] \\
< [s(x_1, w, z, \bar{a}) - s(x_1, w, z, a)] + \theta [s(x_2, w, z, \bar{a}) - s(x_1, w, z, \bar{a})] \\
= y_1 - s(x_1, w, z, a)
\]

and similarly \( y_2 - s(x_2, w, z, a) < -\Delta_s \). Since Condition D1 implies

\[
\mathbb{P}(s(x_t, w_t, z, A) + \varepsilon_t \leq y_t, \\
\quad s(x'_{t'}, w'_{t'}, z, A) + \varepsilon_{t'} > y_{t'} \mid X_1 = x_1, X_2 = x_2, W_1 = w_1, W_2 = w_2, Z = z) \\
= \mathbb{P}(s(x_t, w_t, z, A) + \varepsilon_t \leq y_t, s(x'_{t'}, w'_{t'}, z, A) + \varepsilon_{t'} > y_{t'} \mid Z = z),
\]

Equality (1.A.1) reduces to

\[
F_{Y_1|X_1W_1Z}(y_1|x_1, w_1, z) - F_{Y_1Y_2|X_1X_2W_1W_2Z}(y_1, y_2|x_1, x_2, w_1, w_2, Z = z) \\
= \mathbb{P}(s(x_t, w_t, z, A) + \varepsilon_t \leq y_t, s(x'_{t'}, w'_{t'}, z, A) + \varepsilon_{t'} > y_{t'} \mid Z = z) (1.A.2)
\]

for all \((y_1, y_2, x_1, x_2, w_1, w_2, z) \in \mathcal{Y}^2 \mathcal{X}^2 \mathcal{W}^2 \mathcal{Z} \). It follows that

\[
F_{Y_1|X_1W_1Z}(y_1|x_1, w, z) - F_{Y_1Y_2|X_1X_2W_1W_2Z}(y_1, y_2|x_1, x_2, w, w, z) \\
= \mathbb{P}(s(x_t, w, z, A) + \varepsilon_1 \leq y_1, s(x_2, w, z, A) + \varepsilon_2 > y_2 \mid Z = z) \\
\geq \int_{\mathcal{M}_0} \mathbb{P}(\varepsilon_1 \leq \Delta_s, \varepsilon_2 > -\Delta_s \mid A = a, Z = z) \text{ d}F_{A|Z}(a \mid z) \\
\geq \int_{\mathcal{M}_0} \mathbb{P}(\varepsilon_1 < 0, \varepsilon_2 > 0 \mid A = a, Z = z) \text{ d}F_{A|Z}(a \mid z).
\]

Note that \( \mathbb{P}(\varepsilon_1 < 0, \varepsilon_2 > 0 \mid A = a, Z = z) > 0 \) for every \( a \in \mathcal{M}_0 \) and \( \mathbb{P}(A \in \mathcal{M}_0 \mid Z = z) > 0 \) because \( \bar{a} \in \mathcal{A}^z \). Following arguments analogous to those in Corollary 4.10 of Bartle (1966), we obtain

\[
F_{Y_1|X_1W_1Z}(y_1|x_1, w, z) - F_{Y_1Y_2|X_1X_2W_1W_2Z}(y_1, y_2|x_1, x_2, w, w, z) \\
\geq \int_{\mathcal{M}_0} \mathbb{P}(\varepsilon_1 < 0, \varepsilon_2 > 0 \mid A = a, Z = z) \text{ d}F_{A|Z}(a \mid z) \\
> 0.
\]

Consequently, the shape constraints in Corollary 1.1 do not hold.
Therefore, the shape constraints in Corollary 1.1 do not hold.

**Proof.** Suppose the structural function is in $S_{2m}$. There exists $(x_1, x_2, w, z, \bar{a}) \in X^2WZA$ with $x_1 > x_2$ such that $\bar{x} \equiv s(x_2, w, z, \bar{a}) - s(x_1, w, z, \bar{a}) - 2m > 0$ and $\mathbb{P}(A \in \mathcal{N} | Z = y) > 0$ for all neighborhood $\mathcal{N}$ of $\bar{a}$. Let $\bar{\epsilon} \equiv \sup \mathcal{E}$ and $\underline{\epsilon} \equiv \inf \mathcal{E}$. By Conditions D2 and D4$''$(ii), $m \geq \bar{\epsilon} - \underline{\epsilon} \in (0, \infty)$. Let $y_1 \equiv \bar{\epsilon} + [\vartheta s(x_1, w, z, \bar{a}) + (1 - \vartheta)s(x_2, w, z, \bar{a})]$, and $y_2 \equiv \underline{\epsilon} + [\vartheta s(x_2, w, z, \bar{a}) + (1 - \vartheta)s(x_1, w, z, \bar{a})]$ for some $\vartheta \in \left(\frac{4m + 2\bar{x}}{4m + 2\bar{x}}, 1\right)$. Conditions D4$''$(i) and D4$''$(iii) imply that both $y_1$ and $y_2$ are in $\mathcal{Y}$. Since,

$$\begin{align*}
y_2 - y_1 &= -m + (2\vartheta - 1)[s(x_2, w, z, \bar{a}) - s(x_1, w, z, \bar{a})] \\
&= -m + (2\vartheta - 1)[2m + \bar{x}] \\
&= m + \vartheta(4m + 2\bar{x}) - (4m + \bar{x}) \\
&> m,
\end{align*}$$

we obtain $(y_1, y_2, x_1, x_2, w, w, z) \in R_1(m)$.

In addition, the continuity of $s(x_t, w, z, \cdot)$ at $\bar{a}$ implies the existence of a neighborhood $\mathcal{N}_{\bar{a}}$ of $\bar{a}$ such that

$$s(x_1, w, z, a) < \vartheta s(x_1, w, z, \bar{a}) + (1 - \vartheta)s(x_2, w, z, \bar{a})$$
and
$$s(x_2, w, z, a) > \vartheta s(x_2, w, z, \bar{a}) + (1 - \vartheta)s(x_1, w, z, \bar{a})$$
for all $a \in \mathcal{N}_{\bar{a}}$. It follows from Equality (1.A.2) that

$$\begin{align*}
F_{Y_1|X,W,Z}(y_1|x_1, w, z) - F_{Y_1|X,W,Z}(y_1, y_2|x_1, x_2, w, w, z) \\
&= \mathbb{P}(s(x_1, w, z, A) + \varepsilon_1 \leq y_1 \text{ and } s(x_2, w, z, A) + \varepsilon_2 > y_2 | Z = z) \\
&\geq \mathbb{P}(s(x_1, w, z, A) \leq y_1 - \bar{\epsilon} \text{ and } s(x_2, w, z, A) > y_2 - \underline{\epsilon} | Z = z) \\
&= \mathbb{P}
\left(s(x_1, w, z, A) < \vartheta s(x_1, w, z, \bar{a}) + (1 - \vartheta)s(x_2, w, z, \bar{a})
\right.
\left.\text{ and } s(x_2, w, z, A) > \vartheta s(x_2, w, z, \bar{a}) + (1 - \vartheta)s(x_1, w, z, \bar{a}) | Z = z \right)
\geq \mathbb{P}(A \in \mathcal{N}_{\bar{a}} | Z = z)
> 0.
\end{align*}$$

Therefore, the shape constraints in Corollary 1.1 do not hold. \hfill \Box

**1.A.5 Proof of Theorem 1.1**

**Proof.** Fix $t \in \{1, 2\}$ and $\delta > 0$. There is an integer $n_1 \in \mathbb{N}$ such that $v_n < \delta$ for all $n \geq n_1$. Let $u \in \mathbb{R}$ and consider $n \geq n_1$. Note that $\mathbb{P}(\hat{R}_{tn}^0(\cdot) \leq u)$ is weakly increasing because $\mathcal{R}_t(q') \subseteq \mathcal{R}_t(q)$ when $q' \geq q$. Similarly, with probability one,
\( \mathbb{P}(\hat{T}_{tn}(\cdot) \leq u \mid D_n) \) is weakly increasing on the set \( \{ |\hat{m}_n - m| \leq v_n \} \). It follows that on the set \( \{ |\hat{m}_n - m| \leq v_n \} \),

\[
\mathbb{P}(\hat{T}_{tn}^0(\hat{m}_n + \delta) \leq u) - \mathbb{P}(\hat{T}_{tn}(\hat{m}_n + \delta) \leq u \mid D_n) \\
\leq \mathbb{P}(\hat{T}_{tn}(m + v_n + \delta) \leq u) - \mathbb{P}(\hat{T}_{tn}(m - v_n + \delta) \leq u \mid D_n) + \mathbb{P}(\hat{m}_n - m > v_n) \\
\leq \mathbb{P}(\hat{T}_{tn}(m + v_n + \delta) \leq u) - \mathbb{P}(\hat{T}_{tn}(m - v_n + \delta) \leq u) + \mathbb{P}(\hat{m}_n - m > v_n) \\
+ \sup_{u \in \mathbb{R}} \left| \mathbb{P}(\hat{T}_{tn}^0(m + v_n + \delta) \leq u) - \mathbb{P}(\hat{T}_{tn}(m + v_n + \delta) \leq u) \right| \\
+ \sup_{u \in \mathbb{R}} \left| \mathbb{P}(\hat{T}_{tn}^0(m - v_n + \delta) \leq u) - \mathbb{P}(\hat{T}_{tn}(m - v_n + \delta) \leq u) \right|. 
\tag{1.A.3}
\]

Similarly, on the set \( \{ |\hat{m}_n - m| \leq v_n \} \),

\[
\mathbb{P}(\hat{T}_{tn}(\hat{m}_n + \delta) \leq u \mid D_n) - \mathbb{P}(\hat{T}_{tn}(\hat{m}_n + \delta) \leq u) \\
\leq \mathbb{P}(\hat{T}_{tn}(m + v_n + \delta) \leq u) - \mathbb{P}(\hat{T}_{tn}(m - v_n + \delta) \leq u) + \mathbb{P}(\hat{m}_n - m > v_n) \\
+ \sup_{u \in \mathbb{R}} \left| \mathbb{P}(\hat{T}_{tn}^0(m + v_n + \delta) \leq u) - \mathbb{P}(\hat{T}_{tn}(m + v_n + \delta) \leq u) \right| \\
+ \sup_{u \in \mathbb{R}} \left| \mathbb{P}(\hat{T}_{tn}^0(m - v_n + \delta) \leq u) - \mathbb{P}(\hat{T}_{tn}(m - v_n + \delta) \leq u) \right|. 
\tag{1.A.4}
\]

Combining (1.A.3) and (1.A.4), we obtain that on the set \( \{ |\hat{m}_n - m| \leq v_n \} \),

\[
\sup_{u \in \mathbb{R}} \left| \mathbb{P}(\hat{T}_{tn}^0(\hat{m}_n + \delta) \leq u) - \mathbb{P}(\hat{T}_{tn}(\hat{m}_n + \delta) \leq u \mid D_n) \right| \\
\leq \sup_{u \in \mathbb{R}} \left| \mathbb{P}(\hat{T}_{tn}(m + v_n + \delta) \leq u) - \mathbb{P}(\hat{T}_{tn}(m - v_n + \delta) \leq u) \right| + \mathbb{P}(\hat{m}_n - m > v_n) \\
+ \sup_{u \in \mathbb{R}} \left| \mathbb{P}(\hat{T}_{tn}^0(m + v_n + \delta) \leq u) - \mathbb{P}(\hat{T}_{tn}(m + v_n + \delta) \leq u) \right| \\
+ \sup_{u \in \mathbb{R}} \left| \mathbb{P}(\hat{T}_{tn}^0(m - v_n + \delta) \leq u) - \mathbb{P}(\hat{T}_{tn}(m - v_n + \delta) \leq u) \right| \\
+ \sup_{u \in \mathbb{R}} \left| \mathbb{P}(\hat{T}_{tn}^0(m + v_n + \delta) \leq u \mid D_n) - \mathbb{P}(\hat{T}_{tn}(m + v_n + \delta) \leq u) \right| \\
+ \sup_{u \in \mathbb{R}} \left| \mathbb{P}(\hat{T}_{tn}^0(m - v_n + \delta) \leq u \mid D_n) - \mathbb{P}(\hat{T}_{tn}(m - v_n + \delta) \leq u) \right|. 
\tag{1.A.5}
\]
It follows from Inequality (1.A.5) that for any \( \eta > 0 \),

\[
\mathbb{P} \left( \sup_{u \in \mathbb{R}} \left| \mathbb{P}(\hat{T}_{tn}^0(m + \delta) \leq u) - \mathbb{P}(\hat{T}_{tn}^*(m + \delta) \leq u \mid \mathcal{D}_n) \right| > \eta \right) \\
\leq \mathbb{P} \left( \sup_{u \in \mathbb{R}} \left| \mathbb{P}(\hat{T}_{tn}^0(m + \delta) \leq u) - \mathbb{P}(\hat{T}_{tn}^*(m + \delta) \leq u \mid \mathcal{D}_n) \right| > \eta \text{ and } |\hat{m}_n - m| \leq v_n \right) \\
+ \mathbb{P} \left( |\hat{m}_n - m| > v_n \right) \\
\leq 1 \sup_{u \in \mathbb{R}} \mathbb{P}(\hat{T}_{tn}(m + v_n + \delta) \leq u) - \mathbb{P}(\hat{T}_{tn}(m - v_n + \delta) \leq u) > \eta/6 \] \\
+ 1 \mathbb{P}(\hat{m}_n - m > v_n) + \mathbb{P} \left( |\hat{m}_n - m| > v_n \right) \\
+ 1 \sup_{u \in \mathbb{R}} \mathbb{P}(\hat{T}_{tn}^0(m + v_n + \delta) \leq u) - \mathbb{P}(\hat{T}_{tn}^*(m + v_n + \delta) \leq u) > \eta/6 \] \\
+ 1 \sup_{u \in \mathbb{R}} \mathbb{P}(\hat{T}_{tn}^0(m - v_n + \delta) \leq u) - \mathbb{P}(\hat{T}_{tn}^*(m - v_n + \delta) \leq u) > \eta/6 \] \\
+ \mathbb{P} \left( \sup_{u \in \mathbb{R}} \mathbb{P}(\hat{T}_{tn}^*(m - v_n + \delta) \leq u \mid \mathcal{D}_n) - \mathbb{P}(\hat{T}_{tn}(m - v_n + \delta) \leq u) > \eta/6 \right) \\
+ \mathbb{P} \left( \sup_{u \in \mathbb{R}} \mathbb{P}(\hat{T}_{tn}^*(m + v_n + \delta) \leq u \mid \mathcal{D}_n) - \mathbb{P}(\hat{T}_{tn}(m + v_n + \delta) \leq u) > \eta/6 \right). \tag{1.A.6}
\]

Condition EB ensures that this upper bound converges to zero as \( n \) tends to \( \infty \). \( \square \)

### 1.A.6 Proof of Theorem 1.2

**Proof.** Fix \( t \in \{-1, 1\} \) and \( \delta > 0 \). Since the region \( \mathcal{R}_t(m + \delta) \) is compact and the centered Gaussian process \( \mathcal{G}_t \) has continuous paths with respect to the Euclidean distance with probability one by Condition EB2'(iii), the supremum \( \mathbb{T}_t(m + \delta) \) is finite with probability one and \( \mathbb{E}[\mathbb{T}_t(m + \delta)] \) is also finite by Theorems 2.1.1 and 2.1.2 of Adler and Taylor (2007). In addition, for each \( t \in \{1, 2\} \) and \( \ell \in \{-1, 1\} \), the correspondence \( \mathbb{V}_{t, \ell} : [0, \delta] \rightarrow \mathcal{R}_t(m) \) by \( \mathbb{V}_{t, \ell}(v) \equiv \mathcal{R}_t(m + \ell v + \delta) \) is compact-valued and continuous at \( v = 0 \) by Lemma 1.1. Berge’s maximum theorem implies that for each \( \ell \in \{-1, 1\} \),

\[
\lim_{v \rightarrow 0} \mathbb{T}_t(m + \ell v + \delta) \equiv \mathbb{T}_t(m + \delta)
\]

almost surely. It follows that for any \( r_1 > 0 \) and \( r_2 > 0 \), there exists a constant \( v_0 = v_0(r_1, r_2) < \delta \) such that

\[
\max_{\ell \in \{-1, 1\}} \mathbb{P}(|\mathbb{T}_t(m + \ell v_0 + \delta) - \mathbb{T}_t(m + \delta)| > r_1) \leq r_2.
\]
Lemma 2.1 of Chernozhukov et al. (2016) and Lemma A.1 of Chernozhukov et al. (2014b) imply that
\[
\sup_{u \in \mathbb{R}} |\mathbb{P}(T_t(m + \nu_0 + \delta) \leq u) - \mathbb{P}(T_t(m - \nu_0 + \delta) \leq u)|
\leq 2 \sup_{u \in \mathbb{R}} \mathbb{P}(|T_t(m + \delta) - u| \leq r_1) + 2r_2
\leq C_1 r_1 \left\{ \mathbb{E}[T_t(m + \delta)] + \sqrt{1 + \log (C_2/r_1)} \right\} + 2r_2
\]
f\text{or some constants } C_1 \text{ and } C_2.

Fix } \eta > 0. \text{ We can set } \bar{\nu}_0 = \nu_0(\bar{r}_1, \bar{r}_2) \text{ for some sufficiently small constants } \bar{r}_1 \text{ and } \bar{r}_2 \text{ such that }
\sup_{u \in \mathbb{R}} |\mathbb{P}(T_t(m + \nu_0 + \delta) \leq u) - \mathbb{P}(T_t(m - \nu_0 + \delta) \leq u)| < \eta/6.

Replacing } \nu_n \text{ and } T_{tn} \text{ with } \bar{\nu}_0 \text{ and } T_t, \text{ respectively, in (1.A.6) yields that }
\mathbb{P}\left( \sup_{u \in \mathbb{R}} \mathbb{P}\left( \hat{T}^0_{tn}(\hat{m}_n + \delta) \leq u \right) - \mathbb{P}(\hat{T}^*_tn(\hat{m}_n + \delta) \leq u \mid \mathcal{D}_n) \right) > \eta
\leq 1 \mathbb{P}(|\hat{m}_n - m| > \bar{\nu}_0) + \mathbb{P}(|\hat{m}_n - m| > \bar{\nu}_0)
+ 1_{\sup_{u \in \mathbb{R}} \mathbb{P}(\hat{T}^0_{tn}(m + \bar{\nu}_0 + \delta) \leq u) - \mathbb{P}(T_t(m + \bar{\nu}_0 + \delta) \leq u) > \eta/6}
+ 1_{\sup_{u \in \mathbb{R}} \mathbb{P}(\hat{T}^0_{tn}(m - \bar{\nu}_0 + \delta) \leq u) - \mathbb{P}(T_t(m - \bar{\nu}_0 + \delta) \leq u) > \eta/6}
+ \mathbb{P}\left( \sup_{u \in \mathbb{R}} \mathbb{P}\left( \hat{T}^*_tn(m - \bar{\nu}_0 + \delta) \leq u \mid \mathcal{D}_n \right) - \mathbb{P}(T_t(m - \bar{\nu}_0 + \delta) \leq u) > \eta/6 \right)
+ \mathbb{P}\left( \sup_{u \in \mathbb{R}} \mathbb{P}\left( \hat{T}^*_tn(m + \bar{\nu}_0 + \delta) \leq u \mid \mathcal{D}_n \right) - \mathbb{P}(T_t(m + \bar{\nu}_0 + \delta) \leq u) > \eta/6 \right).

By Condition EB1, the first two terms go to zero as } n \text{ tends to infinity. In addition, Conditions EB2' and EB3', together with the continuous mapping theorem and Pólya's lemma, imply the last four terms also go to zero as } n \text{ tends to infinity.}

\textbf{Lemma 1.1.} \textit{Let } m \text{ and } \delta \text{ be positive real numbers. Suppose the regions } \mathcal{R}_1(m) \text{ and } \mathcal{R}_2(m) \text{ are bounded. For each } t \in \{1, 2\} \text{ and } \ell \in \{-1, 1\}, \text{ the correspondence } \mathbb{V}_{t,\ell} : [0, \delta] \rightarrow \mathcal{R}_t(m) \text{ by } \mathbb{V}_{t,\ell}(v) \equiv \mathcal{R}_t(m + \ell v + \delta) \text{ is compact-valued and continuous at } v = 0.}

\textbf{Proof.} \text{We only prove the results for } \mathbb{V}_{1,\ell} \text{ because the results for } \mathbb{V}_{2,\ell} \text{ can be established by similar arguments. Fix } \ell \in \{-1, 1\}. \text{ For each } v \in [0, \delta], \mathbb{V}_{1,\ell}(v) = \mathcal{R}_1(m + \ell v + \delta) \text{ is compact because it is a closed subset of the bounded set } \mathcal{R}_1(m). \text{ Consequently, } \mathbb{V}_{1,\ell} \text{ is compact-valued.}

\text{To prove the continuity of } \mathbb{V}_{t,\ell} \text{ at } v = 0, \text{ we first show the upper hemicontinuity. Consider sequences } \{v^j\}_{j=1}^\infty \text{ and } \{(y^j_1, y^j_2, x^j_1, x^j_2, w^j_1, w^j_2, z^j)\}_{j=1}^\infty \text{ with } v^j \in [0, \delta]
and \((y^j_1, y^j_2, x^j_1, x^j_2, w^j_1, w^j_2, z^j) \in \mathcal{R}_1(m + \ell v^j + \delta) = \mathcal{V}_{1, \ell}(v^j)\) for each \(j\). By Bolzano-Weierstrass theorem, there exists a subsequence \(\{(y^j_{1k}, y^j_{2k}, x^j_{1k}, x^j_{2k}, w^j_{1k}, w^j_{2k}, z^j_k)\}_{k=1}^{\infty}\) that converges to \((y_1, y_2, x_1, x_2, w_1, w_2, z)\). Suppose that the sequence \(\{v^j\}_{j=1}^{\infty}\) shrinks to zero. It follows from Proposition E.2 of Ok (2007) that \(\mathcal{V}_{1, \ell}(v)\) is upper hemiuniform at \(v = 0\) because \(y_2 - y_1 = \lim_{k \to \infty} (y^j_{2k} - y^j_1) \geq \lim_{k \to \infty} (m + \ell v^{j_k} + \delta) = m + \delta\), \(x_1 - x_2 = \lim_{k \to \infty} (x^j_{1k} - x^j_2) \geq 0\), and \(w_1 - w_2 = \lim_{k \to \infty} (w^j_{1k} - w^j_{2k}) = 0\); that is, \((y_1, y_2, x_1, x_2, w_1, w_2, z) \in \mathcal{R}_1(m + \delta) = \mathcal{V}_{1, \ell}(0)\). Next, we show the lower hemiuniformity. Fix \((y_1, y_2, x_1, x_2, w_1, w_2, z) \in \mathcal{R}_1(m + \delta) = \mathcal{V}_{1, \ell}(0)\) and consider a sequence \(\{v^j\}_{j=1}^{\infty}\) with \(\lim_{j \to \infty} v^j = 0\). For each \(j\), let \((y^j_1, y^j_2, x^j_1, x^j_2, w^j_1, w^j_2, z^j) \equiv (y_1, y_2 + \ell v^j, x_1, x_2, w_1, w_2, z)\). By construction, \(y^j_2 - y^j_1 \geq m + \ell v^j + \delta\), \(x^j_1 - x^j_2 \geq 0\), and \(w^j_1 - w^j_2 = 0\); that is \((y^j_1, y^j_2, x^j_1, x^j_2, w^j_1, w^j_2, z^j) \in \mathcal{R}_1(m + \ell v^j + \delta) = \mathcal{V}_{1, \ell}(v^j)\) for each \(j\). In addition, \(\lim_{j \to \infty} (y^j_1, y^j_2, x^j_1, x^j_2, x^j_1, w^j_2, z^j) = (y_1, y_2, x_1, x_2, w_1, w_2, z)\). By Proposition E.4 of Ok (2007), the lower hemiuniformity of \(\mathcal{V}_{1, \ell}(v)\) at \(v = 0\) is established.

1.A.7 Proof of Corollary 1.2

Proof. Fix \(t \in \{1, 2\}\) and \(\delta > 0\). There is an integer \(n_1 \in \mathbb{N}\) such that for all \(n \geq n_1\), \(v_n < \delta\) and thus

\[
\hat{T}^0_{tn}(\hat{m}_n + \delta) = \hat{T}_{tn}(\hat{m}_n + \delta)
\]

on the set \(\{|\hat{m} - m| \leq v_n\}\) under the null hypothesis. Therefore,

\[
\mathbb{P}(\hat{T}_{tn}(\hat{m}_n + \delta) \leq u) - \mathbb{P}(\hat{T}^0_{tn}(\hat{m}_n + \delta) \leq u) \leq \mathbb{P}(\hat{T}_{tn}(\hat{m}_n + \delta) \leq u \text{ and } |\hat{m}_n - m| > v_n) - \mathbb{P}(\hat{T}^0_{tn}(\hat{m}_n + \delta) \leq u \text{ and } |\hat{m}_n - m| > v_n)
\]

for all \(u \in \mathbb{R}\) and \(n \geq n_1\). Condition EB1 implies that

\[
\sup_{u \in \mathbb{R}} \left| \mathbb{P}(\hat{T}_{tn}(\hat{m}_n + \delta) \leq u) - \mathbb{P}(\hat{T}^0_{tn}(\hat{m}_n + \delta) \leq u) \right| \leq \mathbb{P}(|\hat{m}_n - m| > v_n) = o(1)
\]

as \(n \to \infty\). Hence, the result follows from Theorem 1.1.

1.A.8 Proof of Proposition 1.4

Proof. Let \(\tilde{m}_n \equiv \max_{i=1, \ldots, n}\{1 \|\|(x_{1i}, w_{1i}) - (x_{1i}, w_{1i})\|\|_{\tau_n} |\varepsilon_{1i} - \varepsilon_{1i}|\} \) and \(\eta\) be a positive number greater than \(2c_L\). Condition M2 implies \(|\hat{m}_n - \tilde{m}_n| \leq \eta \tau_n / 2\) with probability one. By Condition M1, there is an \(n_0 \in \mathbb{N}\) such that \(m > \eta \tau_n / 2\) for all \(n \geq n_0\). It follows
that for all \( n \geq n_0 \),

\[
P(|\hat{m}_n - m| > \eta \tau_n) \leq P(\hat{m}_n \leq m - \eta \tau_n/2)
= [P(\|\triangle\| \leq \tau_n)]^{[\|\varepsilon_2 - \varepsilon_1\| \leq m - \eta \tau_n/2]} \]  \tag{1.1.7}

where \( \triangle = (X_2 - X_1, W_2 - W_1) \). By Conditions D1, D2, and M3, there is a constant \( c_1 > 0 \) such that

\[
P(\|\triangle\| \leq \tau_n)]^{[\|\varepsilon_2 - \varepsilon_1\| \leq m - \eta \tau_n/2} \]
= \mathbb{E} [P(\|\triangle\| \leq \tau_n)]^{[\|\varepsilon_2 - \varepsilon_1\| \leq m - \eta \tau_n/2} \mid X_1, X_2, W_1, W_2, Z)]
= \mathbb{E} [F_{|\varepsilon_2 - \varepsilon_1| \mid Z}(m - \eta \tau_n/2 \mid Z)]^{[\|\triangle\| \leq \tau_n]}]
= \mathbb{E} [1 - \mathbb{P}(\|\triangle\| \leq \tau_n)]^{[\|\varepsilon_2 - \varepsilon_1\| \leq m - \eta \tau_n/2} \mid Z)] \]
\leq \mathbb{E} [1 - c_1 \eta \tau_n]^{[\|\triangle\| \leq \tau_n]} \mid Z)] \]
= 1 - c_1 \eta \tau_n \mathbb{P}(\|\triangle\| \leq \tau_n)]^{[\|\varepsilon_2 - \varepsilon_1\| \leq m - \eta \tau_n/2} \] \tag{1.1.8}

for all \( n \geq n_0 \). In addition, Condition M4 ensures that for all \( n \) sufficiently large,

\[
P(\|\triangle\| \leq \tau_n) \geq \frac{1}{2} f_\triangle(0) \text{Leb}(B(0, \tau_n)) = \frac{\pi \left( \frac{1 + d_W}{2} \right)}{2 \Gamma \left( \frac{3 + d_W}{2} \right)} f_\triangle(0) \tau_n^{(1 + d_W)} \] \tag{1.1.9}

where \( \Gamma(\cdot) \) is the Gamma function and \( \text{Leb}(B(0, \tau_n)) \) is the Lebesgue measure of the ball with origin \( 0 \) and radius \( \tau_n \).

Combining (1.1.7), (1.1.8), and (1.1.9), we obtain that there is a constant \( c_2 = c_2(c_1, d_W, f_\triangle(0)) > 0 \) such that for all \( n \) large,

\[
P(|\hat{m}_n - m| > \eta \tau_n) \leq \left[ 1 - c_2 \eta \tau_n^{(2 + d_W)} \right]^n .
\]

Therefore, we obtain

\[
\limsup_{n \to \infty} \mathbb{P}(|\hat{m}_n - m| > \eta \tau_n) \leq \lim_{n \to \infty} \left[ 1 - c_2 \eta \tau_n^{(2 + d_W)} \right]^n = \exp\{-c_0 c_2 \eta\}
\]

by Condition M1.

\[ \square \]

\section*{Appendix 1.B Smoothed kernel CDF estimation}

In this section, we verify the high-level condition EB2 for the statistics \( \hat{T}_{tn}^0 \) and \( \hat{T}_{tn}^* \) constructed by smoothed kernel CDF estimation. In what follows, we consider continuous explanatory variable and covariates \((X_1, X_2, W_1, W_2, Z)\).
1.B.1 Notation

We begin with some notation for smoothed kernel CDF estimation. For any generic $d_U$-dimensional random vector $U$, its density function is denoted by $f_U$ and the associated product kernel function is denoted by $K_U(u_1, u_2; z) \equiv \prod_{j=1}^{d_U} k\left(\frac{u_{1j}-u_{2j}}{\varsigma}\right)$ for some univariate kernel density function $k$ and bandwidth $\varsigma > 0$. Let $\psi(u) = \int_{-\infty}^{u} k(u) \, du$ be the distribution function derived from the kernel $k$. We consider the smoothed kernel CDF estimators

\[
\hat{F}_{Y|X_iW_iZ}(y_t|x_t, w_t, z) \equiv \frac{\sum_{i=1}^{n} \psi\left(\frac{y_t - Y_{it}}{b_n}\right) K_X(X_{it}, x_t; b_n) K_W(W_{it}, w_t; b_n) K_Z(Z_{it}, z; h_n)}{\sum_{i=1}^{n} K_X(X_{it}, x_t; b_n) K_W(W_{it}, w_t; b_n) K_Z(Z_{it}, z; h_n)}
\]

for each $t \in \{1, 2\}$ and

\[
\hat{F}_{Y_1Y_2|X_1X_2W_1W_2Z}(y_1, y_2|x_1, x_2, w_1, w_2, z) \equiv \frac{\sum_{i=1}^{n} \left[ \prod_{i=1}^{2} \psi\left(\frac{y_i - Y_{iit}}{h_n}\right) K_X(X_{iit}, x_{iti}; h_n) K_W(W_{iit}, w_{iti}; h_n) \right] K_Z(Z_{iti}, z; h_n)}{\sum_{i=1}^{n} \left[ \prod_{i=1}^{2} K_X(X_{iit}, x_{iti}; h_n) K_W(W_{iit}, w_{iti}; h_n) \right] K_Z(Z_{iti}, z; h_n)}
\]

(1.B.1)

where $\{b_n\}_{n=1}^{\infty}$ and $\{h_n\}_{n=1}^{\infty}$ are sequences of bandwidth with $b_n = h_n^2$; in addition, we set $\lambda \equiv 2(1 + d_W) + d_Z$ and the scale $\varphi_n \equiv (nh_n^\lambda)^{1/2}$. We also construct the bootstrap smoothed kernel CDF estimators by replacing the sample $\mathcal{D}_n$ with the bootstrap data $\mathcal{D}_n^*$ in (1.B.1); specifically,

\[
\hat{F}_{Y_i|X_iW_iZ}(y_t|x_t, w_t, z) \equiv \frac{\sum_{i=1}^{n} \psi\left(\frac{y_t - Y_{iit}^*}{b_n}\right) K_X(X_{iit}^*, x_{iti}; b_n) K_W(W_{iit}^*, w_{iti}; b_n) K_Z(Z_{iti}^*, z; h_n)}{\sum_{i=1}^{n} K_X(X_{iit}^*, x_{iti}; b_n) K_W(W_{iit}^*, w_{iti}; b_n) K_Z(Z_{iti}^*, z; h_n)}
\]

for each $t \in \{1, 2\}$ and

\[
\hat{F}_{Y_1Y_2|X_1X_2W_1W_2Z}(y_1, y_2|x_1, x_2, w_1, w_2, z) \equiv \frac{\sum_{i=1}^{n} \left[ \prod_{i=1}^{2} \psi\left(\frac{y_i - Y_{iitt}^*}{h_n}\right) K_X(X_{iitt}^*, x_{iit}; h_n) K_W(W_{iitt}^*, w_{iit}; h_n) \right] K_Z(Z_{iitt}^*, z; h_n)}{\sum_{i=1}^{n} \left[ \prod_{i=1}^{2} K_X(X_{iitt}^*, x_{iit}; h_n) K_W(W_{iitt}^*, w_{iit}; h_n) \right] K_Z(Z_{iitt}^*, z; h_n)}
\]

(1.B.2)

42
Moreover, let

\[ \hat{f}_{X_iW_iZ}(x_t, w_t, z) \]

\[ \equiv \frac{1}{nh^2} \sum_{i=1}^{n} K_X(X_{it}, x_t; b_n) K_W(W_{it}, w_t; b_n) K_Z(Z_i, z; h_n), \]

\[ \hat{f}_{X_1X_2W_1W_2Z}(x_1, x_2, w_1, w_2, z) \]

\[ \equiv \frac{1}{nh^2} \sum_{i=1}^{n} \left[ \prod_{t=1}^{2} K_X(X_{it}, x_t; h_n) K_W(W_{it}, w_t; h_n) \right] K_Z(Z_i, z; h_n), \]

\[ \hat{g}_{Y_1X_1W_1Z}(y_t, x_t, w_t, z) \]

\[ \equiv \frac{1}{nh^2} \sum_{i=1}^{n} \psi \left( \frac{y_t - Y_{it}}{b_n} \right) K_X(X_{it}, x_t; b_n) K_W(W_{it}, w_t; b_n) K_Z(Z_i, z; h_n), \]

\[ \hat{g}_{Y_1Y_2X_1X_2W_1W_2Z}(y_1, y_2, x_1, x_2, w_1, w_2, z) \]

\[ \equiv \frac{1}{nh^2} \sum_{i=1}^{n} \left[ \prod_{t=1}^{2} \psi \left( \frac{y_t - Y_{it}}{b_n} \right) K_X(X_{it}, x_t; h_n) K_W(W_{it}, w_t; h_n) \right] K_Z(Z_i, z; h_n) \]

be the estimators of \( f_{X_iW_iZ}(x_t, w_t, z) \), \( f_{X_1X_2W_1W_2Z}(x_1, x_2, w_1, w_2, z) \),

\[ g_{Y_1X_1W_1Z}(y_t, x_t, w_t, z) = \int_{-\infty}^{y_t} f_{Y_1X_1W_1Z}(u, x_t, w_t, z) \, du, \]

\[ g_{Y_1Y_2X_1X_2W_1W_2Z}(y_1, y_2, x_1, x_2, w_1, w_2, z) \]

\[ = \int_{-\infty}^{y_1} \int_{-\infty}^{y_2} f_{Y_1Y_2X_1X_2W_1W_2Z}(u_1, u_2, x_1, x_2, w_1, w_2, z) \, du_2 \, du_1, \]

respectively.

Similarly, \((\hat{g}_{Y_1X_1W_1Z}^*, \hat{f}_{X_iW_iZ}^*, \hat{g}_{Y_1Y_2X_1X_2W_1W_2Z}^*, \hat{f}_{X_1X_2W_1W_2Z}^*)\) is the bootstrap counterpart of \((\hat{g}_{Y_1X_1W_1Z}, \hat{f}_{X_iW_iZ}, \hat{g}_{Y_1Y_2X_1X_2W_1W_2Z}, \hat{f}_{X_1X_2W_1W_2Z})\) by replacing the sample \( D_n \) with the bootstrap data \( D_n^* \).

To avoid the boundary effect due to kernel estimation, we take suprema \( \hat{T}_{tn}(...) \) and \( \hat{T}_{tn}^0(...) \) over a compact subset of \( R_t(...) \) excluding the boundary range of the joint support \( \mathcal{Y}^2 \mathcal{X}^2 \mathcal{W}^2 \mathcal{Z} \). Specifically, let \( \mathcal{Y}_{\text{trim}} \) (\( \mathcal{X}_{\text{trim}}, \mathcal{W}_{\text{trim}}, \mathcal{Z}_{\text{trim}} \)) be some compact subset of \( \mathcal{Y} \) (\( \mathcal{X}, \mathcal{W}, \mathcal{Z} \), respectively) excluding its boundary range, and let \( \mathcal{R}_t(...) \equiv \mathcal{Y}_{\text{trim}}^2 \mathcal{X}_{\text{trim}}^2 \mathcal{W}_{\text{trim}}^2 \mathcal{Z}_{\text{trim}}^2 \) \( \cap \) \( \mathcal{R}_t(...) \) for each \( t \in \{1, 2\} \). We define the test statistic, the centered counterpart, and the bootstrap counterpart as

\[ \hat{T}_{tn}(...) \equiv \sup \left\{ \hat{Q}_{tn}(y_1, y_2, x_1, x_2, w_1, w_2, z) : (y_1, y_2, x_1, x_2, w_1, w_2, z) \in \mathcal{R}_t(...) \right\}, \]

\[ \hat{T}_{tn}^0(...) \equiv \sup \left\{ \hat{Q}_{tn}^0(y_1, y_2, x_1, x_2, w_1, w_2, z) : (y_1, y_2, x_1, x_2, w_1, w_2, z) \in \mathcal{R}_t(...) \right\}, \]

43
and

\[ \hat{T}^*_{tn}(\cdot) \equiv \sup \left\{ \hat{Q}^*_{tn}(y_1, y_2, x_1, x_2, w_1, w_2, z) : (y_1, y_2, x_1, x_2, w_1, w_2, z) \in \mathcal{R}_t(\cdot) \right\} \]

where \( \hat{Q}_{tn}, \hat{Q}^0_{tn}, \) and \( \hat{Q}^*_{tn} \) are constructed based on the kernel CDF estimators in (1.B.1) and (1.B.2) by the formulas in Section 1.3. Without confusion, we drop arguments of functions below for ease of writing. The linearized counterpart of \( \hat{Q}^0_{tn} \) and that of \( \hat{Q}^*_{tn} \) are denoted by

\[ \tilde{Q}^0_{tn} \equiv \varphi_n \left[ \frac{1}{f_{X_i W_i Z}} (\hat{g}_{Y_i X_i W_i Z} - \mathbb{E}(\hat{g}_{Y_i X_i W_i Z})) \right. \]

\[ - \frac{F_{Y_i X_i W_i Z}}{f_{X_i W_i Z}} \left( \hat{f}_{X_i W_i Z} - \mathbb{E}(\hat{f}_{X_i W_i Z}) \right) \]

\[ - \frac{1}{f_{X_i X_2 W_i W_2 Z}} (\hat{g}_{Y_i Y_2 X_i X_2 W_2 W_2 Z} - \mathbb{E}(\hat{g}_{Y_i Y_2 X_i X_2 W_2 W_2 Z})) \]

\[ + \frac{F_{Y_i Y_2 X_i X_2 W_i W_2 Z}}{f_{X_i X_2 W_i W_2 Z}} \left( \hat{f}_{X_i X_2 W_i W_2 Z} - \mathbb{E}(\hat{f}_{X_i X_2 W_i W_2 Z}) \right) \right] \]

and

\[ \tilde{Q}^*_{tn} \equiv \varphi_n \left[ \frac{1}{f_{X_i W_i Z}} (\hat{g}^*_{Y_i X_i W_i Z} - \hat{g}_{Y_i X_i W_i Z}) \right. \]

\[ - \frac{F_{Y_i X_i W_i Z}}{f_{X_i W_i Z}} \left( \hat{f}^*_{X_i W_i Z} - \hat{f}_{X_i W_i Z} \right) \]

\[ - \frac{1}{f_{X_i X_2 W_i W_2 Z}} (\hat{g}^*_{Y_i Y_2 X_i X_2 W_2 W_2 Z} - \hat{g}_{Y_i Y_2 X_i X_2 W_2 W_2 Z}) \]

\[ + \frac{F_{Y_i Y_2 X_i X_2 W_i W_2 Z}}{f_{X_i X_2 W_i W_2 Z}} \left( \hat{f}^*_{X_i X_2 W_i W_2 Z} - \hat{f}_{X_i X_2 W_i W_2 Z} \right) \right], \]

respectively. Let

\[ \tilde{T}^0_{tn}(\cdot) \equiv \sup \left\{ \tilde{Q}^0_{tn}(y_1, y_2, x_1, x_2, w_1, w_2, z) : (y_1, y_2, x_1, x_2, w_1, w_2, z) \in \mathcal{R}_t(\cdot) \right\} \]

and

\[ \tilde{T}^*_{tn}(\cdot) \equiv \sup \left\{ \tilde{Q}^*_{tn}(y_1, y_2, x_1, x_2, w_1, w_2, z) : (y_1, y_2, x_1, x_2, w_1, w_2, z) \in \mathcal{R}_t(\cdot) \right\}. \]

Furthermore, our task is to find a sequence of tight centered Gaussian processes and
its associated sequence of suprema in Condition EB2. For each \( t \in \{1, 2\} \), let
\[
q_{tn}(Y_1, Y_2, X_1, X_2, W_1, W_2, Z; y_1, y_2, x_1, x_2, w_1, w_2, z)
\equiv \psi \left( \frac{y_t - Y_t}{b_n} \right) K_X(X_t, x_t; b_n) K_W(W_t, w_t; b_n) \frac{K_Z(Z, z; h_n)}{f_{X,W,Z}(x_t, w_t, z)}
- F_{Y_1|X_t,W_t,Z}(y_1|x_t, w_t, z) K_X(X_t, x_t; b_n) K_W(W_t, w_t; b_n) \frac{K_Z(Z, z; h_n)}{f_{X,W,Z}(x_t, w_t, z)}
- \left[ \prod_{t=1}^2 \psi \left( \frac{y_t - Y_t}{h_n} \right) K_X(X_t, x_t; h_n) K_W(W_t, w_t; h_n) \right] \frac{K_Z(Z, z; h_n)}{f_{X,W,Z}(x_1, x_2, w_1, w_2, z)}
+ F_{Y_1 Y_2|X_t X_2 W_1 W_2 Z}(y_1, y_2|x_1, x_2, w_1, w_2, z) \left[ \prod_{t=1}^2 K_X(X_t, x_t; h_n) K_W(W_t, w_t; h_n) \right] \frac{K_Z(Z, z; h_n)}{f_{X,W,Z}(x_1, x_2, w_1, w_2, z)}
\]
and
\[
Q_{tn}(\cdot) \equiv \left\{ q_{tn}(Y_1, Y_2, X_1, X_2, W_1, W_2, Z; y_1, y_2, x_1, x_2, w_1, w_2, z) : (y_1, y_2, x_1, x_2, w_1, w_2, z) \in \overline{R}_t(\cdot) \right\}.
\]
For each \( t \in \{1, 2\} \) and \( n \in \mathbb{N} \), let \( G_{tn} \) be a Gaussian process in \((l^\infty(\overline{R}_t(m)), \|\cdot\|_\infty)\) with mean zero and covariance function
\[
\mathbb{E} [G_{tn}(y_1, y_2, x_1, x_2, w_1, w_2, z) G_{tn}(y_1', y_2', x_1', x_2', w_1', w_2', z')]
= h_n^{-\lambda} \text{Cov} \left[ q_{tn}(Y_1, Y_2, X_1, X_2, W_1, W_2, Z; y_1, y_2, x_1, x_2, w_1, w_2, z),
q_{tn}(Y_1, Y_2, X_1, X_2, W_1, W_2, Z; y_1', y_2', x_1', x_2', w_1', w_2', z') \right]
\]
(1.B.4)
for \((y_1, y_2, x_1, x_2, w_1, w_2, z)\) and \((y_1', y_2', x_1', x_2', w_1', w_2', z')\) in \( \overline{R}_t(m) \).\(^{29}\) Let \( T_{tn}(\cdot) \) be the supremum of \( G_{tn} \) over the region \( \overline{R}_t(\cdot) \); that is,
\[
T_{tn}(\cdot) \equiv \sup \{G_{tn}(y_1, y_2, x_1, x_2, w_1, w_2, z) : (y_1, y_2, x_1, x_2, w_1, w_2, z) \in \overline{R}_t(\cdot) \}.
\]
Finally, we can regard \( \tilde{Q}_{tn}^0 \) and \( \tilde{Q}_{tn}^* \) as an empirical process and an empirical bootstrap process, respectively, indexed by \( \{ h_n^{-\lambda/2} q_{tn} : q_{tn} \in Q_{tn}(\cdot) \} \), \( \|\cdot\|_\infty \). Concretely,
\[
\tilde{Q}_{tn}^0(y_1, y_2, x_1, x_2, w_1, w_2, z)
= \frac{1}{\sqrt{nh_n}} \sum_{i=1}^n \{ q_{tn}(Y_{1,i}, Y_{2,i}, X_{1,i}, X_{2,i}, W_{1,i}, W_{2,i}, Z_i; y_1, y_2, x_1, x_2, w_1, w_2, z)
- \mathbb{E} [q_{tn}(Y_1, Y_2, X_1, X_2, W_1, W_2, Z; y_1, y_2, x_1, x_2, w_1, w_2, z)] \}
\]
\(^{29}\) Such a Gaussian process \( G_{tn} \) exists by Kolmogorov’s extension theorem. We will show in Proposition 1.5 that there exists a version of \( G_{tn} \) that is tight.

45
and
\[
\tilde{Q}_{tn}^*(y_1, y_2, x_1, x_2, w_1, w_2, z) = \frac{1}{\sqrt{n}h_n^2} \sum_{i=1}^n \left\{ q_{tn}(Y_{i1}, Y_{i2}, X_{i1}, X_{i2}, W_{i1}, W_{i2}, Z_i; y_1, y_2, x_1, x_2, w_1, w_2, z) \\
- q_{tn}(Y_{i1}, Y_{i2}, X_{i1}, X_{i2}, W_{i1}, W_{i2}, Z_i; y_1, y_2, x_1, x_2, w_1, w_2, z) \right\}
\]
for \((y_1, y_2, x_1, x_2, w_1, w_2, z)\) in \(\mathcal{R}_t(m)\). Their associated suprema \(\tilde{T}_{tn}^0\) and \(\tilde{T}_{tn}^*\) can thus be coupled by the empirical process method developed by Chernozhukov et al. (2014a, 2016).

For the sake of completeness, we now introduce notation from the empirical process literature. The probability space \((\Omega, \mathcal{F}, P)\) is assumed to be rich enough in the sense that there exists a uniform random variable on \((0, 1)\) defined on \((\Omega, \mathcal{F}, P)\) independent of the sample \(\mathcal{D}_n\) and the empirical bootstrap sample \(\mathcal{D}_n^*\). The notation \(d\) denotes equality in distribution while \(d|\mathcal{D}_n\) denotes equality in conditional distribution given the sample \(\mathcal{D}_n\). For any generic measurable space \((S, \mathcal{S})\) and a class \(\mathcal{F}\) of measurable functions defined on \((S, \mathcal{S})\), we say that the class \(\mathcal{F}\) is pointwise measurable if it contains a countable subset \(\mathcal{F}_0\) such that for every \(f \in \mathcal{F}\), there is a sequence \(\{f_j\}_{j=1}^\infty\) of functions in \(\mathcal{F}_0\) with \(f_j(u) \to f(u)\) for every \(u \in S\). Let \(F\) be a measurable envelope for \(\mathcal{F}\), that is, \(F(u) \geq \sup_{f \in \mathcal{F}} |f(u)|\) for every \(u \in S\). The class \(\mathcal{F}\) is bounded measurable VC with a set of VC characteristics \((C, V) \in \mathbb{R}_+^2\) if it is separable or image admissible Suslin and for every probability measure \(Q\) on \((S, \mathcal{S})\), and every \(\eta \in (0, 1)\),

\[N(\mathcal{F}, L_2(Q), \eta \|F\|_{L_2(Q)}) \leq \left( \frac{C}{\eta} \right)^V,\]

where \(N(\mathcal{F}, L_2(Q), \eta)\) is the \(\eta\)-covering number of the metric space \((\mathcal{F}, L_2(Q))\). See van der Vaart and Wellner (1996) for the definition of the covering number. Throughout this chapter, the class \(\mathcal{F}\) of functions is equipped with the uniform \(\|f\|_\infty \equiv \sup_{u \in S} |f(u)|\). For any real numbers \(u_1\) and \(u_2\), \(u_1 \leq u_2\) means \(u_1 \leq C u_2\) for some universal constant \(C > 0\); similarly, \(u_1 \gtrsim u_2\) is equivalent to \(u_2 \lesssim u_1\).

1.B.2 Asymptotic theory

As shown in Figure 1.3, our strategy to validate Condition EB2 for the statistics \(\hat{T}_{tn}^0\) and \(\hat{T}_{tn}^*\) constructed by the smoothed kernel CDF estimators consists of the fourfold couplings: \(\hat{T}_{tn}^0(m + \ell v_n + \delta)\) is first coupled with \(\hat{T}_{tn}^0(m + \ell v_n + \delta)\), which is then coupled with \(T_{tn}(m + \ell v_n + \delta)\); similarly, \(\hat{T}_{tn}^*(m + \ell v_n + \delta)\) is first coupled with \(\hat{T}_{tn}^*(m + \ell v_n + \delta)\), which is then coupled with \(T_{tn}(m + \ell v_n + \delta)\). To establish these couplings, we impose
the following conditions on the kernel, densities, and bandwidth.

**Condition [K]:**

K1 The kernel function $k$ is Lipschitz continuous and of bounded variation on its support $[-1,1]$.

K2 The kernel function $k$ is of second order; specifically, $\int k(u)\,du = 1$, $\int uk(u)\,du = 0$, and $\int u^2k(u)\,du < \infty$.

**Condition [S]:**

S1 The functions $f_{X_1W_1Z}$ and $f_{X_2W_2Z}$ are bounded above on $\mathcal{XWZ}$, and the function $f_{X_1X_2W_1W_2Z}$ is also bounded above on $\mathcal{X}^2\mathcal{W}^2\mathcal{Z}$.

S2 The functions $f_{X_1W_1Z}$, $f_{X_2W_2Z}$, $f_{X_1X_2W_1W_2Z}$, $g_{Y_1X_1W_1Z}$, $g_{Y_2X_2W_2Z}$, and $g_{Y_1Y_2X_1X_2W_1W_2Z}$ are all twice differentiable with continuous and uniformly bounded second derivatives on any compact subset of their joint support, respectively.

S3 The functions $f_{X_1W_1Z}$ and $f_{X_2W_2Z}$ are bounded away from zero on any compact subset of $\mathcal{XWZ}$, and the function $f_{X_1X_2W_1W_2Z}$ is also bounded away from zero on any compact subset of $\mathcal{X}^2\mathcal{W}^2\mathcal{Z}$.

**Condition [R]:**

R1 $\lim_{n \to \infty} h_n = 0$, $\lim_{n \to \infty} \frac{|\log h_n|^2}{n h_n^4} = 0$, and $\lim_{n \to \infty} n h_n^{(\lambda+4)} = 0$.

R2 For some $\kappa_1 > 0$, $\lim_{n \to \infty} \frac{|\log h_n|^2(\log n)^{(1+\kappa_1)}}{n h_n^4} = 0$, $\lim_{n \to \infty} n h_n^{(\lambda+4)}(\log n)^{(1+\kappa_1)} = 0$, and $\lim_{n \to \infty} (\log n)^{(a+\kappa_1)} = 0$.

R3 For some $\kappa_2 \in (0,1)$, $\lim_{n \to \infty} \frac{\log n}{n^{(a+\kappa_2)} h_n^4} = 0$.  

Figure 1.3: Couplings
R4 For some \( \kappa_3 > 0 \), \( \limsup_{n \to \infty} \frac{|\log h_n|}{\log n} \leq \kappa_3 \).

Condition R on the bandwidth is mild; in addition, the requirements on the kernel and densities, imposed by Conditions K and S, are common in the literature on uniform consistency of kernel estimation. See for example Bierens (1983) and Giné and Guillou (2002).

With these conditions, the coupling between \( \hat{T}_n(0) \) and \( \hat{\ell}_n(0) \) is established in Proposition 1.5 below.

**Proposition 1.5.** Suppose Conditions K1-K2, R1, and S1-S3 hold. Then for each \( t \in \{1, 2\} \), \( \ell \in \{-1, 1\} \), and \( \delta > 0 \),

\[
\left| \hat{T}_n(0)(m + \ell v_n + \delta) - \hat{T}_n(0)(m + \ell v_n + \delta) \right| = O_p \left( \frac{|\log h_n|}{\sqrt{nh_n}} + \sqrt{nh_n^{(\lambda+4)}} \right)
\]

as \( n \to \infty \). If in addition Conditions R2-R4 hold and \( \inf_{q \in Q_n(m)} \text{Var}(q) > 0 \) for each \( n \), then for each \( t \in \{1, 2\} \), \( \ell \in \{-1, 1\} \), and \( \delta > 0 \),

\[
\sup_{u \in \mathbb{R}} \left| \mathbb{P} \left( \hat{T}_n(0)(m + \ell v_n + \delta) \leq u \right) - \mathbb{P} \left( \ell_n(0)(m + \ell v_n + \delta) \leq u \right) \right| = o(1)
\]

as \( n \to \infty \).

**Proof.** Fix \( t \in \{1, 2\} \), \( \ell \in \{-1, 1\} \), and \( \delta > 0 \). To prove the first assertion, we aim to find an upper bound of \( \left| \hat{T}_n(0)(m + \ell v_n + \delta) - \hat{T}_n(0)(m + \ell v_n + \delta) \right| \). Simple algebra shows that

\[
\hat{F}_{Y_i|X_i,W_i,Z} - F_{Y_i|X_i,W_i,Z} \\
= \frac{1}{f_{X_i|W_i,Z}} \left[ \hat{g}_{Y_i,X_i,W_i,Z} - \mathbb{E} \left( \hat{g}_{Y_i,X_i,W_i,Z} \right) \right] - \frac{F_{Y_i|X_i,W_i,Z}}{f_{X_i|W_i,Z}} \left[ \hat{f}_{X_i,W_i,Z} - \mathbb{E} \left( \hat{f}_{X_i,W_i,Z} \right) \right] + R_{Y_i,X_i,W_i,Z}
\]

where

\[
R_{Y_i,X_i,W_i,Z} \\
= \frac{1}{f_{X_i|W_i,Z}} \left[ \mathbb{E} \left( \hat{g}_{Y_i,X_i,W_i,Z} \right) - g_{Y_i,X_i,W_i,Z} \right] - \frac{F_{Y_i|X_i,W_i,Z}}{f_{X_i|W_i,Z}} \left[ \mathbb{E} \left( \hat{f}_{X_i,W_i,Z} \right) - f_{X_i,W_i,Z} \right] \\
+ \frac{1}{f_{X_i|W_i,Z}f_{X_i|W_i,Z}} \left[ \hat{f}_{X_i,W_i,Z} - f_{X_i,W_i,Z} \right] \cdot \left\{ F_{Y_i|X_i,W_i,Z} \left[ \hat{f}_{X_i,W_i,Z} - f_{X_i,W_i,Z} \right] - \left[ g_{Y_i,X_i,W_i,Z} - g_{Y_i,X_i,W_i,Z} \right] \right\}.
\]
Similarly, we have
\[
\tilde{F}_{Y_1|Y_2|x_1x_2w_1w_2} - F_{Y_1|Y_2|x_1x_2w_1w_2} = \frac{1}{f_{x_1x_2w_1w_2}} \left( \hat{g}_{Y_1|Y_2|x_1x_2w_1w_2} - E[\hat{g}_{Y_1|Y_2|x_1x_2w_1w_2}] \right)
- \frac{F_{Y_1|Y_2|x_1x_2w_1w_2}}{f_{x_1x_2w_1w_2}} \left( \hat{f}_{x_1x_2w_1w_2} - E[\hat{f}_{x_1x_2w_1w_2}] \right) + R_{Y_1|Y_2|x_1x_2w_1w_2}
\]
where
\[
R_{Y_1|Y_2|x_1x_2w_1w_2} = \frac{1}{f_{x_1x_2w_1w_2}} \left[ E(\hat{g}_{Y_1|Y_2|x_1x_2w_1w_2}) - g_{Y_1|Y_2|x_1x_2w_1w_2} \right]
- \frac{F_{Y_1|Y_2|x_1x_2w_1w_2}}{f_{x_1x_2w_1w_2}} \left[ E(\hat{f}_{x_1x_2w_1w_2}) - f_{x_1x_2w_1w_2} \right]
+ \frac{\hat{f}_{x_1x_1w_1w_2} - f_{x_1x_1w_1w_2}}{f_{x_1x_1w_1w_2}f_{x_1x_1w_1w_2}} \left[ F_{Y_1|Y_2|x_1x_2w_1w_2} \left( \hat{f}_{x_1x_1w_1w_2} - f_{x_1x_1w_1w_2} \right)
- [\hat{g}_{Y_1|Y_2|x_1x_2w_1w_2} - g_{Y_1|Y_2|x_1x_2w_1w_2}] \right].
\]

It follows that
\[
\hat{Q}_{tn}^0 - \tilde{Q}_{tn}^0 = \sqrt{nh_n^\lambda} [R_{Y_1|Y_2|x_1w_1} - R_{Y_1|Y_2|x_1x_2w_1w_2}].
\]
Since the sequence \(\{v_n\}_{n=1}^\infty\) shrinks to zero, there is an \(n_1 \in \mathbb{N}\) such that \(\delta > v_n\) and \(\mathcal{R}_t(m - v_n + \delta) \subset \mathcal{R}_t(m)\) for all \(n \geq n_1\). Therefore, we obtain that for all \(n \geq n_1\),
\[
\left| \hat{T}_{tn}^\lambda(m + \ell v_n + \delta) - \tilde{T}_{tn}^\lambda(m + \ell v_n + \delta) \right| \leq \sqrt{nh_n^\lambda} \sup_{\mathcal{R}_t(m + \ell v_n + \delta)} \left| \hat{Q}_{tn}^0 - \tilde{Q}_{tn}^0 \right|
\leq \sqrt{nh_n^\lambda} \sup_{\mathcal{R}_t(m)} |R_{Y_1|Y_2|x_1w_1} - R_{Y_1|Y_2|x_1x_2w_1w_2}|
\leq O_p \left( \frac{|\log h_n|}{\sqrt{nh_n^\lambda}} + \sqrt{nh_n^{(\lambda+4)}} \right) \quad (1.B.5)
\]
where the last equality holds by Lemma 1.3.

To prove the second assertion, we start by showing the existence of a sequence \(\{G_{tn}\}_{n=1}^\infty\) of tight Gaussian processes in \((L^\infty(\mathcal{R}_t(m)), ||.||_\infty))\) with mean zero and covariance function defined in (1.B.4). Lemma 1.2 shows that for each \(t \in \{1, 2\}\) and \(n \in \mathbb{N}\), the class \(Q_{tn}(m)\) defined in (1.B.3) is bounded measurable VC with its VC characteristics \((C_Q, V_Q)\) independent of \(t\) and \(n\). It follows that
\[
\int_0^\infty \sqrt{\log N(Q_{tn}(m), L_2(\mathbb{P}), u)} \, du < \infty
\]
49
and the class \( \mathcal{Q}_{tn}(m) \) is \( \mathbb{P} \)-pre-Gaussian. Consequently, there is a tight Gaussian process \( \mathbb{B}_{tn} \) in \( (l^\infty(\mathcal{R}_t(m)), \| \cdot \|_\infty) \) with mean zero and covariance function

\[
\mathbb{E} [\mathbb{B}_{tn}(y_1, y_2, x_1, x_2, w_1, w_2, z) \mathbb{B}_{tn}(y_1', y_2', x_1', x_2', w_1', w_2', z')] = \text{Cov} \left[ q_{tn}(Y_1, Y_2, X_1, X_2, W_1, W_2, Z; y_1, y_2, x_1, x_2, w_1, w_2, z),
q_{tn}(Y_1, Y_2, X_1, X_2, W_1, W_2, Z; y_1', y_2', x_1', x_2', w_1', w_2', z') \right]
\]

for \( (y_1, y_2, x_1, x_2, w_1, w_2, z) \) and \( (y_1', y_2', x_1', x_2', w_1', w_2', z') \) in \( \mathcal{R}_t(m) \). Let

\[
\mathbb{G}_{tn}(y_1, y_2, x_1, x_2, w_1, w_2, z) \equiv h_n^{-\lambda} \mathbb{B}_{tn}(y_1, y_2, x_1, x_2, w_1, w_2, z)
\]

for all \( (y_1, y_2, x_1, x_2, w_1, w_2, z) \in \mathcal{R}_t(m) \). Since \( \mathcal{R}_t(m - v_n + \delta) \subset \mathcal{R}_t(m) \) for all \( n \geq n_1 \), the supremum

\[
T_{tn}(m - v_n + \delta) = \sup_{\mathcal{R}_t(m - v_n + \delta)} \mathbb{G}_{tn}
\]

is well-defined for all \( n \geq n_1 \).

Next, for any \( (y_1, y_2, x_1, x_2, w_1, w_2, z) \in \mathcal{R}_t(m - v_n + \delta) \) and \( j \in \{2, 3, 4\} \), we have

\[
\mathbb{E} \left[ \left| K_X(X_t, x_t; b_n) K_W(W_t, w_t; b_n) K_Z(Z, z; h_n) \right|^j \right] \leq \max \left\{ \mathbb{E} \left[ \left| \prod_{t=1}^2 K_X(X_t, x_t; h_n) K_W(W_t, w_t; h_n) \right|^j \right], \right. \\
\left. + \left( \int_{-1}^1 |k(u)|^j \, du \right)^\lambda \sup_{X \times W \times Z} f_{X'WZ} \right\}
\]

Let \( \kappa_1 \) and \( \kappa_2 \) be the positive constants given in Condition R. Applying Theorem 2.1 of Chernozhukov et al. (2016) (with \( \eta = \delta^{(1)}_n \), \( B = 0 \), \( N_B(\eta) = 1 \), \( \gamma = \gamma_n = (\log n)^{-1} \), \( b = O(1) \), and \( \sigma^2 = \sigma^2_n = O(h_n^\lambda) \) in their notation) yields that there is a sequence \( \{\tilde{\zeta}_{tn}\}_{n=1}^\infty \) of random variables such that

\[
\tilde{\zeta}_{tn} \overset{\mathbb{P}}{=} \sup_{\mathcal{R}_t(m - v_n + \delta)} \mathbb{G}_{tn} = T_{tn}(m - v_n + \delta)
\]
and as \( n \to \infty \),
\[
\left| \hat{T}_{tn}^0(m - v_n + \delta) - \tilde{\zeta}_{tn} \right| \\
= O_p \left( n^{(-1/2 + \kappa)} h_n^{-\lambda/2} (\log n)^{(1+\kappa)} + (nh_n^{\lambda})^{-1/6} (\log n) \right) \\
= O_p \left( \frac{|\log h_n|}{\sqrt{nh_n^\lambda}} + \sqrt{nh_n^{(\lambda+4)}} \right) + O_p \left( n^{(-1/2 + \kappa)} h_n^{-\lambda/2} (\log n)^{(1+\kappa)} + (nh_n^{\lambda})^{-1/6} (\log n) \right) \\
= o_p(\gamma_n)
\]
for some constant \( \kappa \in (0, \min \{1/4, \kappa_1/3, \kappa_2/3\} \). Let \( \gamma_n \equiv (\log n)^{-1+\kappa_0}/2 \) for some constant \( \kappa_0 \in (0, \kappa_1/3) \). It follows from (1.B.5) and (1.B.8) that as \( n \to \infty \),
\[
\left| \hat{T}_{tn}^0(m - v_n + \delta) - \tilde{\zeta}_{tn} \right| \\
= O_p \left( \frac{|\log h_n|}{\sqrt{nh_n^\lambda}} + \sqrt{nh_n^{(\lambda+4)}} \right) + O_p \left( n^{(-1/2 + \kappa)} h_n^{-\lambda/2} (\log n)^{(1+\kappa)} + (nh_n^{\lambda})^{-1/6} (\log n) \right) \\
= o_p(\gamma_n)
\]
where the last equality holds by Conditions R2 and R3.

Let \( \eta > 0 \). There is an integer \( n_2 \geq n_1 \) such that
\[
P \left( \left| \hat{T}_{tn}^0(m - v_n + \delta) - \tilde{\zeta}_{tn} \right| > \gamma_n \right) < \eta
\]
for all \( n \geq n_2 \). Following Lemma 2.1 of Chernozhukov et al. (2016) and the anti-concentration inequality in Lemma A.1 of Chernozhukov et al. (2014b), we obtain that for all \( n \) sufficiently large,
\[
\sup_{u \in \mathbb{R}} \left| P \left( \hat{T}_{tn}(m - v_n + \delta) \leq u \right) - P \left( T_{tn}(m - v_n + \delta) \leq u \right) \right| \\
= \sup_{u \in \mathbb{R}} \left| P \left( \hat{T}_{tn}(m - v_n + \delta) \leq u \right) - P \left( \tilde{\zeta}_{tn} \leq u \right) \right| \\
\leq \sup_{u \in \mathbb{R}} P \left( \left| \tilde{\zeta}_{tn} - u \right| \leq \gamma_n \right) + \eta \\
\leq \sup_{u \in \mathbb{R}} P \left( \left| T_{tn}(m - v_n + \delta) - u \right| \leq \gamma_n \right) + \eta \\
\leq \gamma_n \mathbb{E} \left[ T_{tn}(m) \right] + 2\eta.
\]

Finally, we have to show that \( \gamma_n \mathbb{E} \left[ T_{tn}(m) \right] \to 0 \) as \( n \to \infty \). Let
\[
\rho_{tn}((y_1, y_2, x_1, x_2, w_1, w_2, z), (y_1', y_2', x_1', x_2', w_1', w_2', z')) \\
= \sqrt{\mathbb{E} \left[ G_{tn}(y_1, y_2, x_1, x_2, w_1, w_2, z) - G_{tn}(y_1', y_2', x_1', x_2', w_1', w_2', z') \right]^2}
\]
for \((y_1, y_2, x_1, x_2, w_1, w_2, z)\) and \((y_1', y_2', x_1', x_2', w_1', w_2', z')\) in \( \mathcal{R}_t(m) \). It follows from (1.B.7) that the \( \rho_{tn} \)-diameter of \( \mathcal{R}_t(m) \) is bounded for all \( n \) sufficiently large. Furthermore, there exist positive constants \( C_1 \) and \( C_2 \) such that for all \( n \) sufficiently large,
\[
N \left( \mathcal{R}_t(m), \rho_{tn}, \varepsilon \right) \leq N \left( \mathcal{R}_t(m), \| \cdot \|, \frac{h_n^{(\lambda+C_2)/2}}{C_1 \varepsilon} \right).
\]
Following Corollary 2.2.8 of van der Vaart and Wellner (1996), we can show that as \( n \to \infty \),
\[
\mathbb{E}[T_{tn}(m)] \lesssim \sqrt{\log h_n} = O\left(\sqrt{\log n}\right) \tag{1.2.9}
\]
by Condition R4. Therefore, for all \( n \) sufficiently large,
\[
\sup_{u \in \mathbb{R}} \left| \mathbb{P}\left( \hat{T}_{tn}^0(m + \ell v_n + \delta) \leq u \right) - \mathbb{P}(T_{tn}(m + \ell v_n + \delta) \leq u) \right| \lesssim \gamma_n \mathbb{E}[T_{tn}(m)] + 2\eta < 3\eta.
\]
That is, we establish
\[
\lim_{n \to \infty} \sup_{u \in \mathbb{R}} \left| \mathbb{P}\left( \hat{T}_{tn}^0(m + \ell v_n + \delta) \leq u \right) - \mathbb{P}(T_{tn}(m + \ell v_n + \delta) \leq u) \right| = 0
\]
for \( \ell = -1 \). Similar arguments apply to the case \( \ell = 1 \) and the proof is complete. \( \square \)

To complete the coupling between \( \hat{T}_{tn}^0(m + \ell v_n + \delta) \) and \( \hat{T}_{tn}^*(m + \ell v_n + \delta) \) in Condition EB2, we also need to couple \( \hat{T}_{tn}^*(m + \ell v_n + \delta) \) with \( T_{tn}(m + \ell v_n + \delta) \). The coupling between \( \hat{T}_{tn}^*(m + \ell v_n + \delta) \) and \( T_{tn}(m + \ell v_n + \delta) \) is established in Proposition 1.6 below.

**Proposition 1.6.** Suppose Conditions K1-K2, R1-R2, and S1-S3 hold. There is a sequence \( \{\gamma_n\}_{n=1}^{\infty} \) of positive numbers with \( \lim_{n \to \infty} \gamma_n \sqrt{\log n} = 0 \) such that for each \( t \in \{1, 2\} \), \( \ell \in \{-1, 1\} \), and \( \delta > 0 \),
\[
\mathbb{P}\left( \left| \hat{T}_{tn}^*(m + \ell v_n + \delta) - \hat{T}_{tn}^*(m + \ell v_n + \delta) \right| > \gamma_n |\mathcal{D}_n\right) = o_p(1)
\]
as \( n \to \infty \). If in addition Conditions R3-R4 hold and \( \inf_{q \in Q_{tn}(m)} \text{Var}(q) > 0 \) for each \( n \), then for each \( t \in \{1, 2\} \), \( \ell \in \{-1, 1\} \), and \( \delta > 0 \),
\[
\sup_{u \in \mathbb{R}} \left| \mathbb{P}\left( \hat{T}_{tn}^*(m + \ell v_n + \delta) \leq u |\mathcal{D}_n\right) - \mathbb{P}(T_{tn}(m + \ell v_n + \delta) \leq u) \right| = o_p(1)
\]
as \( n \to \infty \).

**Proof.** Let \( \kappa_1 \) be the positive constant given in Condition R2. Let \( \nu_n \equiv \sqrt{\frac{\log h_n}{n \lambda n}} \) and \( \gamma_n \equiv (\log n)^{-\frac{1+\kappa_0}{2}} \) for some constant \( \kappa_0 \in (0, \kappa_1/3) \). Fix \( t \in \{1, 2\} \), \( \ell \in \{-1, 1\} \), and \( \delta > 0 \). In addition, fix \( \eta > 0 \). For ease of exposition, let
\[
p_n(t, \ell, \delta) \equiv \mathbb{P}\left( \left| \hat{T}_{tn}^*(m + \ell v_n + \delta) - \hat{T}_{tn}^*(m + \ell v_n + \delta) \right| > \gamma_n |\mathcal{D}_n\right)
\]

52
and \( \Lambda_n(t, c) \) be the union of the following sets
\[
\left\{ \begin{array}{l}
\sup_{\mathcal{P}_c(m)} |f_{X,W_iZ} - \hat{f}_{X,W_iZ}| > c(n + h^2) \\
\sup_{\mathcal{P}_c(m)} |g_{Y,X,W_iZ} - \hat{g}_{Y,X,W_iZ}| > c(n + h^2) \\
\sup_{\mathcal{P}_c(m)} |\hat{f}_{X_1X_2W_1W_2Z} - f_{X_1X_2W_1W_2Z}| > c(n + h^2) \\
\sup_{\mathcal{P}_c(m)} |\hat{g}_{Y_1Y_2X_1X_2W_1W_2Z} - g_{Y_1Y_2X_1X_2W_1W_2Z}| > c(n + h^2)
\end{array} \right\},
\]
for a generic positive constant \( c \). By Lemma 1.3, there is a constant \( c_0 > 0 \) such that
\[
\limsup_{n \to \infty} \mathbb{P}(\Lambda_n(t, c_0)) < \eta.
\]
It is sufficient to show that
\[
\lim_{n \to \infty} \mathbb{P}(\{p_n(t, \ell, \delta) > \eta\} \cap \{\Lambda_n(t, c_0)^c\}) = 0.
\]
For this purpose, we shall find an upper bound of \( p_n(t, \ell, \delta) \) on the complement of \( \Lambda_n(t, c_0) \). First, some tedious algebra shows that
\[
\tilde{Q}_{tn} - Q_{tn}^* = \varphi_n \left( \frac{f_{X,W_iZ} - \hat{f}_{X,W_iZ}}{f_{X,W_iZ} \hat{f}_{X,W_iZ}} \right) \left( \hat{g}_{Y,X,W_iZ} - \hat{g}_{Y,X,W_iZ} \right)
\-
\varphi_n \left( \frac{\hat{F}_{Y[X,W_iZ]} - \hat{F}_{Y[X,W_iZ]}}{\hat{f}_{X,W_iZ}} \right) \left( \hat{f}_{X,W_iZ} - \hat{f}_{X,W_iZ} \right)
\+ \varphi_n \left( \frac{\hat{f}_{X,W_iZ} - \hat{f}_{X,W_iZ}}{\hat{f}_{X,W_iZ} \hat{f}_{X,W_iZ}} \right) \left[ \hat{F}_{Y[X,W_iZ]} \left( \hat{f}_{X,W_iZ} - \hat{f}_{X,W_iZ} \right) \right] \left( \hat{g}_{Y,X,W_iZ} - \hat{g}_{Y,X,W_iZ} \right)
\-
\varphi_n \left( \frac{f_{X_1X_2W_1W_2Z} - \hat{f}_{X_1X_2W_1W_2Z}}{\hat{f}_{X_1X_2W_1W_2Z} \hat{f}_{X_1X_2W_1W_2Z}} \right) \left( \hat{g}_{Y_1Y_2X_1X_2W_1W_2Z} - \hat{g}_{Y_1Y_2X_1X_2W_1W_2Z} \right)
\+ \varphi_n \left( \frac{\hat{F}_{Y_1Y_2[X,X_1X_2W_1W_2Z]} - \hat{F}_{Y_1Y_2[X,X_1X_2W_1W_2Z]}}{\hat{f}_{X_1X_2W_1W_2Z} \hat{f}_{X_1X_2W_1W_2Z}} \right) \left( \hat{f}_{X_1X_2W_1W_2Z} - \hat{f}_{X_1X_2W_1W_2Z} \right)
\-
\varphi_n \left( \frac{\hat{f}_{X_1X_2W_1W_2Z} - \hat{f}_{X_1X_2W_1W_2Z}}{\hat{f}_{X_1X_2W_1W_2Z} \hat{f}_{X_1X_2W_1W_2Z}} \right) \left[ \hat{F}_{Y_1Y_2[X,X_1X_2W_1W_2Z]} \left( \hat{f}_{X_1X_2W_1W_2Z} - \hat{f}_{X_1X_2W_1W_2Z} \right) \right. 
\quad \left. - \left( \hat{g}_{Y_1Y_2X_1X_2W_1W_2Z} - \hat{g}_{Y_1Y_2X_1X_2W_1W_2Z} \right) \right].
\]
It follows from Conditions S1 and S3 that on the complement of $\Lambda_n(t, c_0)$,

$$
\left| \hat{T}_{tn}^*(m + \ell v_n + \delta) - \bar{T}_{tn}^*(m + \ell v_n + \delta) \right|
\leq \sup_{\mathcal{F}_t(m)} \left| \hat{Q}_{tn}^* - \bar{Q}_{tn}^* \right|
\leq \varphi_n(v_n + h_n^2) \sup_{\mathcal{F}_t(m)} \left| \hat{g}_{Y_1X_1W_1Z} - \bar{g}_{Y_1X_1W_1Z} \right| + \varphi_n(v_n + h_n^2) \sup_{\mathcal{F}_t(m)} \left| \hat{f}_{X_1W_1Z} - \bar{f}_{X_1W_1Z} \right|
\leq \varphi_n(v_n + h_n^2) \sup_{\mathcal{F}_t(m)} \left| \hat{g}_{Y_1Y_2X_1X_2W_1W_2Z} - \bar{g}_{Y_1Y_2X_1X_2W_1W_2Z} \right|
+ \varphi_n(v_n + h_n^2) \sup_{\mathcal{F}_t(m)} \left| \hat{f}_{X_1X_2W_1W_2Z} - \bar{f}_{X_1X_2W_1W_2Z} \right|
\leq \varphi_n(v_n + h_n^2) \sup_{\mathcal{F}_t(m)} \left| \hat{g}_{Y_1Y_2X_1X_2W_1W_2Z} - \bar{g}_{Y_1Y_2X_1X_2W_1W_2Z} \right|
+ \varphi_n(v_n + h_n^2) \sup_{\mathcal{F}_t(m)} \left| \hat{f}_{X_1X_2W_1W_2Z} - \bar{f}_{X_1X_2W_1W_2Z} \right|
$$

for all $n$ sufficiently large. We obtain the following upper bound of $p_n(t, \ell, \delta)$ on the complement of $\Lambda_n(t, c_0)$ for all $n$ large:

$$
p_n(t, \ell, \delta) = P\left( \left| \hat{T}_{tn}^*(m + \ell v_n + \delta) - \bar{T}_{tn}^*(m + \ell v_n + \delta) \right| > \gamma_n \bigg| \mathcal{G}_n \right)
\leq P\left( \varphi_n(v_n + h_n^2)v_n \gtrsim \gamma_n \bigg| \mathcal{G}_n \right)
\leq P\left( \sup_{\mathcal{F}_t(m)} \left| \hat{g}_{Y_1X_1W_1Z} - \bar{g}_{Y_1X_1W_1Z} \right| > c_1 \nu_n \bigg| \mathcal{G}_n \right)
\leq P\left( \sup_{\mathcal{F}_t(m)} \left| \hat{f}_{X_1W_1Z} - \bar{f}_{X_1W_1Z} \right| > c_1 \nu_n \bigg| \mathcal{G}_n \right)
\leq P\left( \sup_{\mathcal{F}_t(m)} \left| \hat{f}_{X_1X_2W_1W_2Z} - \bar{f}_{X_1X_2W_1W_2Z} \right| > c_1 \nu_n \bigg| \mathcal{G}_n \right)
\leq P\left( \sup_{\mathcal{F}_t(m)} \left| \hat{g}_{Y_1Y_2X_1X_2W_1W_2Z} - \bar{g}_{Y_1Y_2X_1X_2W_1W_2Z} \right| > c_1 \nu_n \bigg| \mathcal{G}_n \right)
\leq P\left( \sup_{\mathcal{F}_t(m)} \left| \hat{g}_{Y_1Y_2X_1X_2W_1W_2Z} - \bar{g}_{Y_1Y_2X_1X_2W_1W_2Z} \right| > c_1 \nu_n \bigg| \mathcal{G}_n \right)
$$

54
where the constant $c_1$ is selected such that

$$
\lim_{n \to \infty} \mathbb{P} \left( \sup_{\mathcal{R}_t(m)} \left| \hat{f}_{X_i} - \hat{f}_{X_i \mid \mathcal{D}_n} \right| > c_1 \nu_n \right) > \eta/5
$$

$$
= \lim_{n \to \infty} \mathbb{P} \left( \sup_{\mathcal{R}_t(m)} \left| \hat{g}_{Y_i} - \hat{g}_{Y_i \mid \mathcal{D}_n} \right| > c_1 \nu_n \right) > \eta/5
$$

$$
= \lim_{n \to \infty} \mathbb{P} \left( \sup_{\mathcal{R}_t(m)} \left| \hat{f}_{X_1X_2W_1W_2Z} - \hat{f}_{X_1X_2W_1W_2Z} > c_1 \nu_n \right| \mathcal{D}_n \right) > \eta/5
$$

$$
= \lim_{n \to \infty} \mathbb{P} \left( \sup_{\mathcal{R}_t(m)} \left| \hat{g}_{Y_1Y_2X_1X_2W_1W_2Z} - \hat{g}_{Y_1Y_2X_1X_2W_1W_2Z} > c_1 \nu_n \right| \mathcal{D}_n \right) > \eta/5
$$

= 0.

The existence of $c_1$ is guaranteed by Lemma 1.4. Therefore, applying Markov’s inequality yields

$$
\lim_{n \to \infty} \mathbb{P} \left( \{p_n(t, \ell, \delta) > \eta \} \cap \{\Lambda_n(t, c_0)\}^c \right)
$$

$$
= \lim_{n \to \infty} \mathbb{P} \left( \mathbb{P} \left( \varphi_n(\nu_n + h_n^2) \nu_n \geq \gamma_n | \mathcal{D}_n \right) > \eta/5 \right)
$$

$$
= (5/\eta) \cdot \lim_{n \to \infty} \mathbb{P} \left( \varphi_n(\nu_n + h_n^2) \nu_n \geq \gamma_n \right)
$$

= 0

by Conditions R1 and R2.

Our task now is to prove the second assertion. Let $\kappa_2$ be the positive constant given in Condition R3. As shown in the proof of Proposition 1.5, the Gaussian process $\mathbb{G}_{tn}$ is well-defined via (1.B.6) for all $n$ sufficiently large; in addition, for any $(y_1, y_2, x_1, x_2, w_1, w_2, z) \in \mathcal{R}_t(m - v_n + \delta)$ and $j \in \{2, 3, 4\}$,

$$
\mathbb{E} \left[ q_{tn}(Y_1, Y_2, X_1, X_2, W_1, W_2, Z; y_1, y_2, x_1, x_2, w_1, w_2, z) \right] \lesssim h_n^\lambda.
$$

Applying Theorem 2.3 of Chernozhukov et al. (2016) (with $\eta = \delta_n^{(3)}$, $B = 0$, $N_B(\eta) = 1$, $\gamma = \gamma_n = (\log n)^{-1}$, $b = O(1)$, and $\sigma^2 = \sigma_n^2 = O(h_n^\lambda)$ in their notation) yields that there is a sequence $\{\zeta_{tn}^*\}_{n=1}^\infty$ of random variables such that

$$
\zeta_{tn}^* \overset{d}{=} \varphi_n \sup_{\mathcal{R}_t(m - v_n + \delta)} \mathbb{G}_{tn} = T_{tn}(m - v_n + \delta)
$$

and as $n \to \infty$,

$$
\left| \hat{T}_{tn}^*(m - v_n + \delta) - \zeta_{tn}^* \right|
$$

$$
= O_p \left( n^{-1/2 + \kappa} h_n^{\lambda/2} (\log n)^{(2 + \kappa)} + (nh_n^\lambda)^{-1/6} (\log n) + (nh_n^\lambda)^{-1/4} (\log n)^{\kappa+7/4} \right)
$$
for some constant $\kappa \in (0, \min\{1/4, \kappa_1/12, \kappa_2/6\})$. Conditions R2 and R3 imply that as $n \to \infty$,
\[
\left| \hat{T}_{tn}(m - v_n + \delta) - \tilde{\zeta}_{tn}^* \right| = O_p \left( (nh_n^\lambda)^{-1/6} (\log n) \right)
\]
\[
= o_p (\gamma_n) .
\]

It follows from Markov’s inequality that
\[
\mathbb{P} \left( \left| \hat{T}_{tn}^*(m - v_n + \delta) - \tilde{\zeta}_{tn}^* \right| > \gamma_n|\mathcal{D}_n \right) = o_p (1)
\]
as $n \to \infty$. Following Lemma 2.1 of Chernozhukov et al. (2016), we obtain
\[
\sup_{u \in \mathbb{R}} \mathbb{P} \left( \hat{T}_{tn}(m - v_n + \delta) \leq u |\mathcal{D}_n \right) - \mathbb{P} \left( \tilde{T}_{tn}(m - v_n + \delta) \leq u |\mathcal{D}_n \right)
\]
\[
\leq \sup_{u \in \mathbb{R}} \mathbb{P} \left( \left| \tilde{\zeta}_{tn}^* - u \right| \leq 2\gamma_n|\mathcal{D}_n \right) + \mathbb{P} \left( \left| \hat{T}_{tn}(m - v_n + \delta) - \tilde{\zeta}_{tn}^* \right| > 2\gamma_n|\mathcal{D}_n \right)
\]
\[
\leq \sup_{u \in \mathbb{R}} \mathbb{P} \left( \left| \hat{T}_{tn}(m - v_n + \delta) - \tilde{T}_{tn}(m - v_n + \delta) \right| - \left| \tilde{\zeta}_{tn}^* \right| > \gamma_n|\mathcal{D}_n \right)
\]
\[
= o_p (1)
\]
as $n \to \infty$. By the anti-concentration inequality in Lemma A.1 of Chernozhukov et al. (2014b), as $n \to \infty$,
\[
\sup_{u \in \mathbb{R}} \mathbb{P} \left( \left| \hat{T}_{tn}(m - v_n + \delta) - u \right| \leq 2\gamma_n \right) \leq O \left( \gamma_n \mathbb{E} \left[ \hat{T}_{tn}(m) \right] \right) = o(1)
\]
where the last equality holds by (1.B.9). Similar arguments apply to the case $\ell = 1$ and the proof is complete.

The discussion after Condition EB in Subsection 1.3.2 suggests that to verify Condition EB3, it is sufficient to show for each $t \in \{1, 2\}$ and $\ell \in \{-1, 1\}$,
\[
\sup_{\mathcal{P}_t(m+v_n+\delta)} \mathcal{G}_{tn} - \sup_{\mathcal{P}_t(m+\delta)} \mathcal{G}_{tn} = O_p (\gamma_n) \quad (1.B.10)
\]
for some sequence $\{\gamma_n\}_{n=1}^\infty$ with $\gamma_n \mathbb{E} \left[ \hat{T}_{tn}(m) \right] = o(1)$ as $n \to \infty$. If the matching estimator $\hat{m}_n$ in (1.7) is used, then $v_n$ can shrink to zero at a polynomial rate by Remark 1 in Subsection 1.3.3. In this case, we conjecture that (1.B.10) is valid if the sequence $\{\mathcal{G}_{tn}\}_{n=1}^\infty$ of Gaussian processes is smooth enough. This conjecture could be viewed as some random version of an envelope theorem, which states conditions to establish smoothing properties of the value of a parameterized optimization problem. See for example Morand, Reflett, and Tarafdar (2015) and references cited therein.
1.B.3 Auxiliary lemmas

Throughout this subsection, we write $S_0$ for a generic bounded subset excluding the boundary range of the joint support $XWZ$ and $S_\psi$ for a generic bounded subset excluding the boundary range of the joint support $XYWZ$. For each $n \in \mathbb{N}$, we consider two classes of functions

$$K_n \equiv \{ K_X(\cdot, x; b_n)K_W(\cdot, w; b_n)K_Z(\cdot, z; h_n) : (x, w, z) \in S_0 \} \quad (1.B.11)$$

and

$$\Psi_n \equiv \left\{ \psi \left( \frac{y - \cdot}{b_n} \right) K_X(\cdot, x; b_n)K_W(\cdot, w; b_n)K_Z(\cdot, z; h_n) : (y, x, w, z) \in S_\psi \right\}. \quad (1.B.12)$$

**Lemma 1.2.** Suppose Condition K1 holds. There exist positive numbers $C_K$ and $V_K$ such that for each $n \in \mathbb{N}$, the class $K_n$ defined in (1.B.11) is bounded measurable VC with the VC characteristics $(C_K, V_K)$. Similarly, there exist positive numbers $C_\psi$ and $V_\psi$ such that for each $n \in \mathbb{N}$, the class $\Psi_n$ defined in (1.B.12) is bounded measurable VC with the VC characteristics $(C_\psi, V_\psi)$.

If in addition Conditions S1-S3 hold, then for each $t \in \{1, 2\}$ and $n \in \mathbb{N}$, the class $Q_{tn}(\cdot)$ defined in (1.B.3) is a pointwise measurable class of functions uniformly bounded; moreover, there exist positive numbers $C_Q$ and $V_Q$ such that the class $Q_{tn}(\cdot)$ is bounded measurable VC with the VC characteristics $(C_Q, V_Q)$.

**Proof.** Consider an arbitrary sequence $\{\varsigma_n\}_{n=1}^\infty$ of positive numbers. Lemma 22 (ii) of Nolan and Pollard (1987) implies that there are constants $c_1 > 0$ and $v_1 > 0$ such that for each $n \in \mathbb{N}$, probability measure $Q$, and $u \in (0, 1),$

$$N \left( \left\{ k \left( \frac{\cdot - x}{\varsigma_n} \right) : x \in X \right\} , L_2(Q), u \right) \leq \left( \frac{c_1}{u} \right)^{v_1}.$$

Similarly, since the function $\psi$ is weakly increasing and bounded, Lemma 22 (ii) of Nolan and Pollard (1987) also implies that there are constants $c_2 > 0$ and $v_2 > 0$ such that for each $n \in \mathbb{N}$, probability measure $Q$, and $u \in (0, 1),$

$$N \left( \left\{ \psi \left( \frac{y - \cdot}{\varsigma_n} \right) : y \in Y \right\} , L_2(Q), u \right) \leq \left( \frac{c_2}{u} \right)^{v_2}.$$

It follows that $\{k \left( \frac{\cdot - x}{\varsigma_n} \right) : x \in X\}$ and $\{\psi ((y - \cdot)/\varsigma_n) : y \in Y\}$ are bounded measurable VC with the VC characteristics $(c_1, v_1)$ and $(c_2, v_2)$, respectively. So are both $K_n$ and $\Psi_n$ by Corollary A.1 of Chernozhukov et al. (2014b). In particular, there exist positive
numbers $C_K$, $V_K$, $C_\Psi$, and $V_\Psi$ such that for each $n \in \mathbb{N}$, probability measure $Q$, and $u \in (0, 1)$,

$$N(K_n, L_2(Q), u) \leq \left( \frac{C_K}{u} \right)^{V_K} \text{ and } N(\Psi_n, L_2(Q), u) \leq \left( \frac{C_\Psi}{u} \right)^{V_\Psi}.$$  

It is straightforward to show that for each $t \in \{1, 2\}$ and $n \in \mathbb{N}$, the class $Q_{tn}(m)$ is a pointwise measurable class of functions uniformly bounded under Conditions K1 and S1-S3. The Lipschitz continuity of $[f_{X,t}x_{W_1}w_{2}z]^{-1}$ on the compact set $X_{\text{trim}}^2 W_{\text{trim}}^2 Z_{\text{trim}}$ and $g_{Y_1 Y_2 X_1}x_{W_1}w_{2}z$ on the compact set $Y_{\text{trim}}^2 X_{\text{trim}}^2 W_{\text{trim}}^2 Z_{\text{trim}}$ implies that the classes

$$\{[f_{X,t}x_{W_1}w_{2}z(x_1, x_2, w_1, w_2, z)]^{-1} : (x_1, x_2, w_1, w_2, z) \in X_{\text{trim}}^2 W_{\text{trim}}^2 Z_{\text{trim}}\}$$

and

$$\{g_{Y_1 Y_2 X_1}x_{W_1}w_{2}z(y_1, y_2, x_1, x_2, w_1, w_2, z) : (y_1, y_2, x_1, x_2, w_1, w_2, z) \in Y_{\text{trim}}^2 X_{\text{trim}}^2 W_{\text{trim}}^2 Z_{\text{trim}}\}$$

are bounded measurable VC. Similarly, for each $t \in \{1, 2\}$, the classes

$$\{[f_{X,t}w_{z}(x, w, z)]^{-1} : (x, w, z) \in X_{\text{trim}} W_{\text{trim}} Z_{\text{trim}}\}$$

and

$$\{g_{Y_1 X_1}w_{z}(y, x, w, z) : (y, x, w, z) \in Y_{\text{trim}} X_{\text{trim}} W_{\text{trim}} Z_{\text{trim}}\}$$

are bounded measurable VC. The result follows by preservation properties of bounded measurable VC classes.

\[\square\]

Lemma 1.3. Suppose Conditions K1-K2, R1, and S1-S2 hold.

(i) For each $t \in \{1, 2\}$, sup $|\hat{f}_{X,t}w_z - f_{X,t}w_z| = O_p \left( \sqrt{\frac{\log h_n}{nh_n^2}} + h_n^2 \right)$, where the supremum is taken over any bounded subset excluding the boundary range of the joint support $XWZ$.

(ii) For each $t \in \{1, 2\}$, sup $|\bar{g}_{Y_1 X_1}w_z - g_{Y_1 X_1}w_z| = O_p \left( \sqrt{\frac{\log h_n}{nh_n^2}} + h_n^2 \right)$, where the supremum is taken over any bounded subset excluding the boundary range of the joint support $YXWZ$.

(iii) sup $|\hat{f}_{X_1 X_2 X_1}w_2z - f_{X_1 X_2 X_1}w_2z| = O_p \left( \sqrt{\frac{\log h_n}{nh_n^2}} + h_n^2 \right)$, where the supremum is taken over any bounded subset excluding the boundary range of the joint support $X^2W^2Z$.

(iv) sup $|\bar{g}_{Y_1 Y_2 X_1 X_2 X_1}w_2z - g_{Y_1 Y_2 X_1 X_2 X_1}w_2z| = O_p \left( \sqrt{\frac{\log h_n}{nh_n^2}} + h_n^2 \right)$, where the supremum is taken over any bounded subset excluding the boundary range of the joint support $Y^2X^2W^2Z$. 

58
Proof. As the claims for (iii)-(iv) are similar to those for (i)-(ii), we will only present arguments for the first two statements. For ease of writing, arguments of the estimators and functions are suppressed in this proof.

We begin by proving (i). Following the technique in Lemma 2 of Bierens (1983), we can show that the deterministic part is

\[ \sup_{S_0} \left| \mathbb{E} \left[ \hat{f}_{X,t,W_t,Z} - f_{X,t,W_t,Z} \right] \right| = O \left( h_n^2 \right). \]

The class \( \mathcal{K}_n \) defined in (1.B.11) is bounded measurable VC with a set of VC characteristics independent of \( n \) by Lemma 1.2. Since

\[ \sup_{\xi \in \mathcal{K}_n} \| \xi \|_\infty \leq \left( \sup_{u \in [-1,1]} |k(u)| \right)^{(1+d_W+d_Z)} < \infty \]

and

\[ \sup_{\xi \in \mathcal{K}_n} \text{Var} \xi \leq \sup_{S_0} \mathbb{E} \left( k^2 \left( \frac{X_t - x}{b_n} \right) \prod_{j=1}^{d_W} k^2 \left( \frac{W_{tj} - w_j}{b_n} \right) \prod_{j=1}^{d_Z} k^2 \left( \frac{Z_j - z_j}{h_n} \right) \right) \]

\[ \leq b_n^{(1+d_W+d_Z)} h_n^{d_Z} \left( \int_{-1}^{1} k^2(u) \, du \right)^{(1+d_W+d_Z)} \sup_{X,W,Z} \hat{f}_{X,t,W_t,Z} \]

\[ \lesssim h_n^{\lambda} \]

by Conditions K1 and S1, Theorem 2.1 of Giné and Guillou (2002) implies

\[ \mathbb{E} \left[ \sup_{S_0} \sum_{i=1}^{n} \left( K_X(X_{it},x;b_n)K_W(W_{it},w;b_n)K_Z(Z_i,z;h_n) \right. \right. \]

\[ - \left. \left. \mathbb{E} \left[ K_X(X_{it},x;b_n)K_W(W_{it},w;b_n)K_Z(Z_i,z;h_n) \right] \right) \right] \leq | \log h_n^\lambda | + \sqrt{n h_n^\lambda | \log h_n^\lambda |} \]

\[ \lesssim \sqrt{n h_n^\lambda | \log h_n^\lambda |} \]

for \( n \) large by Condition R1. Applying Markov's inequality yields the stochastic part

\[ \sup_{S_0} \left| \hat{f}_{X,t,W_t,Z} - \mathbb{E} \left[ \hat{f}_{X,t,W_t,Z} \right] \right| = O_p \left( \sqrt{\frac{| \log h_n^\lambda |}{n h_n^\lambda}} \right). \]
We can now proceed analogously to the proof of (ii). By Conditions K2 and S2,
\[
\sup_{s_y} \left| \mathbb{E} [\tilde{g}_{Y_i,X,W,Z} - g_{Y_i,X,W,Z}] \right|
\]
\[
= \sup_{s_y} \int \{ g_{Y_i,X,W,Z}(y - b_n u_y, x + b_n u_x, w + b_n u_1, z + h_n u_2) - g_{Y_i,X,W,Z}(y, x, w, z) \} \cdot k(u_y) k(u_x) \prod_{j=1}^{d_w} k(u_{1j}) \prod_{j=1}^{d_z} k(u_{2j}) \, d(u_y, u_x, u_1, u_2)
\]
\[
\lesssim h_n^2.
\]

The class \( \Psi_n \) defined in (1.B.12) is bounded measurable VC with a set of VC characteristics independent of \( n \) by Lemma 1.2. Since
\[
\sup_{\xi \in \Psi_n} \|\xi\|_\infty \leq \left( \sup_{u \in [-1,1]} |k(u)| \right)^{(1+d_w+d_z)} < \infty
\]
and
\[
\sup_{\xi \in \Psi_n} \text{Var} \xi
\]
\[
\leq \sup_{s_y} \mathbb{E} \left( \psi^2 \left( \frac{y - Y_i}{b_n} \right) K_X^2(X_i, x; b_n) K_W^2(W_i, w; b_n) K_Z^2(Z_i, z; h_n) \right)
\]
\[
\leq \left( \int_{-1}^{1} |k(u)| \, du \right)^2 \sup_{s_y} \mathbb{E} \left( \left( \frac{X_i - x}{b_n} \right)^{d_w} \prod_{j=1}^{d_w} k\left( \frac{W_{ij} - w_{ij}}{b_n} \right) \left( \frac{Z_j - z_j}{h_n} \right)^{d_z} \right)^{(1+d_w+d_z)}
\]
\[
\lesssim b_n^{(1+d_w)} h_n^{d_z} \left( \int_{-1}^{1} |k(u)| \, du \right)^2 \left( \int_{-1}^{1} k^2(u) \, du \right)^{(1+d_w+d_z)} \sup_{x,w,z} f_{X_i,W_i,Z}
\]
\[
\lesssim h_n^\lambda
\]
by Conditions K1 and S1, Theorem 2.1 of Giné and Guillou (2002) suggests
\[
\mathbb{E} \left[ \sup_{s_y} \sum_{i=1}^{n} \left( \psi \left( \frac{y - Y_i t}{b_n} \right) K_X(X_i t, x; b_n) K_W(W_i t, w; b_n) K_Z(Z_i, z; h_n) - \mathbb{E} \left[ \psi \left( \frac{y - Y_i t}{b_n} \right) K_X(X_i t, x; b_n) K_W(W_i t, w; b_n) K_Z(Z_i, z; h_n) \right] \right) \right]
\]
\[
\lesssim |\log h_n^\lambda| + \sqrt{nh_n^\lambda |\log h_n^\lambda|}
\]
\[
\lesssim \sqrt{nh_n^\lambda |\log h_n^\lambda|}
\]
for \( n \) large by Condition R1. Applying Markov’s inequality yields the stochastic part
\[
\sup_{s_y} |\tilde{g}_{Y_i,X,W,Z} - \mathbb{E} [\tilde{g}_{Y_i,X,W,Z}]| = O_P \left( \sqrt{\frac{|\log h_n|}{n h_n^\lambda}} \right).
\]
Lemma 1.4. Let \( \nu_n = \sqrt{\frac{\log h_n}{nh_n^2}} \). Suppose Conditions K1, R1, and S1 hold. For any \( \eta > 0 \), there exists a constant \( c > 0 \) such that

(i) For each \( t \in \{1, 2\} \),
\[
\lim_{n \to \infty} \mathbb{P} \left( \mathbb{P} \left( \sup \left| \hat{f}_{X_1 W_1 Z} - \hat{f}_{X_1 W_1 Z} \right| > c \nu_n \left| \mathcal{D}_n \right. > \eta \right. \right) = 0,
\]
where the supremum is taken over any bounded subset excluding the boundary range of the joint support \( XWZ \).

(ii) For each \( t \in \{1, 2\} \),
\[
\lim_{n \to \infty} \mathbb{P} \left( \mathbb{P} \left( \sup \left| \hat{g}_{Y_1 X_1 W_1 Z} - \hat{g}_{Y_1 X_1 W_1 Z} \right| > c \nu_n \left| \mathcal{D}_n \right. > \eta \right. \right) = 0,
\]
where the supremum is taken over any bounded subset excluding the boundary range of the joint support \( YXWZ \).

(iii) \[
\lim_{n \to \infty} \mathbb{P} \left( \mathbb{P} \left( \sup \left| \hat{f}_{X_1 X_2 W_1 W_2 Z} - \hat{f}_{X_1 X_2 W_1 W_2 Z} \right| > c \nu_n \left| \mathcal{D}_n \right. > \eta \right. \right) = 0,
\]
where the supremum is taken over any bounded subset excluding the boundary range of the joint support \( X^2W^2Z \).

(iv) \[
\lim_{n \to \infty} \mathbb{P} \left( \mathbb{P} \left( \sup \left| \hat{g}_{Y_1 X_1 X_2 W_1 W_2 Z} - \hat{g}_{Y_1 X_1 X_2 W_1 W_2 Z} \right| > c \nu_n \left| \mathcal{D}_n \right. > \eta \right. \right) = 0,
\]
where the supremum is taken over any bounded subset excluding the boundary range of the joint support \( Y^2X^2W^2Z \).

Proof. Since the proofs involve similar arguments, we only present the proof for the first claim.

Fix \( \eta > 0 \) and consider the event \( \Lambda \) that the data \( \mathcal{D}_n \) satisfy
\[
\sup_{S_0} \frac{1}{nh_n^\lambda} \sum_{i=1}^n K^2_X(X_{it}, x; b_n)K^2_W(W_{it}, w; b_n)K^2_Z(Z_i, z; h_n)
- \sup_{S_0} \frac{1}{h_n^\lambda} \mathbb{E} \left[ K^2_X(X_{it}, x; b_n)K^2_W(W_{it}, w; b_n)K^2_Z(Z_i, z; h_n) \right] < 1.
\]

By Lemma 1.2, the class \( \mathcal{K}_n = \{ K_X(\cdot, x; b_n)K_W(\cdot, w; b_n)K_Z(\cdot, z; h_n) : (x, w, z) \in S_0 \} \) is bounded measurable VC with its VC characteristics independent of \( n \). In addition,
\[
\sup_{\xi \in \mathcal{K}_n} \| \xi \|_\infty \leq \left( \sup_{u \in [-1, 1]} \| k(u) \| \right)^{(1 + d_W + d_Z)} < \infty
\]
and on the set $\Lambda$,

$$\sup_{S_0} \text{Var} [K_X(X_{it}^*, x; b_n)K_W(W_{it}^*, w; b_n)K_Z(Z_{it}^*, z; h_n)|\mathcal{D}_n]$$

$$\leq \sup_{S_0} \mathbb{E} [K_X^2(X_{it}^*, x; b_n)K_W^2(W_{it}^*, w; b_n)K_Z^2(Z_{it}^*, z; h_n)|\mathcal{D}_n]$$

$$= \sup_{S_0} \frac{1}{n} \sum_{i=1}^{n} K_X^2(X_{it}, x; b_n)K_W^2(W_{it}, w; b_n)K_Z^2(Z_{it}, z; h_n)$$

$$\leq h_n^\lambda \left\{ 1 + \sup_{S_0} \frac{1}{h_n^\lambda} \mathbb{E} \left[ K_X^2(X_{it}, x; b_n)K_W^2(W_{it}, w; b_n)K_Z^2(Z_{it}, z; h_n) \right] \right\}$$

$$\leq h_n^\lambda \left\{ 1 + \sup_{XWZ} f_{X,W,Z} \left[ \int_{-1}^{1} k^2(u) \, du \right] (1 + d_W + d_Z) \right\}.$$

It follows from Theorem 2.1 of Giné and Guillou (2002) that there are constants $\Delta_1 > 0$ independent of $n$, and $n_1 \in \mathbb{N}$ such that

$$\mathbb{E} \left[ \sup_{S_0} \sum_{i=1}^{n} \left( K_X(X_{it}^*, x; b_n)K_W(W_{it}^*, w; b_n)K_Z(Z_{it}^*, z; h_n) - \mathbb{E} [K_X(X_{it}, x; b_n)K_W(W_{it}, w; b_n)K_Z(Z_{it}, z; h_n)|\mathcal{D}_n] \right) \right] \leq \Delta_1 \left[ |\log h_n| + \sqrt{nh_n^\lambda} |\log h_n| \right].$$

for all $n \geq n_1$. Let $c_0$ be a constant greater than $2\Delta_1/\eta$. Consequently, for any $n \geq n_1$,

$$\mathbb{P} \left( \sup_{S_0} \left| \hat{f}_{X,W,Z} - \hat{f}_{X,W,Z} \right| > c_0 \nu_n |\mathcal{D}_n \right) \leq \frac{\Delta_1}{c_0nh_n^\lambda \nu_n} \left[ |\log h_n| + \sqrt{nh_n^\lambda} |\log h_n| \right]$$

$$\leq \frac{\eta}{2} \left[ \sqrt{\frac{|\log h_n|}{nh_n^\lambda}} + 1 \right]. \quad (1.B.13)$$

We now consider the event $\Lambda^c$, that is, the complement of $\Lambda$. Following the similar arguments in Lemma 1.2, we can show that the class of functions

$$\mathcal{K}_n^2 \equiv \left\{ K_X^2(\cdot, x; b_n)K_W^2(\cdot, w; b_n)K_Z^2(\cdot, z; h_n) : (x, w, z) \in S_0 \right\}$$

is still bounded measurable VC with its VC characteristics independent of $n$. In addition,

$$\sup_{\xi \in \mathcal{K}_n^2} \|\xi\|_\infty \leq \left( \sup_{u \in [-1,1]} |k(u)|^2 \right)^{(1 + d_W + d_Z)} < \infty$$
and

\[ \sup_{\xi \in \mathbb{K}^2_n} \text{Var}(f) \leq \sup_{S_0} \text{Var} \left[ K_X(X_{it}, x; b_n) K_W(W_{it}, w; b_n) K_Z(Z_i, z; h_n) \right] \]

\[ \leq \sup_{S_0} \mathbb{E} \left[ K_X^2(X_{it}, x; b_n) K_W^2(W_{it}, w; b_n) K_Z^2(Z_i, z; h_n) \right] \]

\[ \leq h_n^\lambda \sup_{XWZ} f_{X,W,Z} \left( \int_{-1}^{1} k^2(u) \, du \right)^{(1+d_W+d_Z)} < \infty. \]

It follows from Theorem 2.1 of Giné and Guillou (2002) again that there are constants \( \Delta_2 > 0 \) independent of \( n \), and \( n_2 \in \mathbb{N} \) such that

\[ \mathbb{P}(\Lambda^c) \]

\[ \leq \mathbb{P} \left( \sup_{S_0} \sum_{i=1}^{n} \left[ K_X^2(X_{it}, x_t; b_n) K_W^2(W_{it}, w_t; b_n) K_Z^2(Z_i, z; h_n) \right. \right. \]

\[ - \mathbb{E} \left[ K_X^2(X_{it}, x_t; b_n) K_W^2(W_{it}, w_t; b_n) K_Z^2(Z_i, z; h_n) \right] \left. \right] \geq n h_n^\lambda \right) \]

\[ \leq \Delta_2 \left[ \frac{|\log h_n|}{nh_n^\lambda} + \sqrt{\frac{|\log h_n|}{nh_n^\lambda}} \right] \] (1.B.14)

for all \( n \geq n_2 \).

Combining (1.B.13) and (1.B.14) yields

\[ \limsup_{n \to \infty} \mathbb{P} \left( \sup_{S_0} \left| \hat{f}_{X,W,Z} - \hat{f}_{X,W,Z} \right| > c_0 \nu_n \right) > \eta \]

\[ \leq \limsup_{n \to \infty} \mathbb{P} \left( \left\{ \sup_{S_0} \left| \hat{f}_{X,W,Z} - \hat{f}_{X,W,Z} \right| > c_0 \nu_n \right\} \cap \Lambda \right) \]

\[ + \limsup_{n \to \infty} \mathbb{P}(\Lambda^c) \]

\[ \leq \limsup_{n \to \infty} \left[ \frac{\eta}{2} \left( \sqrt{\frac{|\log h_n|}{nh_n^\lambda}} + 1 > \eta \right) \right] + \limsup_{n \to \infty} \Delta_2 \left[ \frac{|\log h_n|}{nh_n^\lambda} + \sqrt{\frac{|\log h_n|}{nh_n^\lambda}} \right] \]

\[ = 0. \]

by Condition R1.

\[ \Box \]

**Appendix 1.C  Robustness checks**

To check whether the proposed test is robust to the location shift and specific distribution of error terms, we replace the uniform error terms in DGP1 and DGP2 with the two-sided truncated and centered Gaussian error terms, and implement Algorithm 1 with the
For the truncated Gaussian error terms in DGP3 and DGP4, the matching estimator \( \hat{m}_n \) still has downward bias, as shown in Figure 1.4. Since the rejection rates in Tables 1.5-1.8 are similar to those in Tables 1.1-1.4 respectively, the proposed test is robust to the location shift and choice of distributions of error terms.
<table>
<thead>
<tr>
<th></th>
<th>First Region</th>
<th>Second Region</th>
<th>Both Regions</th>
</tr>
</thead>
<tbody>
<tr>
<td>(m) known†</td>
<td>(\tilde{T}<em>{1n}(m + \delta_0) &gt; c^*</em>{1n}(5%, \delta_0))</td>
<td>(\tilde{T}<em>{2n}(m + \delta_0) &gt; c^*</em>{2n}(5%, \delta_0))</td>
<td>(\tilde{T}<em>{1n}(m + \delta_0) &gt; c^*</em>{1n}(5%, \delta_0))</td>
</tr>
<tr>
<td>(m = 0.5^a)</td>
<td>0.020</td>
<td>0.016</td>
<td>0.036</td>
</tr>
<tr>
<td>(n = 250)</td>
<td>0.024</td>
<td>0.030</td>
<td>0.054</td>
</tr>
<tr>
<td>(n = 1000)</td>
<td>0.028</td>
<td>0.032</td>
<td>0.058</td>
</tr>
</tbody>
</table>

| \(m\) unknown‡ | \(\tilde{T}_{1n}(\hat{m}_n + \delta_0) > c^*_{1n}(5\%, \delta_0)\) | \(\tilde{T}_{2n}(\hat{m}_n + \delta_0) > c^*_{2n}(5\%, \delta_0)\) | \(\tilde{T}_{1n}(\hat{m}_n + \delta_0) > c^*_{1n}(5\%, \delta_0)\) |
| \(m = 0.5^a\) | 0.046        | 0.054        | 0.094        |
| \(n = 250\)   | 0.058        | 0.058        | 0.112        |
| \(n = 1000\)  | 0.058        | 0.064        | 0.122        |

| \(m\) unknown§ | \(\tilde{T}_{1n}(\hat{m}_n + \delta_1) > c^*_{1n}(5\%, \delta_0')\) | \(\tilde{T}_{2n}(\hat{m}_n + \delta_0) > c^*_{2n}(5\%, \delta_0')\) | \(\tilde{T}_{1n}(\hat{m}_n + \delta_1) > c^*_{1n}(5\%, \delta_0')\) |
| \(m = 0.5^a\) | 0.034        | 0.038        | 0.070        |
| \(n = 250\)   | 0.042        | 0.040        | 0.082        |
| \(n = 1000\)  | 0.036        | 0.054        | 0.090        |

† For each \(t \in \{1, 2\}\), \(c^*_{tn}(5\%, \delta_0)\) is the 95% quantile of \(\tilde{T}^*_tn(m + \delta_0)\).
‡ For each \(t \in \{1, 2\}\), \(c^*_{tn}(5\%, \delta_0)\) is the 95% quantile of \(\tilde{T}^*_tn(\hat{m}_n + \delta_0)\).
§ For each \(t \in \{1, 2\}\), \(c^*_{tn}(5\%, \delta_0')\) is the 95% quantile of \(\tilde{T}^*_tn(\hat{m}_n + \delta_0')\).

*a The trimmed set is \(Y^2_{\text{trim}}, X^2_{\text{trim}} = [-0.15, 0.41]^2 \times [0.05, 0.95]^2\).
Table 1.6: Rejection Rates for DGP3 with $\bar{a} \in \{0, 1, 2\}$

<table>
<thead>
<tr>
<th>$m$ is known $\dagger$</th>
<th>First Region</th>
<th>Second Region</th>
<th>Both Regions</th>
</tr>
</thead>
<tbody>
<tr>
<td>$m$ is known $\dagger$</td>
<td>$\hat{T}<em>{1n}(m + \delta_0) &gt; c</em>{1a}^*(5%, \delta_0)$</td>
<td>$\hat{T}<em>{2n}(m + \delta_0) &gt; c</em>{2a}^*(5%, \delta_0)$</td>
<td>$\bar{a} = 0^a$</td>
</tr>
<tr>
<td>$n = 250$</td>
<td>0.020</td>
<td>0.016</td>
<td>0.036</td>
</tr>
<tr>
<td>$n = 500$</td>
<td>0.024</td>
<td>0.030</td>
<td>0.054</td>
</tr>
<tr>
<td>$n = 1000$</td>
<td>0.028</td>
<td>0.032</td>
<td>0.058</td>
</tr>
<tr>
<td>$\bar{a} = 1^b$</td>
<td>$\hat{T}_{1n}(\hat{m}<em>n + \delta_0) &gt; c</em>{1a}^*(5%, \delta_0)$</td>
<td>$\hat{T}_{2n}(\hat{m}<em>n + \delta_0) &gt; c</em>{2a}^*(5%, \delta_0)$</td>
<td>$n = 250$</td>
</tr>
<tr>
<td>$n = 500$</td>
<td>0.062</td>
<td>0.060</td>
<td>0.118</td>
</tr>
<tr>
<td>$n = 1000$</td>
<td>0.116</td>
<td>0.126</td>
<td>0.214</td>
</tr>
<tr>
<td>$\bar{a} = 2^c$</td>
<td>$\hat{T}_{1n}(\hat{m}<em>n + \delta_0) &gt; c</em>{1a}^*(5%, \delta_0)$</td>
<td>$\hat{T}_{2n}(\hat{m}<em>n + \delta_0) &gt; c</em>{2a}^*(5%, \delta_0)$</td>
<td>$n = 250$</td>
</tr>
<tr>
<td>$n = 500$</td>
<td>0.990</td>
<td>0.998</td>
<td>1.000</td>
</tr>
<tr>
<td>$n = 1000$</td>
<td>0.986</td>
<td>0.990</td>
<td>1.000</td>
</tr>
<tr>
<td>$m$ is unknown $\ddagger$</td>
<td>$\hat{T}_{1n}(\hat{m}<em>n + \delta_0) &gt; c</em>{1a}^*(5%, \delta_0)$</td>
<td>$\hat{T}_{2n}(\hat{m}<em>n + \delta_0) &gt; c</em>{2a}^*(5%, \delta_0)$</td>
<td>$\bar{a} = 0^a$</td>
</tr>
<tr>
<td>$n = 250$</td>
<td>0.046</td>
<td>0.054</td>
<td>0.094</td>
</tr>
<tr>
<td>$n = 500$</td>
<td>0.058</td>
<td>0.058</td>
<td>0.112</td>
</tr>
<tr>
<td>$n = 1000$</td>
<td>0.058</td>
<td>0.064</td>
<td>0.122</td>
</tr>
<tr>
<td>$\bar{a} = 1^b$</td>
<td>$\hat{T}_{1n}(\hat{m}<em>n + \delta_0) &gt; c</em>{1a}^*(5%, \delta_0)$</td>
<td>$\hat{T}_{2n}(\hat{m}<em>n + \delta_0) &gt; c</em>{2a}^*(5%, \delta_0)$</td>
<td>$n = 250$</td>
</tr>
<tr>
<td>$n = 500$</td>
<td>0.218</td>
<td>0.248</td>
<td>0.390</td>
</tr>
<tr>
<td>$n = 1000$</td>
<td>0.238</td>
<td>0.276</td>
<td>0.426</td>
</tr>
<tr>
<td>$\bar{a} = 2^c$</td>
<td>$\hat{T}_{1n}(\hat{m}<em>n + \delta_0) &gt; c</em>{1a}^*(5%, \delta_0)$</td>
<td>$\hat{T}_{2n}(\hat{m}<em>n + \delta_0) &gt; c</em>{2a}^*(5%, \delta_0)$</td>
<td>$n = 250$</td>
</tr>
<tr>
<td>$n = 500$</td>
<td>1.000</td>
<td>0.998</td>
<td>1.000</td>
</tr>
<tr>
<td>$n = 1000$</td>
<td>0.998</td>
<td>0.996</td>
<td>1.000</td>
</tr>
</tbody>
</table>

$\dagger$ For each $t \in \{1, 2\}$, $c_{ta}^*(5\%, \delta_0)$ is the 95% quantile of $\hat{T}_{tn}(m + \delta_0)$.
$\ddagger$ For each $t \in \{1, 2\}$, $c_{ta}^*(5\%, \delta_0)$ is the 95% quantile of $\hat{T}_{tn}(\hat{m}_n + \delta_0)$.
a The trimmed set is $\mathcal{Y}_{2\text{trim}}^2 \times \mathcal{X}_{2\text{trim}}^2 = [-0.15, 0.41]^2 \times [0.05, 0.95]^2$.
b The trimmed set is $\mathcal{Y}_{2\text{trim}}^2 \times \mathcal{X}_{2\text{trim}}^2 = [-0.41, 0.15]^2 \times [0.05, 0.95]^2$.
c The trimmed set is $\mathcal{Y}_{2\text{trim}}^2 \times \mathcal{X}_{2\text{trim}}^2 = [-0.83, 0.06]^2 \times [0.05, 0.95]^2$. 
<table>
<thead>
<tr>
<th></th>
<th>First Period</th>
<th>Second Period</th>
<th>Both Periods</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>GSV</strong></td>
<td>$\hat{T}<em>{1n}^{GSV}&gt;c</em>{1n}^{GSV}(5%)$</td>
<td>$\hat{T}<em>{2n}^{GSV}&gt;c</em>{2n}^{GSV}(5%)$</td>
<td>$\hat{T}<em>{1n}^{GSV}&gt;c</em>{1n}^{GSV}(5%)$ or $\hat{T}<em>{2n}^{GSV}&gt;c</em>{2n}^{GSV}(5%)$</td>
</tr>
<tr>
<td>$n=250$</td>
<td>0.068</td>
<td>0.064</td>
<td>0.132</td>
</tr>
<tr>
<td>$n=500$</td>
<td>0.070</td>
<td>0.046</td>
<td>0.110</td>
</tr>
<tr>
<td>$n=1000$</td>
<td>0.050</td>
<td>0.062</td>
<td>0.110</td>
</tr>
</tbody>
</table>

| **PI-A**         | $\hat{T}_{1n}^{PI-A}>c_{1n}^{PI-A}(5\%)$ | $\hat{T}_{2n}^{PI-A}>c_{2n}^{PI-A}(5\%)$ | $\hat{T}_{1n}^{PI-A}>c_{1n}^{PI-A}(5\%)$ or $\hat{T}_{2n}^{PI-A}>c_{2n}^{PI-A}(5\%)$ |
| $n=250$          | 0.112        | 0.120         | 0.226        |
| $n=500$          | 0.078        | 0.080         | 0.152        |
| $n=1000$         | 0.052        | 0.032         | 0.082        |

<table>
<thead>
<tr>
<th></th>
<th>First Region</th>
<th>Second Region</th>
<th>Both Regions</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>SM</strong></td>
<td>$\hat{T}<em>{1n}(m+\delta</em>{0})&gt;c_{1n}^{5%(m,\delta_{0})}$</td>
<td>$\hat{T}<em>{2n}(m+\delta</em>{0})&gt;c_{2n}^{5%(m,\delta_{0})}$</td>
<td>$\hat{T}<em>{1n}(m+\delta</em>{0})&gt;c_{1n}^{5%(m,\delta_{0})}$ or $\hat{T}<em>{2n}(m+\delta</em>{0})&gt;c_{2n}^{5%(m,\delta_{0})}$</td>
</tr>
<tr>
<td>$m = 0.2$ is known†</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$n=250$</td>
<td>0.452</td>
<td>0.440</td>
<td>0.650</td>
</tr>
<tr>
<td>$n=500$</td>
<td>0.612</td>
<td>0.686</td>
<td>0.862</td>
</tr>
<tr>
<td>$n=1000$</td>
<td>0.800</td>
<td>0.812</td>
<td>0.966</td>
</tr>
<tr>
<td><strong>SM</strong></td>
<td>$\hat{T}<em>{1n}(\hat{m}</em>{n}+\delta_{0})&gt;c_{1n}^{5%(\hat{m}<em>{n},\delta</em>{0})}$</td>
<td>$\hat{T}<em>{2n}(\hat{m}</em>{n}+\delta_{0})&gt;c_{2n}^{5%(\hat{m}<em>{n},\delta</em>{0})}$</td>
<td>$\hat{T}<em>{1n}(\hat{m}</em>{n}+\delta_{0})&gt;c_{1n}^{5%(\hat{m}<em>{n},\delta</em>{0})}$ or $\hat{T}<em>{2n}(\hat{m}</em>{n}+\delta_{0})&gt;c_{2n}^{5%(\hat{m}<em>{n},\delta</em>{0})}$</td>
</tr>
<tr>
<td>$m = 0.2$ is unknown‡</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$n=250$</td>
<td>0.486</td>
<td>0.454</td>
<td>0.668</td>
</tr>
<tr>
<td>$n=500$</td>
<td>0.698</td>
<td>0.676</td>
<td>0.864</td>
</tr>
<tr>
<td>$n=1000$</td>
<td>0.824</td>
<td>0.788</td>
<td>0.956</td>
</tr>
</tbody>
</table>

---

* The supremum test statistics $\hat{T}_{1n}^{GSV}$ and $\hat{T}_{2n}^{GSV}$ of Ghosal et al. (2000) are calculated based on the first-period and second-period data, respectively.

* The plug-in adaptive test statistics $\hat{T}_{1n}^{PI-A}$ and $\hat{T}_{2n}^{PI-A}$ of Chetverikov (2017) are calculated based on the first-period and second-period data, respectively.

* The trimmed set is $\mathcal{Y}_{\text{trim}}^{2} = \mathcal{X}_{\text{trim}}^{2} = \pm 4.41, 1.27 | 0.05, 0.95 | 2$.

† For each $t \in \{1, 2\}$, $c^{5\%(m,\delta_{0})}$ is the 95% quantile of $\hat{T}_{tn}^{*}(m+\delta_{0})$.

‡ For each $t \in \{1, 2\}$, $c^{5\%(\hat{m}_{n},\delta_{0})}$ is the 95% quantile of $\hat{T}_{tn}^{*}(\hat{m}_{n}+\delta_{0})$. 

---

67
Table 1.8: Rejection Rates for DGP4 with \( m \in \{0.2, 0.45, 0.7\} \)

<table>
<thead>
<tr>
<th>( m ) is known(^\dagger )</th>
<th>First Region</th>
<th>Second Region</th>
<th>Both Regions</th>
</tr>
</thead>
<tbody>
<tr>
<td>( m = 0.2^a )</td>
<td>( \tilde{T}<em>{1n}(m+\delta_0) &gt; c</em>{1n}^* (5%, \delta_0) )</td>
<td>( \tilde{T}<em>{2n}(m+\delta_0) &gt; c</em>{2n}^* (5%, \delta_0) )</td>
<td>( \tilde{T}<em>{1n}(m+\delta_0) &gt; c</em>{1n}^* (5%, \delta_0) ) or ( \tilde{T}<em>{2n}(m+\delta_0) &gt; c</em>{2n}^* (5%, \delta_0) )</td>
</tr>
<tr>
<td>( n = 250 )</td>
<td>0.452</td>
<td>0.440</td>
<td>0.650</td>
</tr>
<tr>
<td>( n = 500 )</td>
<td>0.612</td>
<td>0.686</td>
<td>0.862</td>
</tr>
<tr>
<td>( n = 1000 )</td>
<td>0.800</td>
<td>0.812</td>
<td>0.966</td>
</tr>
<tr>
<td>( m = 0.45^b )</td>
<td>( \tilde{T}<em>{1n}(m+\delta_0) &gt; c</em>{1n}^* (5%, \delta_0) )</td>
<td>( \tilde{T}<em>{2n}(m+\delta_0) &gt; c</em>{2n}^* (5%, \delta_0) )</td>
<td>( \tilde{T}<em>{1n}(m+\delta_0) &gt; c</em>{1n}^* (5%, \delta_0) )</td>
</tr>
<tr>
<td>( n = 250 )</td>
<td>0.080</td>
<td>0.062</td>
<td>0.126</td>
</tr>
<tr>
<td>( n = 500 )</td>
<td>0.104</td>
<td>0.078</td>
<td>0.172</td>
</tr>
<tr>
<td>( n = 1000 )</td>
<td>0.114</td>
<td>0.086</td>
<td>0.192</td>
</tr>
<tr>
<td>( m = 0.7^c )</td>
<td>( \tilde{T}<em>{1n}(m+\delta_0) &gt; c</em>{1n}^* (5%, \delta_0) )</td>
<td>( \tilde{T}<em>{2n}(m+\delta_0) &gt; c</em>{2n}^* (5%, \delta_0) )</td>
<td>( \tilde{T}<em>{1n}(m+\delta_0) &gt; c</em>{1n}^* (5%, \delta_0) )</td>
</tr>
<tr>
<td>( n = 250 )</td>
<td>0.052</td>
<td>0.046</td>
<td>0.094</td>
</tr>
<tr>
<td>( n = 500 )</td>
<td>0.070</td>
<td>0.046</td>
<td>0.114</td>
</tr>
<tr>
<td>( n = 1000 )</td>
<td>0.078</td>
<td>0.044</td>
<td>0.118</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>( m ) is unknown(^\ddagger )</th>
<th>First Region</th>
<th>Second Region</th>
<th>Both Regions</th>
</tr>
</thead>
<tbody>
<tr>
<td>( m = 0.2^a )</td>
<td>( \tilde{T}<em>{1n}(m_n+\delta_0) &gt; c</em>{1n}^* (5%, \delta_0) )</td>
<td>( \tilde{T}<em>{2n}(m_n+\delta_0) &gt; c</em>{2n}^* (5%, \delta_0) )</td>
<td>( \tilde{T}<em>{1n}(m_n+\delta_0) &gt; c</em>{1n}^* (5%, \delta_0) ) or ( \tilde{T}<em>{2n}(m_n+\delta_0) &gt; c</em>{2n}^* (5%, \delta_0) )</td>
</tr>
<tr>
<td>( n = 250 )</td>
<td>0.486</td>
<td>0.454</td>
<td>0.668</td>
</tr>
<tr>
<td>( n = 500 )</td>
<td>0.698</td>
<td>0.676</td>
<td>0.864</td>
</tr>
<tr>
<td>( n = 1000 )</td>
<td>0.824</td>
<td>0.788</td>
<td>0.956</td>
</tr>
<tr>
<td>( m = 0.45^b )</td>
<td>( \tilde{T}<em>{1n}(m_n+\delta_0) &gt; c</em>{1n}^* (5%, \delta_0) )</td>
<td>( \tilde{T}<em>{2n}(m_n+\delta_0) &gt; c</em>{2n}^* (5%, \delta_0) )</td>
<td>( \tilde{T}<em>{1n}(m_n+\delta_0) &gt; c</em>{1n}^* (5%, \delta_0) )</td>
</tr>
<tr>
<td>( n = 250 )</td>
<td>0.092</td>
<td>0.098</td>
<td>0.176</td>
</tr>
<tr>
<td>( n = 500 )</td>
<td>0.116</td>
<td>0.100</td>
<td>0.210</td>
</tr>
<tr>
<td>( n = 1000 )</td>
<td>0.152</td>
<td>0.110</td>
<td>0.230</td>
</tr>
<tr>
<td>( m = 0.7^c )</td>
<td>( \tilde{T}<em>{1n}(m_n+\delta_0) &gt; c</em>{1n}^* (5%, \delta_0) )</td>
<td>( \tilde{T}<em>{2n}(m_n+\delta_0) &gt; c</em>{2n}^* (5%, \delta_0) )</td>
<td>( \tilde{T}<em>{1n}(m_n+\delta_0) &gt; c</em>{1n}^* (5%, \delta_0) )</td>
</tr>
<tr>
<td>( n = 250 )</td>
<td>0.046</td>
<td>0.046</td>
<td>0.088</td>
</tr>
<tr>
<td>( n = 500 )</td>
<td>0.082</td>
<td>0.038</td>
<td>0.114</td>
</tr>
<tr>
<td>( n = 1000 )</td>
<td>0.086</td>
<td>0.046</td>
<td>0.128</td>
</tr>
</tbody>
</table>

\(^\dagger \) For each \( t \in \{1, 2\} \), \( c_{tn}^* (5\%, \delta_0) \) is the 95% quantile of \( \hat{T}_{tn}^* (m + \delta_0) \).

\(^\ddagger \) For each \( t \in \{1, 2\} \), \( c_{tn}^* (5\%, \delta_0) \) is the 95% quantile of \( \hat{T}_{tn}^* (m_n + \delta_0) \).

\(^a\) The trimmed set is \( Y_{2\text{trim}}^2 X_{2\text{trim}}^2 = [-0.44, 1.27]^2 \times [0.05, 0.95]^2 \).

\(^b\) The trimmed set is \( Y_{2\text{trim}}^2 X_{2\text{trim}}^2 = [-0.47, 1.28]^2 \times [0.05, 0.95]^2 \).

\(^c\) The trimmed set is \( Y_{2\text{trim}}^2 X_{2\text{trim}}^2 = [-0.53, 1.30]^2 \times [0.05, 0.95]^2 \).
Chapter 2

Counterfactual Inference in Duration Models with Random Censoring

Abstract

A counterfactual Kaplan-Meier estimator is proposed to evaluate policy effects of an exogenous manipulation of covariates on a duration outcome that is subject to random censoring. Under some regularity conditions, we establish the joint weak convergence of the proposed counterfactual estimator and the unconditional Kaplan-Meier (1958) estimator. Applying the functional delta method, we make inference on the cumulative hazard policy effect, that is, the change of duration dependence in response to a counterfactual policy. We also evaluate the finite sample performance of the proposed counterfactual estimation method in a Monte Carlo study.

2.1 Introduction

Policy evaluation is one of the important areas in social science. Counterfactual analysis is an approach that provides policy recommendations when a policy is not implemented yet or when a similar quasi-experiment is infeasible. Recent studies on counterfactual analysis, for example Rothe (2010) and Chernozhukov et al. (2013), emphasize the unconditional distributional effect of an exogenous manipulation of covariates on an outcome variable of interest. Methods in these studies are usually based on data that are completely observed; however, sampling schemes may generate incomplete data and thus restrict their applicability. For example, duration data on unemployment spells, collected by the Current Population Survey, are commonly believed to be subject to right censoring, as explained by Kiefer (1988).
The main objectives of this chapter are to estimate the unconditional distribution of a duration variable affected by a counterfactual policy that exogenously manipulates covariates, and to evaluate associated policy effects by the comparison between the counterfactual and original unconditional distribution of the duration variable. Specifically, we consider a nonseparable model

$$T = \varphi(X, \varepsilon),$$

(2.1)

where $T$ is a nonnegative duration variable of interest, $X$ is a $d$-dimensional vector of time-invariant covariates, $\varepsilon$ is individual unobserved heterogeneity in an arbitrary measurable space of unrestricted dimensionality, and $\varphi$ is a structural function that is unknown to researchers. In addition, the right censoring may make $T$ unobserved; instead, the observable data are the vector $X$ of covariates,

$$Y = \min\{T, C\} \quad \text{and} \quad \delta = \mathbb{1}_{\{T \leq C\}},$$

(2.2)

where $C$ is a censoring random variable, which is only observed for censored observations, and $\mathbb{1}_{\{\cdot\}}$ is an indicator function. Policy makers consider the counterfactual scenario that exogenously changes $X$ to $X^*$ and leads to the counterfactual duration variable

$$T^* = \varphi(X^*, \varepsilon),$$

(2.3)

and attempt to evaluate the policy effect

$$\nu(F_{T^*}) - \nu(F_T),$$

where $F_T$ and $F_{T^*}$ are the cumulative distribution functions (CDFs) of $T$ and $T^*$, respectively, and $\nu$ is some functional defined on the collection of all CDFs. Such an effect $\nu(F_{T^*}) - \nu(F_T)$ may be important in policy evaluation. For example, although unemployment insurance benefits may smooth the income fluctuation of the unemployed, it may discourage the unemployed from searching for jobs. Therefore, policy makers would be interested in the effect of reducing wage replacement ratio on the cumulative hazard rate of unemployment spells. In this case, $T$ is the unemployment duration, $X$ is the wage replacement ratio, and the functional $\nu$ is a map from a CDF to its cumulative hazard function, that is, $\nu: F \mapsto \int_{[0,1]} \frac{1}{1-F} \, dF$.

We propose a nonparametric estimation method of the unconditional CDF $F_T$ arising from an exogenous manipulation of covariates $X$ on the duration variable $T$. This proposed nonparametric estimation method allows researchers to conduct

\footnote{For any càdlàg function $F$, we write $F^-$ for its left-continuous version, that is, $F^-(t) \equiv \lim_{s \uparrow t} F(s)$.}
a counterfactual policy analysis, rather than just a descriptive analysis, of duration data.² Specifically, we evaluate \( \nu(F_{T^*}) - \nu(F_T) \) by replacing \( F_{T^*} \) and \( F_T \) with their nonparametric estimator, respectively. On the one hand, we construct the unconditional Kaplan-Meier (1958) estimator \( \hat{F}_{T^*} \). On the other hand, under regularity conditions, the unconditional CDF of \( T^* \) is recovered by \( F_{T^*}(t) = \mathbb{E}(F_{T|X}(t|X^*)) \) where \( F_{T|X} \) is the conditional CDF of \( T \) given \( X \); thus, we follow the analogy principle to propose a two-stage fully nonparametric estimator of the counterfactual CDF of \( T^* \). In particular, we first construct a variant of Beran’s (1981) conditional Kaplan-Meier estimator \( \hat{F}_{T|X} \) and then take average of \( \hat{F}_{T|X} \) with respect to the empirical distribution of \( X^* \) to obtain a counterfactual estimator \( \hat{F}_{T^*} \). The first-stage nonparametric estimation can avoid the misspecification of the conditional CDF, which is emphasized in Rothe and Wied (2013). Moreover, the proposed first-stage estimator, instead of the kernel CDF estimator in Rothe (2010), is essential to avoid the estimation bias in the presence of censoring. Indeed, our simulation experiments show that the proposed estimator \( \hat{F}_{T^*} \), compared with Rothe’s counterfactual CDF estimator, has smaller mean integrated absolute error (MIAE) and root mean integrated squared error (RMISE) when duration data are subject to censoring.

To establish the validity of the proposed approach, we show that under some regularity conditions, the vector \( (\hat{F}_{T^*} - F_{T^*}, \hat{F}_T - F_T)^\top \) converges weakly to a two dimensional centered Gaussian process over a specific compact subset of \( \mathbb{R}^2_+ \) at the rate \( \sqrt{n} \). This convergence rate avoids the curse of dimensionality even though the first-stage estimator \( \hat{F}_{T|X} \) converges at a rate less than \( \sqrt{n} \). Applying the functional delta method, we can obtain the asymptotic distribution of the counterfactual policy effect \( \sqrt{n}(\nu(\hat{F}_{T^*}) - \nu(\hat{F}_T)) \), which allows us to evaluate the effect of counterfactual policy intervention on the cumulative hazard function.

The proposed method also complements the literature on decomposition methods. Decomposition methods are usually used to explain the difference in unconditional distributional features of an outcome variable across two different demographic groups or time periods. The between-group difference is usually decomposed into a structure effect and a composition effect.³ In the presence of complete data, Rothe (2010) proposes a two-stage fully nonparametric estimation of the composition effect, whereas Chernozhukov et al. (2013) develop a two-stage semiparametric estimation of this effect by either

---

² Lancaster (1992) and Cameron and Trivedi (2005) indicate that the nonparametric Kaplan-Meier estimation is traditionally viewed as a descriptive analysis.

³ A structure effect arises because structural functions are different between two groups, and a composition effect reflects the differences in covariates between two groups. The early development in decomposition methods is well surveyed by Fortin, Lemieux, and Firpo (2011). Recently, Rothe (2015) further investigates a detailed decomposition, which attributes the composition effect to each covariate.
distribution regression or quantile regression. Taking random censoring into account, García-Suaza (2016) studies the effect based on the proportional hazard specification. In contrast, the method in this chapter is fully nonparametric; additionally, as explained by Rothe (2010), we can regard $X^*$ as observable covariates of a different group, and $
u(F_{T^*}) - 
u(F_T)$ as the composition effect in the setup (2.1)-(2.3) of random censoring.

Throughout this chapter, all random variables are defined on the same probability space $(\Omega, \mathcal{A}, P)$. We denote $\mathbb{D}[c_1, c_2]$ and $l^\infty[c_1, c_2]$ by the set of càdlàg and bounded functions defined on the interval $[c_1, c_2]$, respectively. We write $\Rightarrow$ for weak convergence in a function space equipped with the uniform norm, and $a \wedge b$ for the minimum of $a$ and $b$. We also denote the density of $X$ by $m$, and the density of $X^*$ by $m^*$. For a generic random variable $U$, we write $F_U$ for the CDF of $U$, $f_U$ for the derivative of $F_U$, $F_{U|X}$ for the conditional CDF of $U$ given $X$, and $f_{U|X}(u|x)$ for the derivative with respect to $u$ of $F_{U|X}(u|x)$; additionally, let $F_{U}(u) = P(U \leq u, \delta = 1)$ and $F_{U|X}(u|x) = P(U \leq u, \delta = 1|X = x)$. We assume that $T, C, X,$ and $X^*$ are absolutely continuous random variables. The absolute continuity of the duration variable $T$ is reasonable because $T$ is expected to be generated by a transition process, which is usually modeled in continuous time. (See Cameron and Trivedi (2005) and Florens, Fougeré, and Mouchart (2008) for example.) Furthermore, we only consider absolutely continuous covariates for ease of exposition because the proposed estimation method can be revised to include discrete covariates. Alternatively, in the case of a binary policy variable, Sant’Anna (2016) extends the method of Kaplan-Meier integrals and studies various treatment effects when the outcome may be right censored.

The remainder of this chapter is organized as follows. Section 2.2 discusses the setup of duration analysis, the objects of interest, and the counterfactual Kaplan-Meier estimator. Section 2.3 shows the asymptotic theory of the proposed estimator and statistical inference on the associated policy effects. Section 2.4 presents the results of Monte Carlo simulation. Section 2.5 concludes. Technical proofs are deferred to Appendix 2.A.

2.2 Model and estimation

2.2.1 Setup and objects of interest

The flexible duration model in (2.1) can avoid several types of model misspecification. First, Lu and White (2014) point out that the nonseparability of $\varepsilon$ enables treatment effect and marginal effect could depend on unobservable heterogeneity. In addition, the unrestricted dimensionality of $\varepsilon$ can avoid incorrect inference due to the inclusion of
limited heterogeneity, as argued by Browning and Carro (2007). Hoderlein and Mammen (2009) also indicate that $\varepsilon$ can be viewed as an element of an infinitely dimensional function space; for example, it could be individual preference for leisure in the analysis of unemployment spells. Finally, both the marginal distribution of $\varepsilon$ and the conditional distribution of $T$ given $\varepsilon$ are unspecified to avoid inappropriate inference caused by parametric assumptions.\footnote{Lancaster (1992) documents many alternatives of parametric assumptions about the hazard function in mixture models. Parametric specification of the heterogeneity distribution and duration dependence is, however, a well-known issue in econometrics. See the discussion in Hausman and Woutersen (2014).}

The random censoring feature in (2.2) is prevalent in duration analysis and usually arises because of sampling schemes, for example, a random failure to follow up an individual during the study period. We refer readers to Moore (2016) for more underlying reasons of random censoring. In this chapter, we consider the simple counterfactual scenario that policy intervention does not affect the structural function $\varphi$ in (2.3); however, we allow a change in the censoring variable $C$ after policy intervention.\footnote{As suggested in Fortin et al. (2011), policy intervention may result in an alternative structural function $\varphi^*$ in general equilibrium.}

A counterfactual policy that changes the duration variable from $T$ to $T^*$ yields the distribution policy effect

$$\Delta F(t) \equiv F_{T^*}(t) - F_T(t).$$

For instance, since the shape of the distribution of unemployment spells may affect the government expenditures on unemployment insurance, the distribution policy effect matters for policy makers concerning fiscal deficits. In addition, the literature on duration models especially emphasizes the duration dependence, that is, the shape of the hazard function.\footnote{The hazard function of a nonnegative duration variable $T$ is defined as

$$\lambda_T(t) \equiv \lim_{u \to 0} \frac{\mathbb{P}(t \leq T < t + u | T \geq t)}{u}.$$}

The change of duration dependence in response to a counterfactual policy can be answered by the cumulative hazard policy effect

$$\Delta \Lambda(t) \equiv \Lambda_{T^*}(t) - \Lambda_T(t),$$

where $\Lambda_{T^*}(t) = \int_0^t \frac{F_{T^*}(du)}{1-F_{T^*}(u)}$ and $\Lambda_T(t) = \int_0^t \frac{F_T(du)}{1-F_T(u)}$ are the cumulative hazard functions of $T^*$ and $T$, respectively. In the case of unemployment spells, policy makers would be interested in the cumulative hazard policy effect because it would evaluate whether a counterfactual policy is beneficial for a target group, for example the long-term unemployed, to escape the unemployment trap. Other counterfactual policy effects,
such as quantile policy effect and Lorenz curve policy effect, may also be of interest. See Bhattacharya (2007) and Rothe (2010) for treatment of these and further examples. When the objects of interest are the aforementioned policy effects, the identification of $\varphi$ is not necessary, as indicated by Rothe (2010); thus, we maintain the flexible specification of the structural function in (2.1).

### 2.2.2 Nonparametric identification and estimation

Since a counterfactual policy effect can be generally written as $\nu(F_{T^*}) - \nu(F_T)$ for some specific functional $\nu$, we start by identifying the CDFs $F_T$ and $F_{T^*}$. We first introduce the following assumptions.

**Assumption D (Data)**

- **D1** Both $\{(Y_i, T_i, C_i, \delta_i, X_i)\}_{i=1}^n$ and $\{X_j^*\}_{j=1}^{n^*}$ are independent and identically distributed across individuals.

- **D2** (i) $\{(Y_i, \delta_i, X_i)\}_{i=1}^n$ are observable; (ii) $\{X_j^*\}_{j=1}^{n^*}$ are observable and $n^* = n$.

Assumption D1 is common in models of cross-sectional data. Assumption D2(i) is also common in duration models where researchers know whether the observed duration variable is censored. Assumption D2(ii) is innocuous when we consider the counterfactual policy that shifts $X$ to $X^* = \pi(X)$ for some measurable function $\pi$, whereas this assumption is imposed for convenience when we treat $X^*$ as observable covariates of a different group in the analysis of the composition effect.

**Assumption I (Identification)**

- **I1** $T$ and $C$ are independent.

- **I2** $T$ and $C$ are conditionally independent given $X$.

- **I3** $\varepsilon$ is independent of both $X$ and $X^*$.

- **I4** The support of $X^*$ is a subset of the support of $X$.

Assumptions I1 and I2 are commonly imposed in survival analysis, for example Lancaster (1992) and Kalbfleisch and Prentice (2002) for Assumption I1 and Dabrowska (1989), Iglesias-Pérez and González-Manteiga (1999), and Gneyou (2014) for Assumption I2. Assumption I1 ensures that the random censoring is non-informative; that is, Matzkin (2003) provides conditions such that in the absence of censoring, the structural function $\varphi$ can be identified if $\varphi(x, \varepsilon)$ is strictly increasing in unobserved scalar heterogeneity $\varepsilon$ for each $x$.  

---

7 Matzkin (2003) provides conditions such that in the absence of censoring, the structural function $\varphi$ can be identified if $\varphi(x, \varepsilon)$ is strictly increasing in unobserved scalar heterogeneity $\varepsilon$ for each $x$.  

---
C does not provide any information about T, and vice versa. Assumption I2 holds if random censoring is non-informative when covariates are controlled. Note that neither Assumption I1 nor Assumption I2 is stronger.\(^8\) If C is independent of (X, T), then Assumptions I1 and I2 are satisfied; however, these two assumptions are not sufficient for the independence between C and X.\(^9\) Assumptions I3 and I4 are imposed for counterfactual analysis as in Rothe (2010). Assumption I3 requires that all covariates are exogenous and thus may be strong in some empirical studies. If we are interested in the effects arising from the manipulation of some policy variables, this assumption can be weaken by conditional exogeneity of \(\varepsilon\) given observable covariates. To be precise, let \(X = (X_p, X_c)^\top\) where \(X_p\) and \(X_c\) are the vector of policy variables and vector of covariates, respectively. Policy intervention changes \(X_p\) to \(X_p^*\) but keeps \(X_c\) unchanged. Proposition 2.1 below is still valid if Assumption I3 is replaced with the assumption that \(\varepsilon\) is independent of \((X_p, X_p^*)\) conditional on \(X_c\).\(^10\) Assumptions I2 and I3 imply that censoring occurs exogenously provided that \(\varphi(x, \cdot)\) is invertible for all \(x\).\(^11\) Since nonparametric analyses of counterfactual policy effects resulting from an extrapolation of covariates may be invalid, we impose the overlap condition in Assumption I4.

Assumption I guarantees the identification of \((F_T, F_T^*)\) over a subset of \(\mathbb{R}_+^2\). Stute and Wang (1993) show that under Assumption I1, \(F_T(t)\) is identified for each \(t < \tau \equiv \inf\{t : F_Y(t) = 1\}\). Under Assumptions I3 and I4, we can express \(F_T^*\) as the population average, taken with respect to the distribution of \(X^*\), of the conditional CDF of \(T\) given \(X\); to be precise, \(F_T^*(t) = \mathbb{E}(F_T|X(t|X^*))\). The identification of \(F_T^*(t)\) thus follows that of \(F_T|X(t|x)\), and the latter is achieved under Assumption I2 for \((t, x) \in [0, \tau) \times \mathbb{R}^d\). Alternatively, the joint CDF of \((T, X)\) can be identified on \([0, \tau) \times \mathbb{R}^d\) by replacing Assumption I2 with the assumption that \(\delta\) and \(X\) are conditionally independent given \(T\); that is, given the duration, covariates provide no further information on whether censoring occurs.\(^12\) This assumption is imposed in recent studies on duration analysis,

\(^8\) See the examples on page 65 of Stoyanov (2014).
\(^9\) The independence between \(C\) and \(X\) holds if Assumptions I1 and I2 hold and the family of distributions of \(T\) given \(X\) is boundedly complete; that is, for a bounded function \(g\), \(\mathbb{E}[g(T)|X] = 0\) almost surely implies \(g(T) = 0\) almost surely. See the discussion in Dawid (1998).
\(^10\) Similarly, Assumptions I3 can be relaxed by the control function approach proposed by Blundell and Powell (2003) and Imbens and Newey (2009). The application of the control function approach is however beyond the scope this chapter. See Lee (2015) for the analysis of counterfactual effects by the control function approach in the absence of random censoring.
\(^11\) Suppose that invertibility of \(\varphi(x, \cdot)\) holds for all \(x\). Assumption I2 implies that \(\varepsilon\) and \(C\) are conditionally independent given \(X\) by Lemmas 4.1 and 4.2 of Dawid (1979). If moreover Assumption I3 holds, then \(\varepsilon\) is independent of \((C, X)\) by Lemma 4.2 of Dawid (1979).
\(^12\) Suppose that Assumption I1 holds. If \(\delta\) and \(X\) are conditionally independent given \(T\), then the joint distribution of \((T, X)\) on \([0, \tau) \times \mathbb{R}^d\) can be recovered by

\[
\mathbb{P}(T \leq t, X \leq x) = \int I_{[s \leq t]} I_{[s \leq x]} \exp \left\{ \int_0^s \frac{H^0_Y(dy)}{1 - F_Y(g)} \right\} H^*_Y \mathbb{I}_X(ds, dz)
\]
for example Sant’Anna (2016, 2017), García-Suaza (2016), and references cited therein. We summarize the discussion in the following proposition.

**Proposition 2.1.** Suppose that Assumptions D1 and D2 hold. Under Assumption I1, $F_{T^*}(t)$ is identified for $t \in [0, \tau)$. If in addition Assumption I2 holds, then $F_{T|X}(t|x)$ is identified for $(t, x) \in [0, \tau) \times \mathbb{R}^d$. Moreover, if Assumptions I3 and I4 are also satisfied, we have $F_{T^*}(t) = \mathbb{E}(F_{T|X}(t|X^*))$ for $t \in [0, \tau)$.

Proposition 2.1 suggests that we follow the analogy principle to construct an estimator of $F_{T^*}(t)$ by

$$
\hat{F}_{T^*}(t; h_n) = \frac{1}{n} \sum_{i=1}^{n} \hat{F}_{T|X}(t|X_i^*; h_n),
$$

where $\hat{F}_{T|X}$ is the variant of Beran’s (1981) conditional Kaplan-Meier estimator, that is

$$
\hat{F}_{T|X}(t|x; h) = 1 - \prod_{j=1}^{n} \exp \left\{ - \sum_{\ell=1}^{n} \frac{1}{\mathbb{1}_{[Y_j \leq t, \delta_j = 1]} B_{n_\ell}(x; h)} \right\},
$$

where $\{B_{n_\ell}(x; h)\}_{\ell=1}^{n}$ are appropriate weights and $h$ is a tuning parameter. Different choices of weights are documented in the literature on the conditional Kaplan-Meier estimator. In this chapter, we construct the counterfactual Kaplan-Meier estimator in (2.4) based on the Nadaraya-Watson weights

$$
B_{n_\ell}(x; h_n) = \frac{K \left( \frac{x - X_{\ell}}{h_n} \right)}{\sum_{i=1}^{n} K \left( \frac{x - X_{i}}{h_n} \right)}, \quad \ell = 1, 2, \ldots, n,
$$

for some kernel function $K$ and bandwidth $h_n$. For ease of notation, we suppress the dependence on $h_n$ for $\hat{F}_{T^*}$ and $\hat{F}_{T|X}$ hereafter. To estimate $F_T(t)$, we adopt the unconditional Kaplan-Meier (1958) estimator

$$
\hat{F}_T(t) = 1 - \prod_{j=1}^{n} \left( \frac{n - j}{n - j + 1} \right)^{1_{[Y_j \leq t, \delta_j = 1]}}
$$

where $\{(Y_j, \delta_j)\}_{j=1}^{n}$ are the $n$ pairs of observations ordered on the order statistics of $\{Y_j\}_{j=1}^{n}$.

where $H^T_{Y}(y) = \mathbb{P}(Y \leq y, \delta = 0)$ and $H^T_{Y,X}(y, x) = \mathbb{P}(Y \leq y, X \leq x, \delta = 1)$. The availability of data on $\{(Y, \delta, X)\}_{i=1}^{n}$ implies that $F_{T|X}(t|x)$ is identified for $(t, x) \in [0, \tau) \times \mathbb{R}^d$. More general results are shown in Equation (1.2) of Stute (1996).

In fact, the estimator $\hat{F}_{T|X}$ is the exponential transformation of the Nalson-Aalen estimator of the cumulative hazard function of $F_{T|X}$. Additionally, Beran’s (1981) conditional Kaplan-Meier estimator

$$
\hat{F}^{KN}_{T|X}(t|x; h) = 1 - \prod_{j=1}^{n} \left\{ 1 - \frac{B_{n_\ell}(x; h)}{\sum_{\ell=1}^{n} \mathbb{1}_{[Y_j \leq t]} B_{n_\ell}(x; h)} \right\}^{1_{[Y_j \leq t, \delta_j = 1]}}
$$

can be viewed as the first-order Taylor series approximation of $\hat{F}_{T|X}$.

13 These choices include Gasser-Muller weights and Nadaraya-Watson weights. See for example Gonzalez-Manteiga and Cadarso-Suarez (1994), Dabrowska (1989), and Iglesias-Pérez and González-Manteiga (1999).
2.3 Asymptotic theory

2.3.1 Representations

Asymptotic properties of the unconditional Kaplan-Meier estimator $\hat{F}_T$ in (2.6) have been studied extensively in survival analysis. One attractive feature is that $\hat{F}_T(t) - F_T(t)$ can be approximated by an average of independent and identically distributed random variables with mean zero.\textsuperscript{15} We state this representation in the following proposition for completeness.

**Proposition 2.2.** Under Assumptions D1-D2 and I1, for any $\zeta < \tau = \inf\{t : F_Y(t) = 1\}$ and $t \in [0, \zeta]$, 
\[
\sqrt{n} \left(\hat{F}_T(t) - F_T(t)\right) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \xi(Y_i, \delta_i; t) + R_n(t),
\]
where 
\[
\xi(y, \delta; t) = [1 - F_T(t)] \left[ \mathbb{1}_{[y \leq t, \delta = 1]} \frac{1}{1 - F_Y(y)} - \int_{0}^{y \wedge t} \frac{F_Y^\delta(du)}{(1 - F_Y(u))^2} \right],
\]
and 
\[
\sup_{t \in [0, \zeta]} |R_n(t)| = o_p(1).
\]

The influence function of $\hat{F}_T$ in Proposition 2.2 is centered at zero. Moreover, the precise rate of approximation error $\sup_{t \in [0, \zeta]} |R_n(t)|$ differs if different assumptions about the data generating process are imposed (cf. Lo and Singh (1986), Cai (1998), Chen and Lo (1997)).

In addition to the representation of $\hat{F}_T$, a similar representation of $\hat{F}_{T|X}$ in (2.5) has been established in the case of a univariate covariate by Iglesias-Pérez and González-Manteiga (1999), and further extended to the case of multivariate covariates and dependent data by Liang, de Uña-Álvarez, and Iglesias-Pérez (2012). Since the counterfactual Kaplan-Meier estimator $\hat{F}_{T^*}$ in (2.4) is constructed by taking average of $\hat{F}_{T|X}$ with respect to the empirical distribution of $X^*$, applying the representation of $\hat{F}_{T|X}$ allows us to approximate $\hat{F}_{T^*}$ by an average of independent and identically distributed random variables with mean zero. To obtain the approximation of $\hat{F}_{T^*}$, we need the following assumptions about the kernel and bandwidth.

\textsuperscript{15} Another appealing feature is the strong approximation for $\sqrt{n}(\hat{F}_T - F_T)$ by a sequence of Gaussian processes. See for example Burke, Csörgő, and Horváth (1988) and Major and Rejto (1988).
**Assumption K** (Kernel)
The kernel function $K : \mathbb{R}^d \to \mathbb{R}$ satisfies the following conditions.

1. **K1** $K$ is of bounded variation, vanishes outside $[-1, 1]^d$, and satisfies $\int K(u) \, du = 1$.

2. **K2** There is a positive integer $r \geq 2$ such that
   
   \[ \int \left( \prod_{\ell=1}^{d} u_{\ell}^{\lambda_{\ell}} \right) K(u) \, du = 0 \]
   
   for any $d$-dimensional vector $\lambda = (\lambda_1, \ldots, \lambda_d)^\top$ of nonnegative integers with $\sum_{\ell=1}^{d} \lambda_{\ell} \leq r - 1$.

3. **K3** For $u \in [-1, 1]^d$, $K(u)$ is $r$-times differentiable with respect to $u$ and the derivatives are uniformly continuous and bounded.

4. **K4** For $u \in [-1, 1]^d$, $K(u) = K(|u|)$.

**Assumption B** (Bandwidth)
The sequence $\{h_n\}_{n=1}^{\infty}$ of bandwidths satisfies the following conditions.

1. **B1** $h_n \to 0$.

2. **B2** $n^{1/2} \left( \frac{\log n}{nh_n} \right)^{3/4} \to 0$.

3. **B3** $n^{1/2} h_n^r \to 0$.

These assumptions about the kernel and bandwidth are mild. Assumption K3 restricts the choice of kernels so that the estimator $\hat{F}_{T|x}(t|x)$ is $r$-times differentiable with respect to $x$ and these derivatives are uniformly continuous and bounded. Assumptions B2 implies the remainder term in the representation of $\hat{F}_{T|x}$ in Proposition 2.3 below is of order less than $n^{-1/2}$. Assumptions K2-K3 and B3 are imposed to ensure the bias terms of $\hat{F}_{T|x}$ in Proposition 2.3 are also of order less than $n^{-1/2}$. Note that a necessary condition to make Assumptions B2 and B3 valid simultaneously is $3d < 2r$. Thus, we use a higher order kernel to construct $\hat{F}_{T|x}$ under Assumption K2 if multivariate policy variables are of interest, that is, $d \geq 2$. Assumption K4 is valid if $K$ is a product kernel function $K(u) = \prod_{\ell=1}^{d} k_\ell(u_{\ell})$ and each $k_\ell$ is a univariate kernel function that is symmetric around zero.

Moreover, we need conditions about the support and smoothness of densities and distributions as follows.

**Assumption SP** (Support)
SP1 The support of $X^*$ is the compact subset $J^* \equiv \prod_{\ell=1}^{d}[X^*_\ell, \bar{X}^*_\ell]$ of the interior of the support of $X$, say $J \equiv \prod_{\ell=1}^{d}[X_\ell, \bar{X}_\ell]$.

SP2 There exist positive numbers $u_0$ and $v_0$ such that $\inf \{m(x) : x \in J_{v_0}^*\} \geq u_0$ where $J_{v_0}^* = \prod_{\ell=1}^{d}[X^*_\ell - v_0, \bar{X}^*_\ell + v_0]$.

SP3 There exist positive numbers $\zeta^*$ and $v^*$ such that $\inf\{1 - F_{Y|X}(\zeta^*|x) : x \in J\} \geq v^*$.

Assumption SP1, stronger than Assumption I4, requires the support of $X^*$ to be a proper subset of the support of $X$. When $J^*$ is close to $J$, Assumption SP2 is valid provided the density of $X$ on the boundary of $J^*$ is still bounded away from zero. Assumption SP3 requires the conditional survival function of $Y$ given $X$ is uniformly bounded away from zero on $[0, \zeta^*] \times J$; in addition, it implies $\zeta^* < \tau = \inf\{t : F_Y(t) = 1\}$.

**Assumption SM** (Smoothness)

SM1 The function $m(x)$ is $r$-times differentiable with respect to $x$ on the interior of $J$, and its derivatives are bounded and uniformly continuous.

SM2 For all $(t, x) \in \mathbb{R} \times J^*$, the first $r$ partial derivative with respect to $x$ of $F_{T|X}(t|x)$, $F_{C|X}(t|x)$, $f_{T|X}(t|x)$ and $f_{C|X}(t|x)$ are bounded.

SM3 For all $(t, x) \in \mathbb{R} \times J^*$, the first derivative with respect to $t$ of $f_{T|X}(t|x)$ and $f_{C|X}(t|x)$ are bounded.

SM4 The function $m^*(x)$ is $r$-times differentiable with respect to $x$ on the interior of $J$, and its derivatives are bounded and uniformly continuous.$^{16}$

SM5 Both $\int (\sup_{s \in [0, \zeta^*]} f_{T|X}(s|x))^2 m(x) \, dx$ and $\int [m^*(x)]^2 / m(x) \, dx$ are finite.

Assumptions SM1-SM3 are imposed to obtain the representation of $\hat{F}_{T|X}$ in Proposition 2.3. Similar conditions are used in Iglesias-Pérez and González-Manteiga (1999) and Liang et al. (2012). As in Rothe (2010), we impose Assumption SM4 to establish the representation of $\hat{F}_{T^*}$ in Proposition 2.3 by standard kernel smoothing techniques. Assumption SM5 is technical and valid if the second moments of $\sup_{s \in [0, \zeta^*]} f_{T|X}(s|X)$ and $m^*(X)/m(X)$ are finite.

**Proposition 2.3.** (i) Under Assumptions D1-D2, I1-I2, K1-K3, B1-B3, SP1-SP3, and SM1-SM3, for $(t, x) \in [0, \zeta^*] \times J^*$,

$$
\hat{F}_{T|X}(t|x) - F_{T|X}(t|x) = \sum_{i=1}^{n} \xi^*(Y_i, \delta_i; t, x) B_n(x) + r_n(t, x),
$$

$^{16}$Let $m^*(x) = 0$ if $x \notin J^*$. 

79
where
\[
\xi^*(y, \delta; t, x) = \left[1 - F_{T|X}(t|x)\right] \left[\frac{\mathbb{I}_{[y \leq t, \delta = 1]} - F_{\hat{T}|X}(y|x)}{1 - F_{Y|X}(y|x)} - \int_0^{y \wedge t} \frac{F_{\delta|X}^*(du|x)}{(1 - F_{Y|X}(u|x))^2}\right],
\]
and
\[
\sup_{(t,x) \in [0,\zeta^*] \times J^*} |r_n(t, x)| = O_{as}\left(\left(\frac{\log n}{nh^d}\right)^{3/4}\right).
\]

(ii) If in addition Assumptions I3-I4, K4, and SM4-SM5 hold, then for \( t \in [0, \zeta^*] \),
\[
\sqrt{n} \left(\hat{F}_{T^*}(t) - F_{T^*}(t)\right) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \left(F_{T|X}(t|X_i^*) - F_{T^*}(t)\right) + \frac{1}{\sqrt{n}} \sum_{i=1}^n \xi^*(Y_i, \delta_i; t, X_i) \frac{m^*(X_i)}{m(X_i)} + R_n^*(t),
\]
where
\[
\xi^*(y, \delta; t, x) = \left[1 - F_{T|X}(t|x)\right] \left[\frac{\mathbb{I}_{[y \leq t, \delta = 1]} - F_{\hat{T}|X}(y|x)}{1 - F_{Y|X}(y|x)} - \int_0^{y \wedge t} \frac{F_{\delta|X}^*(du|x)}{(1 - F_{Y|X}(u|x))^2}\right].
\]
and
\[
\sup_{(t,x) \in [0,\zeta^*]} |R_n^*(t)| = o_p(1).
\]

The influence function of \( \hat{F}_{T^*} \) in Proposition 2.3 has mean zero and can be decomposed into two components: the first part arises from the sample variation in \( X^* \), and the second part results from the estimate of \( F_{T|X}(t|x) \). The influence function of \( \hat{F}_{T^*} \) is different from that of the counterfactual CDF estimator in Rothe (2010) because we use the estimator \( \hat{F}_{T|X}(t|x) \) in (2.5) to recover the conditional CDF \( F_{T|X}(t|x) \) in the presence of random censoring.

### 2.3.2 Asymptotic properties

Propositions 2.2 and 2.3 show that the estimators \( \hat{F}_T(t) \) and \( \hat{F}_{T^*}(t) \) can be represented by the average of functions of independent and identically distributed random variables \((Y, \delta, X, X^*)^\top\) plus asymptotic negligible terms. The counterfactual estimator \( \hat{F}_{T^*} \) is uniformly consistent for \( F_{T^*} \) on \([0, \zeta^*]\) because the two classes \( \{x \mapsto F_{T|X}(t|x) : t \in [0, \zeta^*]\} \) and \( \{(y, \delta, x) \mapsto \xi^*(y, \delta; t, x) : t \in [0, \zeta^*]\} \) are both Euclidean under the assumptions imposed. Moreover, for each \( t \in [0, \zeta^*] \), the proposed estimator \( \hat{F}_{T^*}(t) \) can
avoid the curse of dimensionality, namely convergence at the usual parametric rate \( \sqrt{n} \), even if the first-stage estimator \( \hat{F}_{T|x}(t|x) \) converges to \( F_{T|x}(t|x) \) at a rate slower than \( \sqrt{n} \) for each \( x \in J^* \).

The representations of \( \hat{F}_T \) and \( \hat{F}_{T*} \) further allow us to apply techniques in the literature on empirical processes to show that the random map

\[
t \mapsto \sqrt{n} \left( \hat{F}(t) - F(t) \right)
\]

converges weakly to a two dimensional centered Gaussian process, where \( \hat{F} \equiv (\hat{F}_{T*}, \hat{F}_T)^\top \), \( F \equiv (F_{T*}, F_T)^\top \), and \( t = (t_1, t_2)^\top \). Let \( Z = (Y, \delta, X, X^*)^\top \) and

\[
\Psi(t, Z) = \begin{bmatrix}
\psi_1(t_1, Z) \\
\psi_2(t_2, Z)
\end{bmatrix} = \begin{bmatrix}
(F_{T|x}(t_1|X^*) - F_{T*}(t_1)) + \xi^*(Y, \delta; t_1, X) \\
\xi(Y, \delta; t_2)
\end{bmatrix}
\]

(2.8)

where \( \xi \) and \( \xi^* \) are defined in Propositions 2.2 and 2.3, respectively. We establish the weak convergence of \( \hat{F} \) as follows.

**Theorem 2.1.** If Assumptions D, I, K, B, SP, and SM hold, then in \( \mathbb{D}[0, \zeta^*] \times \mathbb{D}[0, \zeta^*] \),

\[
\sqrt{n} \left( \hat{F}(\cdot) - F(\cdot) \right) \Rightarrow \mathbb{F}(\cdot),
\]

where \( \mathbb{F} \) is a two dimensional centered Gaussian process with covariance function \( \Sigma(s, t) = \mathbb{E} \left( \Psi(s, Z) \Psi(t, Z)^\top \right) \) for every \( s, t \in [0, \zeta^*] \times [0, \zeta^*] \).

Theorem 2.1 demonstrates that when the sample size is large, we can approximate the random map in (2.7) by the two dimensional centered Gaussian process \( \mathbb{F} \) with the covariance function

\[
\Sigma(s, t) = \mathbb{E} \left( \Psi(s, Z) \Psi(t, Z)^\top \right) = \begin{bmatrix}
\Sigma_{11}(s_1, t_1) & \Sigma_{12}(s_1, t_2) \\
\Sigma_{21}(s_2, t_1) & \Sigma_{22}(s_2, t_2)
\end{bmatrix}
\]

where

\[
\Sigma_{11}(u, u') = \mathbb{E} \left[ \left( F_{T|x}(u|X^*) - F_{T*}(u) \right) \left( F_{T|x}(u'|X^*) - F_{T*}(u') \right) \right]
\]

\[
= \mathbb{E} \left[ \left( m^*(X) \right)^2 \left[ 1 - F_{T|x}(u|X) \right] \left[ 1 - F_{T|x}(u'|X) \right] \int_0^{u \wedge u'} \frac{F^\delta_{Y|x}(d \bar{u}|X)}{(1 - F_{Y|x}(\bar{u}|X))^2} \right],
\]

\[
\Sigma_{22}(u, u') = \mathbb{E} \left[ \left( m^*(X) \right)^2 \left[ 1 - F_{T}(u) \right] \left[ 1 - F_{T}(u') \right] \int_0^{u \wedge u'} \frac{F^\delta_{Y}(d \bar{u})}{(1 - F_{Y}(\bar{u}))^2} \right],
\]

\( ^{17} \) Details can be found in Dabrowska (1989), Iglesias-Pérez and González-Manteiga (1999), Iglesias-Pérez (2003), and Liang et al. (2012).
and
\[
\Sigma_{12}(u, u') = \Sigma_{21}(u', u)
= \mathbb{E} \left[ \left( 1 - F_{T|X}(u|X) \right) \left[ \frac{\mathbb{1}_{[Y \leq u, \delta = 1]}}{1 - \hat{F}_Y|X}(Y|X) - \int_0^{Y \wedge u} \frac{F_{Y|X}(d\tilde{u}|X)}{(1 - F_{Y|X}(\tilde{u}|X))^2} \right] \cdot \left( \frac{m^*(X)}{m(X)} \right) \left[ \frac{\mathbb{1}_{[Y \leq u', \delta = 1]}}{1 - \hat{F}_Y(Y)} - \int_0^{Y \wedge u'} \frac{F_{Y}(d\tilde{u})}{(1 - F_{Y}(\tilde{u}))^2} \right] \right] [1 - F_T(u')]
+ \mathbb{E} \left[ (F_{T|X}(u|X^*) - F_{T^*}(u)) \left[ \frac{\mathbb{1}_{[Y \leq u', \delta = 1]}}{1 - \hat{F}_Y(Y)} - \int_0^{Y \wedge u'} \frac{F_{Y}(d\tilde{u})}{(1 - F_{Y}(\tilde{u}))^2} \right] \right] [1 - F_T(u')].
\]

The second term in the last line is zero if $X^*$ is independent of $(Y, \delta, X)$, which is expected to be valid in the analysis of the composition effect. In contrast, if we consider a counterfactual manipulation with $X^* = \pi(X)$, then the second term should not be omitted in general.

We can make inference on the distribution policy effect $\Delta_F(t)$ by $\hat{\Delta}_F(t) \equiv \hat{F}_T^*(t) - \hat{F}_T(t)$ because $\sqrt{n} \left( \hat{\Delta}_F(\cdot) - \Delta_F(\cdot) \right)$ converges weakly to $(1, -1)\mathbb{F}(\cdot)$ in $\mathbb{D}[0, \zeta^*] \times \mathbb{D}[0, \zeta^*]$ by Theorem 2.1 and the continuous mapping theorem. The covariance function $\Sigma(s, t)$ can be estimated by replacing the unknown functions with associated consistent estimators. For example, a plug-in estimator of $\Sigma_{11}(u, u')$ is
\[
\hat{\Sigma}_{11}(u, u')
= \frac{1}{n} \sum_{i=1}^n \left[ \hat{F}_{T|X}(u|X_i^*) - \hat{F}_{T^*}(u) \right] \left[ \hat{F}_{T|X}(u'|X_i^*) - \hat{F}_{T^*}(u') \right]
+ \frac{1}{n} \sum_{i=1}^n \left( \frac{m^*(X_i)}{m(X_i)} \right)^2 \left[ 1 - \hat{F}_{T|X}(u|X_i) \right] \left[ 1 - \hat{F}_{T|X}(u'|X_i) \right] \int_0^{u \wedge u'} \frac{\hat{F}_{Y|X}(d\tilde{u}|X_i)}{(1 - \hat{F}_{Y|X}(\tilde{u}|X_i))^2}
\]
where $\hat{F}_{T^*}$ is defined in (2.4), $\hat{F}_{T|X}$ is defined in (2.5),
\[
\hat{F}_{Y|X}(y|x) \equiv \frac{\sum_{i=1}^n \mathbb{1}_{[Y_i \leq y]}K \left( \frac{x - X_i}{h_n} \right)}{\sum_{i=1}^n K \left( \frac{x - X_i}{h_n} \right)},
\]
\[
\hat{F}_{Y|X}^\delta(y|x) \equiv \frac{\sum_{i=1}^n \mathbb{1}_{[Y_i \leq y, \delta_i = 1]}K \left( \frac{x - X_i}{h_n} \right)}{\sum_{i=1}^n K \left( \frac{x - X_i}{h_n} \right)},
\]
\[
\hat{m}(x) \equiv \frac{1}{nh_d} \sum_{i=1}^n K \left( \frac{x - X_i}{h_n} \right), \text{ and}
\]
\[
\hat{m}^*(x) \equiv \frac{1}{nh_d} \sum_{i=1}^n K \left( \frac{x - X_i}{h_n} \right).
\]
To analyze other counterfactual policy effects, we apply the functional delta method as follows.

**Theorem 2.2.** Let $\nu$ be a functional mapping from a subset of $\mathbb{D}[0, \zeta^*] \times \mathbb{D}[0, \zeta^*]$ to some normed space $\mathcal{V}$. Suppose that $\nu$ is Hadamard differentiable at $F$ with derivative $\nu_F'$. Let $Z = (Y, \delta, X, X^*)^\top$ and $\Psi'(t, Z) = \nu_F'(\Psi(t, Z))$, where $\Psi = (\psi_1, \psi_2)^\top$ is defined in (2.8). Under the assumptions of Theorem 2.1, we have that in $\mathcal{V},$

$$
\sqrt{n} \left(\nu(\hat{F})(\cdot) - \nu(F)(\cdot)\right) \Rightarrow \nu_F'(\mathbb{F})(\cdot) \equiv \mathbb{G}(\cdot),
$$

where $\mathbb{G}$ is a two dimensional centered Gaussian process with covariance function

$$
\Sigma'(s, t) = \mathbb{E} \left(\Psi'(s, Z)\Psi'(t, Z)^\top\right).
$$

Let $\hat{\Lambda}_T = \nu(\hat{F}_T)$ and $\hat{\Lambda}_T^* = \nu(\hat{F}_{T^*})$ for the functional $\nu : F \mapsto \int_{[0, 1]} \frac{1}{1-F} \, dF$. In addition, let $\Lambda \equiv (\hat{\Lambda}_T, \hat{\Lambda}_T)^\top$ and $\Lambda \equiv (\Lambda_T, \Lambda_T)^\top$. Theorem 2.2 immediately implies the following corollary. We can make inference on the cumulative hazard policy effect $\Delta_\Lambda(t)$ by $\hat{\Delta}_\Lambda(t) \equiv \hat{\Lambda}_T(t) - \hat{\Lambda}_T(t)$ because $\sqrt{n} \left(\hat{\Delta}_\Lambda(\cdot) - \Delta_\Lambda(\cdot)\right)$ converges weakly in $\mathbb{D}[0, \zeta^*] \times \mathbb{D}[0, \zeta^*]$ by the continuous mapping theorem.

**Corollary 2.1.** Under the assumptions of Theorem 2.1,

$$
\sqrt{n} \left(\hat{\Lambda}(\cdot) - \Lambda(\cdot)\right) \Rightarrow \left[\int_0^1 \frac{1}{1-F_{T^*}(u)} \mathbb{F}_1(du) + \int_0^1 \frac{\mathbb{F}_1(u)}{(1-F_{T^*}(u))^2} \mathbb{F}_{T^*}(du)\right] \equiv \mathbb{A}(\cdot)
$$

in $\mathbb{D}[0, \zeta^*] \times \mathbb{D}[0, \zeta^*]$. The two dimensional process $\mathbb{A}$ is centered Gaussian with covariance function $\Sigma^A(s, t) = \mathbb{E} \left(\Psi^A(s, Z)\Psi^A(t, Z)^\top\right)$, where $\Psi^A(t, Z) = \left(\psi_1(t, Z), \psi_2(t, Z)^\top\right)^\top$, $Z = (Y, \delta, X, X^*)^\top$, and $(\psi_1, \psi_2)^\top$ is defined in (2.8). \hfill \square

### 2.4 Monte Carlo simulation

In this section, we evaluate the small-sample performance of the proposed estimator $\hat{F}_{T^*}$ and its associated counterfactual policy effects by Monte Carlo simulation. We consider the following data generating process (DGP):

$$
Y = \min \{T, C\},
$$

$$
T = 5 - 3X_1 + 2X_2 + \varepsilon \cdot \sqrt{X_1^2 + X_2^2},
$$

where the covariates $(X_1, X_2)$ follow the Beta distribution with shape parameters $(2, 2)$, the unobserved heterogeneity $\varepsilon$ is exponentially distributed with mean 2, and $C$ is log-normally distributed with parameters $(2.5, 1)$; additionally, $(X_1, X_2, \varepsilon, C)$ are mutually

83
independent. The censoring rate in this design is approximate 23.45%. We study the policy intervention

\[(X_1^*, X_2^*)^\top = \pi(X_1, X_2) = 0.05 + 0.9 \cdot (X_1, X_2)^\top\]

and this intervention does not affect \((C, \varepsilon)\). The DGP and counterfactual policy are not meant to mimic any data set in empirical studies; instead, they are only used to illustrate the proposed method.

We consider the sample sizes \(n = 100, 200, 400,\) and 800. The number of simulation replications is \(S = 1000\). The criteria of evaluation include the mean integrated absolute error (MIAE) and the root mean integrated squared error (RMISE).\(^{18}\) The CDF and cumulative hazard estimators in this Monte Carlo study are calculated over the equidistant grids \(\{4.25, 4.30, 4.35, \ldots, 8.10, 8.15\}\), and the numerical integration in MIAE and RMISE is taken over \([4.25, 8.15]\), where 4.25 and 8.15 are the 10% and 90% quantile of \(T\), respectively.

We evaluate the estimation of the CDFs \((F_{T^*}, F_T)\) and the estimation of the cumulative hazard functions \((\Lambda_{T^*}, \Lambda_T)\). We estimate \(F_T\) by the unconditional Kaplan-Meier estimator \(\hat{F}_T\) in (2.6). To estimate \(F_{T^*}\), we use the proposed estimator \(\hat{F}_{T^*}\) in (2.4) with the fourth order product kernel function \(K(u_1, u_2) = k(u_1)k(u_2)\) where \(k(u) = (15/32)(3 - 10u^2 + 7u^4)1_{|u|<1}\) and the bandwidth \(h_n = 3n^{-1/7}\).\(^{19}\) Table 2.1 shows that the MIAE and RMISE of \((\hat{F}_{T^*}, \hat{F}_T)\) shrink as the sample size increases. The MIAE and RMISE of \(\hat{F}_{T^*}\) halves as the sample size quadruples; namely, this estimator converges at the rate \(\sqrt{n}\). This confirms the theoretical analysis that the proposed estimator \(\hat{F}_{T^*}\) does not suffer from the curse of dimensionality. We also consider an oracle estimator \(\tilde{F}_{T^*}\), which is the unconditional Kaplan-Meier estimator of \(F_{T^*}\) if \(Y^* \equiv \min\{T^*, C\}\) and \(\delta^* = 1_{[T^* \leq C]}\) are observed. Surprisingly, this oracle estimator \(\tilde{F}_{T^*}\) does not outweigh considerably the proposed estimator \(\hat{F}_{T^*}\) in terms of MIAE and RMISE; however, \(\tilde{F}_{T^*}\) is infeasible because \(Y^*\) and \(\delta^*\) are unobserved in practice. Moreover, we consider Rothe’s (2010) counterfactual estimator \(F^+_T\), which is constructed under the assumption that

\(^{18}\) For an estimator \(\hat{f}\) of a generic real-valued function \(f\), the mean integrated absolute error of \(\hat{f}\) is

\[\text{MIAE}(\hat{f}) = \mathbb{E} \left[ \int |\hat{f}(u) - f(u)| \, du \right]\]

and the root mean integrated squared error of \(\hat{f}\) is

\[\text{RMISE}(\hat{f}) = \sqrt{\mathbb{E} \left[ \int |\hat{f}(u) - f(u)|^2 \, du \right]}\]

\(^{19}\) Assumptions K1-K3 and B1-B3 are satisfied under this choice of kernel function and bandwidth.
Table 2.1: Estimation of the CDFs

<table>
<thead>
<tr>
<th></th>
<th>Unconditional CDF</th>
<th>Counterfactual CDF</th>
</tr>
</thead>
<tbody>
<tr>
<td>n</td>
<td>( \hat{F}_T )</td>
<td>( \hat{F}_{T^*} )</td>
</tr>
<tr>
<td>100</td>
<td>0.1454</td>
<td>0.1482</td>
</tr>
<tr>
<td>200</td>
<td>0.1061</td>
<td>0.1089</td>
</tr>
<tr>
<td>400</td>
<td>0.0741</td>
<td>0.0761</td>
</tr>
<tr>
<td>800</td>
<td>0.0528</td>
<td>0.0547</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Unconditional CDF</th>
<th>Counterfactual CDF</th>
</tr>
</thead>
<tbody>
<tr>
<td>n</td>
<td>( \hat{F}_T )</td>
<td>( \hat{F}_{T^*} )</td>
</tr>
<tr>
<td>100</td>
<td>0.0922</td>
<td>0.0942</td>
</tr>
<tr>
<td>200</td>
<td>0.0674</td>
<td>0.0693</td>
</tr>
<tr>
<td>400</td>
<td>0.0475</td>
<td>0.0489</td>
</tr>
<tr>
<td>800</td>
<td>0.0336</td>
<td>0.0349</td>
</tr>
</tbody>
</table>

data are not censored.\(^{20}\) As shown in Table 2.1, the neglect of censoring results in larger MIAE and RMISE of \( F_{T^*}^\dagger \), compared with those of \( \hat{F}_{T^*} \) and \( \tilde{F}_{T^*} \). Table 2.2 reports the MIAE and RMISE of the estimated cumulative hazard functions

\[
\hat{\Lambda}_T \equiv \nu(\hat{F}_T), \quad \hat{\Lambda}_{T^*} \equiv \nu(\hat{F}_{T^*}), \quad \tilde{\Lambda}_{T^*} \equiv \nu(\tilde{F}_{T^*}), \quad \text{and} \quad \Lambda_{T^*}^\dagger \equiv \nu(F_{T^*}^\dagger)
\]

where \( \nu \) is the functional that maps \( F \) to \( -\log(1 - F) \). Similarly, the simulation results provide evidence that the proposed cumulative hazard function \( \hat{\Lambda}_{T^*} \) converges at the rate \( \sqrt{n} \). Moreover, \( \hat{\Lambda}_{T^*} \) performs as well as the oracle estimator \( \tilde{\Lambda}_{T^*} \). The neglect of censoring, however, causes relatively large bias of \( \Lambda_{T^*}^\dagger \).

2.5 Conclusion

We have proposed a two-stage fully nonparametric estimator of the counterfactual CDF for a duration variable, which is subject to the random censoring. Since the nonseparable heterogeneity is of unrestricted dimensionality and its marginal distribution is unspecified, the duration analysis in this chapter would avoid several types of model misspecification in empirical studies. The incorporation of covariates also enables researchers to evaluate

\(^{20}\) Since the support of \((X_1^*, X_2^*)\) is a proper subset of the support of \((X_1, X_2)\), we construct Rothe's estimator based on the aforementioned fourth order kernel function and bandwidth \( h_n = 3n^{-1/7} \).
Table 2.2: Estimation of the Cumulative Hazard Functions

<table>
<thead>
<tr>
<th></th>
<th>Unconditional CDF</th>
<th>Counterfactual CDF</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n$</td>
<td>$\hat{\Lambda}_T$</td>
<td>$\hat{\Lambda}_{T^*}$</td>
</tr>
<tr>
<td>100</td>
<td>0.5382</td>
<td>0.5359</td>
</tr>
<tr>
<td>200</td>
<td>0.3870</td>
<td>0.3945</td>
</tr>
<tr>
<td>400</td>
<td>0.2661</td>
<td>0.2736</td>
</tr>
<tr>
<td>800</td>
<td>0.1862</td>
<td>0.1944</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Unconditional CDF</th>
<th>Counterfactual CDF</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n$</td>
<td>$\hat{\Lambda}_T$</td>
<td>$\hat{\Lambda}_{T^*}$</td>
</tr>
<tr>
<td>100</td>
<td>0.4124</td>
<td>0.3919</td>
</tr>
<tr>
<td>200</td>
<td>0.2861</td>
<td>0.2845</td>
</tr>
<tr>
<td>400</td>
<td>0.1943</td>
<td>0.1974</td>
</tr>
<tr>
<td>800</td>
<td>0.1332</td>
<td>0.1377</td>
</tr>
</tbody>
</table>

the change of duration dependence in response to a counterfactual policy that changes exogenous covariates.

There are some directions of extension to this research. First, we may relax the assumption of exogenous covariates by the control function approach. Second, it would be important to establish the validity of a bootstrap method to construct a confidence band for the counterfactual policy effect. Finally, the inclusion of time-varying covariates might be relevant in some empirical studies.

Appendix 2.A  Technical proofs

2.A.1 Proof of Proposition 2.1

Proof. Under Assumption I1, we can show that

$$F_T(t) = \int \mathbb{E}[\delta|Y = s] \exp \left\{ \int_0^s \frac{1 - \mathbb{E}[\delta|Y = y]}{1 - F_Y(y)} F_Y(dy) \right\} F_Y(ds)$$

for $t < \tau$. See page 1604 of Stute and Wang (1993). Since data on $\{(Y_i, \delta_i)\}_{i=1}^n$ are available, $F_T(t)$ is identified for $t \in [0, \tau)$. Under Assumption I2, for $(t, x) \in [0, \tau) \times \mathbb{R}^d$, the identification of $F_{T|X}(t|x)$ is established by similar arguments in Lemma 25.74 of van der Vaart (1998).
The independence between $\varepsilon$ and $X$ implies
\[
F_{T^*}(t) = \int \mathbb{P}(\varphi(x, \varepsilon) \leq t) F_{X^*}(dx).
\]
Since $\varepsilon$ is also independent of $X^*$, which only takes values in a subset of support of $X$, we have
\[
F_{T^*}(t) = \int \mathbb{P}(\varphi(x, \varepsilon) \leq t | X = x) F_{X^*}(dx) = \mathbb{E}[F_{T|X}(t|X^*)].
\]

2.A.2 Proof of Proposition 2.2


2.A.3 Proof of Proposition 2.3

Proof. (i) Let $\hat{\Lambda}_{T|X}(t|X) = -\log [1 - \hat{F}_{T|X}(t|X)]$. By Theorem 2.1 of Liang et al. (2012),
\[
\sup_{(t,x) \in [0,\zeta] \times J^*} |\hat{\Lambda}_{T|X}(t|x) - \Lambda_{T|X}(t|x)| = o_{a.s.} (1),
\]
where $\Lambda_{T|X}(t|x) = -\log [1 - F_{T|X}(t|x)]$. Hence, with probability one, $\hat{\Lambda}_{T|X}(t|x)$ is well defined on $[0,\zeta] \times J^*$ for $n$ large. Taylor series expansion yields
\[
\hat{F}_{T|X}(t|x) - F_{T|X}(t|x) = [1 - F_{T|X}(t|x)] \left[ \hat{\Lambda}_{T|X}(t|x) - \Lambda_{T|X}(t|x) \right]
+ O \left( \sup_{(t,x) \in [0,\zeta] \times J^*} \left[ \hat{\Lambda}_{T|X}(t|x) - \Lambda_{T|X}(t|x) \right]^2 \right).
\]
The desired result follows from Theorems 2.1 and 2.3 of Liang et al. (2012).

(ii) Let $\mathcal{D}_n \equiv \{Y_i, \delta_i, X_i\}_{i=1}^n$ and $X^*$ be a random variable that is independent of $\mathcal{D}_n$ and follows the same distribution as $X_1^*$ does. Note that
\[
\sqrt{n} \left[ \hat{F}_{T^*}(t) - F_{T^*}(t) \right]
= \frac{1}{\sqrt{n}} \sum_{i=1}^n \left[ \hat{F}_{T|X}(t|X_i^*) - F_{T^*}(t) \right]
= \frac{1}{\sqrt{n}} \sum_{i=1}^n \left[ F_{T|X}(t|X_i^*) - F_{T^*}(t) \right]
+ \sqrt{n} \left\{ \mathbb{E}[\hat{F}_{T|X}(t|X^*)|\mathcal{D}_n] - \mathbb{E}[F_{T|X}(t|X^*)] \right\}
+ \frac{1}{\sqrt{n}} \sum_{i=1}^n \left\{ \left[ \hat{F}_{T|X}(t|X_i^*) - \mathbb{E}[\hat{F}_{T|X}(t|X^*)|\mathcal{D}_n] \right] - \left[ F_{T|X}(t|X_i^*) - \mathbb{E}[F_{T|X}(t|X^*)] \right] \right\}.
\]
Lemma 2.1 shows that
\[
\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{ \left[ \hat{F}_{T|X}(t|X_i^*) - \mathbb{E}[\hat{F}_{T|X}(t|X_i^*)] \right] - \left[ F_{T|X}(t|X_i^*) - \mathbb{E}[F_{T|X}(t|X_i^*)] \right] \right\} = o_p(1)
\]
uniformly in \( t \in [0, \zeta^*] \). The result follows from Lemma 2.2 that
\[
\sqrt{n} \left\{ \mathbb{E}[\hat{F}_{T|X}(t|X^*)|\mathcal{D}_n] - \mathbb{E}[F_{T|X}(t|X^*)] \right\} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \xi^*(Y_i, \delta_i; t, X_i) \frac{m^*(X_i)}{m(X_i)} + r^*_n(t)
\]
and \( \sup_{t \in [0, \zeta^*]} |r^*_n(t)| = o_p(1) \). □

2.A.4 Proof of Theorem 2.1

Proof. The representations in Propositions 2.2 and 2.3 allow us to write \( \sqrt{n} \left( \hat{F} - F \right) \) as a two-dimensional empirical process plus an asymptotically negligible term; that is, for each \( t \in [0, \zeta^*] \times [0, \zeta^*] \), we have
\[
\sqrt{n} \left( \hat{F}(t) - F(t) \right) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} [\phi(t, Z_i) - \mathbb{E}(\phi(t, Z_i))] + o_p(1),
\]
where \( Z_i = (Y_i, \delta_i, X_i, X_i^*)^\top \) for each \( i \in \{1, 2, \ldots, n\} \),
\[
\phi(t, Z) = \begin{bmatrix} F_{T|X}(t|X^*) \\ 0 \end{bmatrix} + \begin{bmatrix} \xi^*(Y, \delta; t, X) \frac{m^*(X)}{m(X)} \\ \xi(Y, \delta; t) \end{bmatrix},
\]
and the two functions \( \xi \) and \( \xi^* \) are defined in Propositions 2.2 and 2.3, respectively. It follows from Lemma 2.3 that the classes \( \{(y, \delta) \mapsto \xi(y, \delta; t) : t \in [0, \zeta^*]\} \), \( \{(y, \delta, x) \mapsto \xi^*(y, \delta; t, x) : t \in [0, \zeta^*]\} \), and \( \{x^* \mapsto F_{T|X}(t|x^*) : t \in [0, \zeta^*]\} \) are all Euclidean. Lemma 2.14 of Pakes and Pollard (1989) further implies that the class \( \Phi = \{\phi(t, \cdot) : t \in [0, \zeta^*] \times [0, \zeta]\} \) is Euclidean under Assumption SM5. Hence, the class \( \Phi \) is Donsker by Theorem 19.14 of van der Vaart (1998), and the process
\[
\sqrt{n} \left( \hat{F}(\cdot) - F(\cdot) \right) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} [\phi(\cdot, Z_i) - \mathbb{E}(\phi(\cdot, Z_i))] + o_p(1)
\]
converges weakly in \( \mathbb{D}[0, \zeta^*] \times \mathbb{D}[0, \zeta^*] \) to a centered Gaussian process with covariance function
\[
\mathbb{E} \left( [\phi(s, Z) - \mathbb{E}(\phi(s, Z))] [\phi(t, Z) - \mathbb{E}(\phi(t, Z))]^\top \right) = \mathbb{E} \left( \Psi(s, Z) \Psi(t, Z)^\top \right)
\]
for each \( s, t \in [0, \zeta^*] \times [0, \zeta^*] \). □
2.A.5 Proof of Theorem 2.2

Proof. The proof is an application of the functional delta method, which is established in Theorem 20.8 of van der Vaart (1998). □

2.A.6 Proof of Corollary 2.1

Proof. Let \( \eta = \min\{1 - F_{T^*}(\zeta^*), 1 - F_T(\zeta^*)\} \). Assumptions I1 implies that \( 1 - F_{T^*}(\zeta^*) \geq 1 - F_T(\zeta^*) > 0 \); in addition, Assumptions I2 and SP3 imply that \( 1 - F_{T^*}(\zeta^*) = \mathbb{E}(1 - F_{T|X}(\zeta^*|X^*)) \geq \nu^* > 0 \). It follows that \( \eta > 0 \). Let \( \mathbb{D}_\eta^2 \) be the set of nondecreasing càdlàg functions \((F_1, F_2)^\top\) such that \( F_\ell : [0, \zeta^*] \to \mathbb{R} \) with \( F_\ell(0) = 0 \) and \( 1 - F_\ell(\zeta^*) \geq \eta \) for each \( \ell \in \{1, 2\} \). Let \( \nu \) be the functional from \( \mathbb{D}_\eta^2 \) to \( \mathbb{D}[0, \zeta^*] \times \mathbb{D}[0, \zeta^*] \) such that \( \nu(F_1, F_2) = (\nu_0(F_1), \nu_0(F_2))^\top \) where \( \nu_0(F)(\cdot) = \int_{[0,1]} F(du) \). From Lemma 20.14 of van der Vaart (1998), the functional \( \nu \) is Hadamard differentiable at \( \mathbf{F} = (F_{T^*}, F_T)^\top \in \mathbb{D}_\eta^2 \). Moreover, it can be shown that the Hadamard derivative is

\[
\nu'_\mathbf{F}(S_1, S_2) = \begin{bmatrix}
\int_0^1 \frac{1}{1-F_{T^*}(u)} S_1(du) + \int_0^1 \frac{S_1^-(u)}{(1-F_{T^*}(u))^2} F_{T^*}(du) \\
\int_0^1 \frac{1}{1-F_T(u)} S_2(du) + \int_0^1 \frac{S_2^-(u)}{(1-F_T(u))^2} F_T(du)
\end{bmatrix}
\]

for \((S_1, S_2)^\top \in \mathbb{D}_\eta^2 \). Applying Theorem 2.2 yields that

\[
\sqrt{n} \left( \hat{\mathbf{A}}(\cdot) - \mathbf{A}(\cdot) \right) = \sqrt{n} \left( \nu(\hat{\mathbf{F}})(\cdot) - \nu(\mathbf{F})(\cdot) \right) \Rightarrow \nu'_\mathbf{F}(\mathbf{F})(\cdot) \equiv \mathbf{A}(\cdot)
\]

in \( \mathbb{D}[0, \zeta^*] \times \mathbb{D}[0, \zeta^*] \). □

2.A.7 Auxiliary lemmas

Lemma 2.1. Let \( \mathcal{D}_n \equiv \{Y, \delta_i, X_i\}_{i=1}^n \) and \( X^* \) be a random variable that is independent of \( \mathcal{D}_n \) and follows the same distribution as \( X^*_1 \) does. Under Assumptions D1-D2, I1-I2, K1-K3, B1-B3, SP1-SP3, and SM1-SM2,

\[
\sup_{t \in [0, \zeta^*]} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^n \left[ \hat{F}_{T|X}(t|X^*_i) - \mathbb{E}[\hat{F}_{T|X}(t|X^*)|\mathcal{D}_n] \right. \right. \\
\left. \left. \mathbb{E}[F_{T|X}(t|X^*_i)] \right] \\n- \left. \left[ F_{T|X}(t|X^*_i) - \mathbb{E}[F_{T|X}(t|X^*)] \right) \right] \right| = o_p(1).
\]

Proof. Let \( \hat{\Gamma}(t|x; h_n) = \hat{F}_{T|X}(t|x; h_n) - F_{T|X}(t|x) \) for \((t, x) \in [0, \zeta^*] \times J^* \). Our goal is to show that

\[
\sup_{t \in [0, \zeta^*]} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^n \left[ \hat{\Gamma}(t|X^*_i; h_n) - \mathbb{E}[\hat{\Gamma}(t|X^*; h_n)|\mathcal{D}_n] \right] \right| = o_p(1).
\]

89
For each \( t \in [0, \zeta^*] \), we have

\[
\mathbb{E} \left[ \left| \hat{\Gamma}(t|X^*; h_n) \right|^2 | \mathcal{D}_n \right] \leq \sup_{(s,x) \in [0,\zeta^*] \times J^*} \left| \hat{\Gamma}(s|x; h_n) \right|^2 = O_{\text{as}} \left( \frac{\log n}{nh^d} \right) = o_p(1)
\]

by Theorem 2.1 of Liang et al. (2012). Let \( \mathcal{C}_r(J^*) \) be the class of real-valued functions defined on \( J^* \) whose partial derivatives up to order \( r \) exist and are bounded by some constant. Example 19.9 of van der Vaart (1998) shows that \( \mathcal{C}_r(J^*) \) is Donsker whenever \( r > d/2 \), which is guaranteed under Assumptions B2 and B3. Note that for each \( t \in [0, \zeta^*] \), \( \{ x \mapsto \hat{\Gamma}(t|x; h_n) \}_{n=1}^{\infty} \) is a sequence of random functions taking their values in \( \mathcal{C}_r(J^*) \) under Assumption K3 and SM2. It follows from Lemma 19.24 of van der Vaart (1998) that for each \( t \in [0, \zeta^*] \),

\[
\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left( \hat{\Gamma}(t|X^*_i; h_n) - \mathbb{E}[\hat{\Gamma}(t|X^*; h_n)|\mathcal{D}_n] \right) = o_p(1)
\]

It remains to show the uniform convergence in probability. Since the dominated convergence theorem implies that with probability one,

\[
\mathbb{E} \left[ \left| \hat{\Gamma}(s|X^*; h_n) - \hat{\Gamma}(s'|X^*; h_n) \right|^2 | \mathcal{D}_n \right] \to 0
\]
as \( s' \to s \), the stochastic equicontinuity holds; specifically, we have for any \( u > 0 \) there is a \( v > 0 \) such that

\[
\limsup_{n \to \infty} \mathbb{P} \left( \sup_{|s-s'| < v} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left( \hat{\Gamma}(s|X^*_i; h_n) - \mathbb{E}[\hat{\Gamma}(s|X^*; h_n)|\mathcal{D}_n] \right) - \left( \hat{\Gamma}(s'|X^*_i; h_n) - \mathbb{E}[\hat{\Gamma}(s'|X^*; h_n)|\mathcal{D}_n] \right) \right| > u \) < u.

Applying Theorem 21.9 of Davidson (1994) yields the desired result. \( \square \)

**Lemma 2.2.** Let \( \mathcal{D}_n \equiv \{ Y_i, \delta_i, X_i \}_{i=1}^{n} \) and \( X^* \) be a random variable that is independent of \( \mathcal{D}_n \) and follows the same distribution as \( X^*_1 \) does. Suppose that the assumptions of Proposition 2.3 hold. Then, for \( t \in [0, \zeta^*] \),

\[
\sqrt{n} \left\{ \mathbb{E}[\hat{F}_{T|X}(t|X^*)|\mathcal{D}_n] - \mathbb{E}[F_{T|X}(t|X^*)] \right\} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \xi^*(Y_i, \delta_i; t, X_i) m^*(X_i) + r_n^*(t)
\]
and \( \sup_{t \in [0, \zeta^*]} |r_n^*(t)| = o_p(1) \).
Proof. Following the first part of Proposition 2.3, we obtain
\[
\sqrt{n} \left\{ \mathbb{E}\left[ \hat{F}_{T|X}(t|X^*) \mid \mathcal{D}_n \right] - \mathbb{E}[F_{T|X}(t|X^*)] \right\}
\]
\[
= \sqrt{n} \int_{J^*} \left[ \hat{F}_{T|X}(t|x) - F_{T|X}(t|x) \right] m^*(x) \, dx
\]
\[
= \sqrt{n} \int_{J^*} \sum_{i=1}^{n} \xi^*(Y_i, \delta_i; t, X_i) B_{n_i}(x)m^*(x) \, dx
\]
\[
+ \sqrt{n} \int_{J^*} \sum_{i=1}^{n} [\xi^*(Y_i, \delta_i; t, x) - \xi^*(Y_i, \delta_i; t, X_i)] B_{n_i}(x)m^*(x) \, dx
\]
\[
+ \sqrt{n} \int_{J^*} r_n(t, x)m^*(x) \, dx
\]
= Term I + Term II + Term III.

Term III is asymptotically uniformly negligible because
\[
\sup_{t \in [0, \xi^*]} \left| \sqrt{n} \int_{J^*} r_n(t, x)m^*(x) \, dx \right| \leq \sqrt{n} \left( \sup_{(t, x) \in [0, \xi^*] \times J^*} |r_n(t, x)| \right) \int_{J^*} m^*(x) \, dx
\]
\[
= o_p(1)
\]
by Assumption B2.

We first show that Term I can be approximated by
\[
\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \xi^*_i(Y_i, \delta_i; t, X_i) \frac{m^*(X_i)}{m(X_i)}.
\]
For each \( i \in \{1, 2, \ldots, n\} \), we abbreviate by writing \( \xi^*_i \equiv \xi^*(Y_i, \delta_i; t, X_i) \), and we have \( B_{n_i}(x) = K(\frac{x-X_i}{h_n})/nh_n^d \hat{m}(x) \). Applying the second order Taylor expansion of \( 1/\hat{m}(x) \) around \( 1/m(x) \) yields

Term I = \[
\sqrt{n} \sum_{i=1}^{n} \xi^*_i \int_{J^*} B_{n_i}(x)m^*(x) \, dx
\]
\[
= \sqrt{n} \sum_{i=1}^{n} \xi^*_i \int_{J^*} \frac{1}{\hat{m}(x)} \frac{m^*(x)}{m(X_i)} \, dx
\]
\[
= \sqrt{n} \sum_{i=1}^{n} \xi^*_i \int_{J^*} \frac{1}{\hat{m}(x)} \frac{m^*(x)}{m(X_i)} \, dx
\]
\[
+ \sqrt{n} \sum_{i=1}^{n} \xi^*_i \int_{J^*} \frac{1}{\hat{m}(x)} \frac{m^*(x)}{m(X_i)} \left[ m(x) - \hat{m}(x) \right] \, dx
\]
\[
+ \sqrt{n} \sum_{i=1}^{n} \xi^*_i \int_{J^*} \frac{1}{\hat{m}(x)} \frac{m^*(x)}{m(X_i)} \left[ m(x) - \hat{m}(x) \right]^2 \, dx
\]
= Term I.a + Term I.b + Term I.c
where $\tilde{m}(x)$ is between $m(x)$ and $\hat{m}(x)$. The last term is asymptotically uniformly negligible because

$$\sup_{t \in [0, \zeta^*]} |\xi_{i,t}^*| \leq \frac{1}{v^*} \left( 1 + \frac{1}{v^*} \right)$$

by Assumption SP3, and

$$|\text{Term I.c}| \leq \frac{2}{u_0} \sup_{x \in J^*} |\tilde{m}(x) - m(x)|^2 \frac{1}{n^{1/2}} \sum_{i=1}^n |\xi_{i,t}^*| \int |K(z)| m^*(X_i + h_nz) \, dz$$

$$= O \left( \frac{n^{1/2}}{n^{1/2}} \sup_{x \in J^*} |\tilde{m}(x) - m(x)|^2 \right)$$

$$= o_p(1)$$

by Assumptions SP2, K1-K3 and B1-B3. In addition, Assumptions SM1, SM4, and K1-K3 imply that the first term satisfies

$$\text{Term I.a} = \frac{1}{\sqrt{n}} \sum_{i=1}^n \xi_{i,t}^* \frac{m^*(X_i)}{m(X_i)} + o_p(1). \quad (2.A.1)$$

It suffices to show that the second term

$$\text{Term I.b} \equiv \frac{1}{\sqrt{n}} \sum_{i=1}^n \xi_{i,t}^* \int_{J^*} \frac{1}{h_n^d} K \left( \frac{x - X_i}{h_n} \right) \frac{m^*(x)}{[m(x)]^2} [m(x) - \tilde{m}(x)] \, dx \quad (2.A.2)$$

is asymptotically uniformly negligible. Let $u(x) \equiv m^*(x)/m(x)$ and $v(x) \equiv m^*(x)/(m(x))^2$.

Assumptions SM1, SM4, and K1-K3 imply that

$$\int_{J^*} \frac{1}{h_n^d} K \left( \frac{x - X_i}{h_n} \right) \frac{m^*(x)}{[m(x)]^2} [m(x) - \tilde{m}(x)] \, dx$$

$$= \frac{1}{h_n^d} K \left( \frac{x - X_i}{h_n} \right) u(x) \, dx - \int_{J^*} \frac{1}{h_n^d} K \left( \frac{x - X_i}{h_n} \right) v(x) \tilde{m}(x) \, dx$$

$$= u(X_i) + O_p(h_n^r) - \frac{1}{nh_n^d} \sum_{j=1}^n \int K(w) v(X_i + h_nw) K \left( \frac{X_i - X_j}{h_n} + w \right) \, dw$$

$$= u(X_i) - \frac{1}{nh_n^d} \sum_{j=1}^n v(X_i) K \left( \frac{X_i - X_j}{h_n} \right) - \frac{1}{n} \sum_{j=1}^n Q(X_i, X_j; h_n) + o_p(n^{-1/2}) \quad (2.A.3)$$

where

$$Q(x_1, x_2; h_n)$$

$$= \frac{1}{h_n^d} \int K(z) \left[ v(x_1 + h_nz) K \left( \frac{x_1 - x_2}{h_n} + z \right) - v(x_1) K \left( \frac{x_1 - x_2}{h_n} \right) \right] \, dz.$$
Let $\bar{Q}(x; h_n) = \mathbb{E}[Q(x, X; h_n)]$. Substituting (2.A.3) into (2.A.2) yields

Term I.b

\[
\frac{1}{n^{3/2}} \sum_{i=1}^{n} \sum_{j=1}^{n} \xi_{i,t}^* \left[ u(X_i) - \frac{1}{h_n} v(X_i) K \left( \frac{X_i - X_j}{h_n} \right) \right] - \frac{1}{n^{1/2}} \sum_{i=1}^{n} \xi_{i,t}^* \bar{Q}(X_i; h_n) - \frac{1}{n^{3/2}} \sum_{i=1}^{n} \sum_{j=1}^{n} \xi_{i,t}^* \left[ Q(X_i, X_j; h_n) - \bar{Q}(X_i; h_n) \right] + o_p(1) = \text{Term I.b1} - \text{Term I.b2} - \text{Term I.b3} + o_p(1) .
\]

(2.A.4)

For simplicity, we introduce further notation. Let $\bar{m}(x; h_n) \equiv \mathbb{E} \left[ \frac{1}{n} K \left( \frac{x - X}{h_n} \right) \right]$. For each $t \in [0, \zeta^*]$, let

\[
\mathcal{L}(W_i, W_j; t, h_n) \equiv \xi_{i,t}^* v(X_i) \left[ h_n^d \bar{m}(X_i; h_n) - K \left( \frac{X_i - X_j}{h_n} \right) \right] ,
\]

(2.A.5)

and

\[
\mathcal{M}(W_i, W_j; t, h_n) \equiv \xi_{i,t}^* h_n^d \left[ Q(X_i, X_j; h_n) - \bar{Q}(X_i; h_n) \right] .
\]

(2.A.6)

where $W_i = (Y_i, \delta_i, X_i)^T$ for each $i \in \{1, 2, \ldots, n\}$. We have

\[
\frac{1}{n^{3/2}} \sum_{i=1}^{n} \sum_{j=1}^{n} \xi_{i,t}^* \left[ u(X_i) - \frac{1}{h_n^d} v(X_i) K \left( \frac{X_i - X_j}{h_n} \right) \right] = \frac{1}{h_n^d} \frac{1}{n^{3/2}} \sum_{i=1}^{n} \mathcal{L}(W_i, W_j; t, h_n) + \frac{1}{n^{1/2}} \sum_{i=1}^{n} \xi_{i,t}^* v(X_i) [m(X_i) - \bar{m}(X_i; h_n)] .
\]

(2.A.7)

Since $\mathcal{L}$ is uniformly bounded and $n^{1/2} h_n^d \to \infty$ by Assumptions B1 and B2, we have

\[
\sup_{t \in [0, \zeta^*]} \left| \frac{1}{h_n^d} \frac{1}{n^{3/2}} \sum_{i=1}^{n} \mathcal{L}(W_i, W_i; t, h_n) \right| = o_p(1) .
\]

In addition, for each $t \in [0, \zeta^*]$ and $h > 0$, $\frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} \mathcal{L}(W_i, W_j; t, h)$ is a degenerate U statistic of order 2 because $\mathbb{E}[\mathcal{L}(W_i, W_j; t, h)|W_i] = \mathbb{E}[\mathcal{L}(W_i, W_j; t, h)|W_j] = 0$ with probability one. Lemma 2.3 shows that $\{\mathcal{L}(\cdot, \cdot; t, h) : t \in [0, \zeta^*], h > 0\}$ is an Euclidean class. Corollary 4 of Sherman (1994) implies that

\[
\sup_{(t,h) \in [0,\zeta^*] \times (0,\infty)} \left| \frac{1}{n} \sum_{i=1}^{n} \sum_{j \neq i} \mathcal{L}(W_i, W_j; t, h) \right| = O_p(1) .
\]
It follows that
\[
\sup_{t \in [0, \zeta^*]} \frac{1}{h_n^d n^{3/2}} \sum_{i=1}^{n} \sum_{j=1}^{n} \mathcal{L}(W_i, W_j; t, h_n) \leq \frac{1}{n^{3/2} h_n^d} \left( \sup_{(t, h) \in [0, \zeta^*] \times (0, \infty)} \left| \sum_{i=1}^{n} \sum_{j \neq i} \mathcal{L}(W_i, W_j; t, h) \right| + \sup_{t \in [0, \zeta^*]} \left| \sum_{i=1}^{n} \mathcal{L}(W_i, W_i; t, h_n) \right| \right) = o_p(1). \tag{2.A.8}
\]

In addition, since \( \bar{m}(x; h_n) = m(x) + O_p(h_n^r) \) by kernel smoothing techniques, we have
\[
\sup_{t \in [0, \zeta^*]} \frac{1}{n^{1/2}} \left| \sum_{i=1}^{n} \xi_{i,t}^* v(X_i) \right| = O_p(n^{1/2} h_n^r) = o_p(1) \tag{2.A.9}
\]
by Assumption B3. Substituting (2.A.8) and (2.A.9) into (2.A.7), we obtain that Term I.b1 is asymptotically uniformly negligible. In addition, after some simple but tedious calculation, we show that under Assumptions K1-K2 and SM1, \( \bar{Q}(X_i; h_n) = O_p(h_n^r) \). It follows that Term I.b2 is asymptotically uniformly negligible; that is,
\[
\sup_{t \in [0, \zeta^*]} \frac{1}{n^{1/2}} \left| \sum_{i=1}^{n} \xi_{i,t}^* \bar{Q}(X_i; h_n) \right| \leq O_p(n^{1/2} h_n^r) = o_p(1). \tag{2.A.10}
\]

It remains to show that Term I.b3 is asymptotically uniformly negligible. For each \( t \in [0, \zeta^*] \) and \( h > 0 \), \( \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} \mathcal{M}(W_i, W_j; t, h) \) is a degenerate U statistic of order 2 because \( \mathbb{E}[\mathcal{M}(W_i, W_j, t, h)|W_i] = \mathbb{E}[\mathcal{M}(W_i, W_j, t, h)|W_j] = 0 \) with probability one. Lemma 2.3 shows that \( \{\mathcal{M}(\cdot, \cdot; t, h) : t \in [0, \zeta^*], h \in (0, 1)\} \) is an Euclidean class; thus, we have
\[
\sup_{(t, h) \in [0, \zeta^*] \times (0, 1)} \left| \frac{1}{n} \sum_{i=1}^{n} \sum_{j \neq i} \mathcal{M}(W_i, W_j; t, h) \right| = O_p(1)
\]
by Corollary 4 of Sherman (1994). We also have
\[
\sup_{t \in [0, \zeta^*]} \frac{1}{n^{3/2} h_n^d} \sum_{i=1}^{n} \mathcal{M}(W_i, W_i; t, h_n) = o_p(1)
\]
because $\mathcal{M}$ is uniformly bounded and $n^{1/2}h_n^d \to \infty$ by Assumptions B1 and B2. It follows that Term I.b3 is asymptotically uniformly negligible; that is,

$$\sup_{t \in [0, \zeta^*]} \left| \frac{1}{n^{3/2}} \sum_{i=1}^{n} \sum_{j=1}^{n} \xi_{i,t}^{*} \left[ Q(X_i, X_j; h_n) - \bar{Q}(X_i; h_n) \right] \right| \leq \frac{1}{n^{3/2}h_n^d} \left( \sup_{(t, h) \in [0, \zeta^*] \times (0, 1)} \left| \sum_{i=1}^{n} \sum_{j \neq i} \mathcal{M}(W_i, W_j; t, h) \right| + \sup_{t \in [0, \zeta^*]} \left| \sum_{i=1}^{n} \mathcal{M}(W_i, W_i; t, h_n) \right| \right) = o_p(1). \quad (2.A.11)$$

Combining (2.A.4)-(2.A.11), we obtain that Term I.b is of order $o_p(1)$ uniformly in $t \in [0, \zeta^*]$. Hence, we obtain

$$\text{Term I} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \xi_{i,t}^{*} \frac{m^{*}(X_i)}{m(X_i)} + o_p(1)$$

where the remainder term is asymptotically uniformly negligible in $t \in [0, \zeta^*]$.

Following similar arguments, we can also show that Term II is asymptotically uniformly negligible.

\begin{proof}

Lemma 2.3. Suppose that the assumptions of Proposition 2.3 hold.

(i) Let $\xi$ be the function defined in Proposition 2.2. The class of functions

$$\{(y, \delta) \mapsto \xi(y, \delta; t) : t \in [0, \zeta^*]\}$$

is Euclidean for some envelope.

(ii) Let $\xi^{*}$ be the function defined in Proposition 2.3. The class of functions

$$\{(y, \delta, x) \mapsto \xi^{*}(y, \delta; t, x) : t \in [0, \zeta^*]\}$$

is Euclidean for some envelope.

(iii) Let $\mathcal{L}$ be the function defined in (2.A.5). The class of functions

$$\{(w_1, w_2) \mapsto \mathcal{L}(w_1, w_2; t, h) : t \in [0, \zeta^*], h > 0\}$$

is Euclidean for some envelope.

(iv) Let $\mathcal{M}$ be the function defined in (2.A.6). The class of functions

$$\{(w_1, w_2) \mapsto \mathcal{M}(w_1, w_2; t, h) : t \in [0, \zeta^*], h \in (0, 1)\}$$

is Euclidean for some envelope.

\end{proof}
Proof. (i) For any \( t, t' \in [0, \xi^*] \), we have
\[
\left| \int_0^t \frac{F^\delta_Y(ds)}{(1 - F_Y(s))^2} - \int_0^{t'} \frac{F^\delta_Y(ds)}{(1 - F_Y(s))^2} \right| \leq \left( \frac{1}{1 - F_Y(\xi^*)} \right)^2 |F^\delta_Y(t) - F^\delta_Y(t')|
\]
\[
\leq \left( \frac{1}{1 - F_Y(\xi^*)} \right)^2 \sup_{s \in [0, \xi^*]} |\partial F^\delta_Y(s) / \partial t| |t - t'|
\]
\[
\leq \left( \frac{1}{1 - F_Y(\xi^*)} \right)^2 \left( \sup_{s \in [0, \xi^*]} f_T(s) \right) |t - t'|
\]
by Assumption I1. Lemmas 2.13 and 2.14 of Pakes and Pollard (1989) imply that the class
\[
\left\{ (y, \delta) \mapsto \int_0^{y,t} \frac{F^\delta_Y(ds)}{(1 - F_Y(s))^2} : t \in [0, \xi^*] \right\}
\]
is Euclidean for the constant envelope \([1 + 2\xi^* \sup_{s \in [0, \xi^*]} f_T(s)] / [1 - F_Y(\xi^*)]^2\) under Assumption SM5 because
\[
\int_0^{y,t} \frac{F^\delta_Y(ds)}{(1 - F_Y(s))^2} = \min \left\{ \int_0^y \frac{F^\delta_Y(ds)}{(1 - F_Y(s))^2}, \int_0^t \frac{F^\delta_Y(ds)}{(1 - F_Y(s))^2} \right\}
\]
In addition, Lemma 19.15 of van der Vaart (1998) implies that the class \( \{ y \mapsto 1_{[y \leq t]} : t \in [0, \xi^*] \} \) is Euclidean for a constant envelope because \( \{ (-\infty, t) : t \in \mathbb{R} \} \) is a VC class. Note that we also have
\[
|F_T(t) - F_T(t')| \leq \left( \sup_{s \in [0, \xi^*]} f_T(s) \right) |t - t'|
\]
for any \( t, t' \in [0, \xi^*] \). Applying Lemmas 2.13 and 2.14 of Pakes and Pollard (1989) again yields that the class \( \{ (y, \delta) \mapsto \xi(y, \delta; t) : t \in [0, \xi^*] \} \) is Euclidean for some envelope.

(ii) For any \( t, t' \in [0, \xi^*] \) and \( x \in J \), we have
\[
|F_{T|X}(t|x) - F_{T|X}(t'|x)| \leq \left( \sup_{s \in [0, \xi^*]} f_{T|X}(s|x) \right) |t - t'|
\]
and
\[
\left| \int_0^t \frac{F^\delta_{Y|X}(ds|x)}{(1 - F_{Y|X}(s|x))^2} - \int_0^{t'} \frac{F^\delta_{Y|X}(ds|x)}{(1 - F_{Y|X}(s|x))^2} \right| \leq \left( \frac{1}{v^*} \right)^2 |F^\delta_{Y|X}(t|x) - F^\delta_{Y|X}(t'|x)|
\]
\[
\leq \left( \frac{1}{v^*} \right)^2 \sup_{s \in [0, \xi^*]} |\partial F^\delta_{Y|X}(s|x) / \partial t| |t - t'|
\]
\[
\leq \left( \frac{1}{v^*} \right)^2 \left( \sup_{s \in [0, \xi^*]} f_{T|X}(s|x) \right) |t - t'|
\]
by Assumption I2. From Lemmas 2.13 and 2.14 of Pakes and Pollard (1989) and under Assumption SM5, the class \( \{ x \mapsto 1 - F_T|X(t|x) : t \in [0, \zeta^*] \} \) is Euclidean for the envelope 
\( 1 + 2 \zeta^* \sup_{s \in [0, \zeta^*]} f_{T|X}(s|x) \), and the class

\[
\left\{ (y, \delta, x) \mapsto \int_0^{yu/t} \frac{F_{Y|X}(ds|x)}{(1 - F_{Y|X}(s|x))^2} : t \in [0, \zeta^*] \right\}
\]

is Euclidean for the envelope \( [1 + 2 \zeta^* \sup_{s \in [0, \zeta^*]} f_{T|X}(s|x)] / (u^*)^2 \) because

\[
\int_0^{yu/t} \frac{F_{Y|X}(ds|x)}{(1 - F_{Y|X}(s|x))^2} = \min \left\{ \int_0^y \frac{F_{Y|X}(ds|x)}{(1 - F_{Y|X}(s|x))^2}, \int_0^t \frac{F_{Y|X}(ds|x)}{(1 - F_{Y|X}(s|x))^2} \right\}.
\]

Since \( \{(-\infty, t] : t \in \mathbb{R} \} \) is a VC class, Lemma 19.15 of van der Vaart (1998) implies that the class \( \{(y, \delta, x) \mapsto \frac{1_{y \leq x} \delta}{1 - F_{Y|X}(y|x)} : t \in [0, \zeta^*] \} \) is Euclidean for a constant envelope. Therefore, the class of functions \( \{(y, x, \delta) \mapsto \xi^*(y, \delta; t, x) : t \in [0, \zeta^*] \} \) is Euclidean for some envelope by Lemma 2.14 of Pakes and Pollard (1989).

(iii) Let \( g(x_1, x_2) = x_1 - x_2 \) and \( K = \{ u \mapsto K(u/h) : h > 0 \} \). Assumption K1 implies that \( K \) is a VC-subgraph class. (See the discussion on page 911 of Giné and Guillou (2002).) It follows from Lemma 2.6.18(vii) of van der Vaart and Wellner (1996) that the class \( \mathcal{F} \circ g = \{(x_1, x_2) \mapsto K((x_1 - x_2)/h) : h > 0 \} \) is a VC-subgraph class; thus \( \mathcal{F} \circ g \) is Euclidean for a constant envelope. Lemma 5 of Sherman (1994) implies that the class \( \{x \mapsto h^d \tilde{m}(x; h) : h > 0 \} \) is also Euclidean for a constant envelope. Since the class \( \{(y, \delta, x) \mapsto \xi^*(y, \delta; t, x) : t \in [0, \zeta^*] \} \) is Euclidean by part (ii), the class \( \{(w_1, w_2) \mapsto L(w_1, w_2; t, h) : t \in [0, \zeta^*], h > 0 \} \) is Euclidean for some envelope by Lemma 2.14 of Pakes and Pollard (1989).

(iv) By part (ii), Lemma 5 of Sherman (1994), and Lemma 2.14 of Pakes and Pollard (1989), it suffices to show that \( \{(x_1, x_2) \mapsto h^d Q(x_1, x_2; h) : h \in (0, 1) \} \) is Euclidean. Since there is a constant \( c \) such that for \( h_1, h_2 \in (0, 1), \)

\[
\sup_{x, z \in J} |v(x + h_1 z) - v(x + h_2 z)| \leq c|h_1 - h_2|
\]

by Assumptions SM1 and SM4, the class \( \{(x, z) \mapsto v(x + h z) : h \in (0, 1) \} \) is Euclidean by Lemma 2.13 of Pakes and Pollard (1989). In addition, Lemma 22(i) of Nolan and Pollard (1987) imply that the class

\[
\left\{ (x_1, x_2, z) \mapsto K\left( \frac{x_1 - x_2}{h} + z \right) : h \in (0, 1) \right\}
\]
is Euclidean by Assumptions K1 and K4. Let $\mathcal{U}$ be the measure on $[-1, 1]^d$ associated with a uniform random variable on $[-1, 1]^d$. Note that

$$h^d Q(x_1, x_2)$$

$$= \int 2^d K(z) \left[ v(x_1 + h z) K \left( \frac{x_1 - x_2}{h} + z \right) - v(x_1) K \left( \frac{x_1 - x_2}{h} \right) \right] \mathcal{U}(dz).$$

It follows from Lemma 2.14 of Pakes and Pollard (1989) and Lemma A.2. of Ghosal et al. (2000) that the class \( \{ (x_1, x_2) \mapsto h^d Q(x_1, x_2) : h \in (0, 1) \} \) is Euclidean. \qed
Chapter 3

Model Selection in Utility-Maximizing Binary Prediction

Abstract

The semiparametric maximum utility (MU) estimation proposed by Elliott and Lieli (2013) can be viewed as cost-sensitive binary classification; thus, its in-sample overfitting issue is similar to that of perceptron learning in the machine learning literature. Based on structural risk minimization, a utility-maximizing prediction rule (UMPR) is constructed to alleviate the in-sample overfitting of the MU estimation. Simulation results show that the UMPR, in comparison to the Lasso-logit, may have larger relative expected utility if the conditional probability of the binary outcome is misspecified.

3.1 Introduction

Making a binary decision based on an uncertain binary outcome is common in modern economic activities. For instance, an investor who considers buying a financial instrument may tend to predict the direction of its price change in the future and decide to buy the instrument if the price is predicted to rise. As suggested by Granger and Machina (2006), decision-making based on the prediction of a binary outcome should be driven by the preference of the decision maker. On the one hand, the utility arising from a mismatch between the binary decision and outcome may differ in the realized outcome; on the other hand, the utility may be affected by observable covariates. In making financial investment decisions, the disutility for the investor who buys the instrument but suffers from a decrease in the price may be greater than that for the investor who does not buy the instrument but finds an increase in the price. In addition, features
of the instrument, for example measures of its price volatility, may affect not only the likelihood of price change but also the investor’s utility.\(^1\)

Although a decision maker’s preference would be important for the decision-making based on binary prediction, traditional methods of pattern classification rarely take the decision maker’s utility into consideration. To incorporate the decision maker’s utility into binary classification, Elliott and Lieli (2013) propose a maximum utility approach. Instead of globally estimating the conditional probability \(p^*(x) \equiv P(Y = 1 | X = x)\), they show the utility-maximizing binary classification problem can be solved by only estimating the sign of \(p^*(x) - c(x)\), where \(c(\cdot)\) is the cutoff function determined by the decision maker’s utility function. Compared to maximum likelihood estimation, their maximum utility estimation is, however, prone to in-sample overfitting.

In this chapter, we show that the in-sample overfitting of maximum utility estimation is similar to that of perceptron learning in the machine learning literature. To alleviate the tendency of fitting the in-sample noise by sophisticated models, we pre-specify a hierarchy of classes of (finite-dimensional) functions and consider a utility-maximizing prediction rule (UMPR), which is a maximum utility estimator that maximizes a complexity penalized empirical utility. Following the suggestion on the complexity penalty in Bartlett, Boucheron, and Lugosi (2002), we demonstrate in Corollary 3.1 that the difference between the maximal expected utility and the (generalized) expected utility of the UMPR can be bounded by an almost optimal trade-off between the complexity penalty and the approximation error, that is, an error due to the approximation of functions in a hierarchy of classes to an optimal decision rule. Hence, whenever the approximation error is equal to zero for some class of functions, the relative expected utility of the UMPR increases as the sample size tends to infinity. This theoretical property is supported by simulation results. Simulation results also suggest that the UMPR may have better performance in terms of relative expected utility than the Lasso-logit, which is a common machine learning predictor, if the conditional probability of the binary outcome is misspecified.

The structure of the remaining chapter is as follows. Section 3.2 describes the maximum utility estimation in Elliott and Lieli’s (2013) model and the issue of its in-sample overfitting. Section 3.3 presents the construction of a utility-maximizing prediction rule and a probabilistic lower bound of its expected utility when the training data are independent and identically distributed. In Section 3.4, a Monte Carlo experiment is carried out to evaluate the finite-sample performance of this utility-maximizing prediction rule. Section 3.5 concludes. Proofs are given in Appendix 3.A.

\(^{1}\)Barberis and Xiong (2012) propose a model to explain the individual investor preference for volatile stocks.
3.2 Maximum utility estimation

3.2.1 Model

We start by describing Elliott and Lieli’s (2013) model of binary decision-making based on binary prediction: Before the realization of a binary outcome \( Y \in \{-1, 1\} \), a decision maker aims to choose a binary decision \( a \in \{-1, 1\} \) to maximize his or her expected utility conditional on a \( d \)-dimensional vector of observed covariates \( X = x \). Concretely, the decision maker solves the optimization problem

\[
\max_{a \in \{-1, 1\}} \mathbb{E}[U(a, Y, X) \mid X = x]. \tag{3.1}
\]

We abbreviate by writing \( u_{a,y}(x) = U(a, y, x) \) for notational simplicity and make the following assumptions that are imposed in Elliott and Lieli (2013).

**Assumptions**

A1 The conditional probability \( p^*(x) \equiv \mathbb{P}(Y = 1 \mid X = x) \) does not depend on the decision \( a \).

A2 For all \( x \) in the support \( \mathcal{X} \subseteq \mathbb{R}^d \) of \( X \), \( u_{1,1}(x) > u_{-1,1}(x) \) and \( u_{-1,-1}(x) > u_{1,-1}(x) \).

A3 For any \( a, y \in \{1, -1\} \), \( u_{a,y}(\cdot) \) is Borel measurable; in addition, there is some \( M > 0 \) such that \( |u_{a,y}(x)| \leq M \) for all \( x \in \mathcal{X} \) and \( a, y \in \{1, -1\} \).

Assumption A1 excludes the possibility of feedback from the binary action to the binary outcome. Take the financial investment in Section 3.1 as an example. Under Assumption A1, investors are price takers whose decisions on buying an instrument do not affect the possibility of price change. Assumption A2 implies that the decision maker obtains higher utility when the decision matches the outcome. This assumption seems plausible in many situations, for example the aforementioned financial investment. Although Assumption A3 imposes a uniform bound on the utility functions, this condition is in general not restrictive because the utility function is used to characterize an individual’s ordinal preference.

Elliott and Lieli (2013) show that under Assumptions A1 and A2, we can obtain an optimal decision rule (after observing \( X = x \))

\[
a^*(x) = \begin{cases} 
1 & \text{if } p^*(x) \geq c(x), \\
-1 & \text{otherwise},
\end{cases} \tag{3.2}
\]

where

\[
c(x) \equiv \frac{u_{-1,-1}(x) - u_{1,-1}(x)}{u_{1,1}(x) - u_{-1,1}(x) + u_{-1,-1}(x) - u_{1,-1}(x)}
\]
is a cutoff function derived from the utility function, which is known in principle to the
decision maker. Thus, knowledge of the correct conditional probability $p^*(x)$ yields the
maximal expected utility of the decision.

Elliott and Lieli’s (2013) insight is that the correct specification of $\text{sign}(p^*(x) - c(x))$
rather than that of $p^*(x)$ is enough to achieve maximal expected utility in (3.1); specifically,
the knowledge of crossing points between the conditional probability $p^*(x)$ and the
cutoff $c(x)$ is sufficient. Moreover, they point out that under Assumptions A1 and A2,
the decision-making problem in (3.1) can be equivalently written as

$$\max_{f} S(f) = \max_{f} \mathbb{E}[b(X)[Y + 1 - 2c(X)] \text{sign}(f(X) - c(X))],$$

(3.3)

where

$$b(x) \equiv u_{1,1}(x) - u_{-1,1}(x) + u_{-1,-1}(x) - u_{1,-1}(x)$$

is the denominator of $c(x)$ and the maximum is taken over all measurable functions
from $\mathcal{X}$ to $\mathbb{R}$.\footnote{For any real number $z$, let $\text{sign}(z) = 1$ if $z \geq 0$ and $\text{sign}(z) = -1$ otherwise.}

Given observations $\{(Y_i, X_i)\}_{i=1}^n$ with the sample size $n$ and a pre-specified class $\mathcal{F}$ of
functions indexed by a finite-dimensional parameter, the maximum of the sample analog,
restricted to $\mathcal{F}$, of (3.3) is achieved by some function $\hat{f}$, which is called a maximum
utility (MU) estimator.\footnote{Since $U(a, y, x) = 0.25b(x)[y + 1 - 2c(x)]a + 0.25b(x)[y + 1 - 2c(x)] + u_{-1, y}(x)$, we normalize the utility
function by setting $u_{-1, y}(x) = -0.25b(x)[y + 1 - 2c(x)]$ for all $x \in \mathcal{X}$ and call $S(f)$ the expected utility of $f$ for
convenience of exposition.}

Concretely,

$$S_n(\hat{f}) = \max_{f \in \mathcal{F}} S_n(f) = \max_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n b(X_i)[Y_i + 1 - 2c(X_i)] \text{sign}(f(X_i) - c(X_i)).$$

(3.4)

The Monte Carlo simulation in Elliott and Lieli (2013) shows that the maximum utility estimation in (3.4), compared to traditional maximum likelihood approaches,
achieves a large improvement in utility especially when the conditional probability $p^*(\cdot)$
is misspecified. However, the in-sample performance of the maximum utility estimation
may be attributed to the overfitting. Elliott and Lieli further make the following comment:

\footnote{Multiplicity of the maximum utility estimator, say the maximizer $\hat{f}$ of (3.4), could be present. Similarly, the
set of maximizers of $S(f)$ over $\mathcal{F}$ may not be a singleton. The analysis in this chapter thus emphasizes the
properties of optimand functions. See Elliott and Lieli (2013) for the discussion about the lack of identification
of optimizers.}
Both ML and MU have a strong tendency to overfit in sample, however the problem seems more severe for the MU method. This creates challenges for model selection.\(^5\)

### 3.2.2 Nature of the overfitting in MU estimation

The in-sample overfitting of maximum utility estimation is similar to that of perceptron learning in the machine learning literature.\(^6\) More precisely, the optimization problem in (3.3) can be viewed as perceptron learning in which the cost of misclassification for each training example may be different. To see this, note that

\[
S(f) = \mathbb{E}[b(X)[Y(1 - 2c(X)) + 1] \text{sign}(f(X) - c(X))]
\]

\[
= \mathbb{E}[b(X)[Y(1 - 2c(X)) + 1]] - 2 \mathbb{E}[b(X)[Y(1 - 2c(X)) + 1] \mathbb{1}_{[Y \neq \text{sign}(f(X) - c(X))]}].
\]

(3.5)

The first term \(G\) in (3.5) represents the (double) expected gain in utility from a correct match between the decision and outcome, whereas the second term \(L(f)\) reflects the (double) expected loss in utility from a mismatch between the decision and outcome.\(^7\) So, we have

\[
\max_f S(f) = G - 2 \min_f L(f)
\]

\[
= G - 2 \min_f \mathbb{E}[b(X)[Y(1 - 2c(X)) + 1] \mathbb{1}_{[Y \neq \text{sign}(f(X) - c(X))]}].
\]

(3.6)

and for the maximum utility estimator in (3.4),

\[
L_n(\hat{f}) = \min_{f \in \mathcal{F}} L_n(f)
\]

\[
= \min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} b(X_i)[Y_i(1 - 2c(X_i)) + 1] \mathbb{1}_{[Y_i \neq \text{sign}(f(X_i) - c(X_i))]}.
\]

(3.7)

---

\(^5\) When the data are independent and identically distributed, Elliott and Lieli (2013) propose a pretest to select MU estimators but do not investigate theoretical properties of their “post-model-selection” MU estimator.

\(^6\) The simple perceptron learning is a method of binary pattern recognition that establishes classification based on the threshold function \(f(x) = \text{sign}(\theta_1^\top x - \theta_0)\), where \(\theta_1 \in \mathbb{R}^d\) and \(\theta_0 \in \mathbb{R}\). More variants of perceptron are well documented in Vapnik (1995) and Anthony and Bartlett (1999).

\(^7\) To see this, notice that for any \(x \in \mathcal{X}\),

\[
b(x)[y(1 - 2c(x)) + 1] = \begin{cases} 2(u_{1,1}(x) - u_{-1,1}(x)) & \text{if } y = 1, \\ 2(u_{-1,-1}(x) - u_{1,-1}(x)) & \text{otherwise.} \end{cases}
\]
The perceptron learning is a special case of (3.7) because the cost of misclassification is identical for each observation whenever there is a constant  \( \bar{u} \in \mathbb{R}_+ \) such that  \( u_{1,1}(x) - u_{-1,1}(x) = u_{-1,-1}(x) - u_{1,-1}(x) = \bar{u} \) for all  \( x \in \mathcal{X} \). Furthermore, even though the cost of misclassification may be different for each observation (\( Y_i, X_i \)), when the in-sample observations (training data set) can be perfectly separated by \( \mathcal{F} \) (i.e., classification without error), the optimization problem in (3.7) boils down to perceptron learning. This is because in this case, the cost of misclassification \( b(X_i)[Y_i(1 - 2c(X_i)) + 1] \) has no effect on \( \min_{f \in \mathcal{F}} L_n(f) \):

\[
L_n(f^\dagger) = \min_{f \in \mathcal{F}} L_n(f) = 0 = \min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}[Y_i \neq \text{sign}(f(X_i) - c(X_i))]
\]

for some prediction rule  \( f^\dagger \in \mathcal{F} \). Although perfect separation of the in-sample observations could be accomplished by a sufficiently large class of functions, such sophisticated models will also fit the (in-sample) noise and thus worsen the out-of-sample performance.

### 3.3 Model selection

Motivated by the possible in-sample overfitting, we adopt the structural risk minimization approach in machine learning to investigate model selection in cost-sensitive binary classification; specifically, our goal is to alleviate the overfitting by adding an appropriate complexity penalty to the objective function in the optimization.

We first introduce notation. Let

\[
\ell(Y,X,f) = b(X)[Y(1 - 2c(X)) + 1]\mathbb{I}[Y \neq \text{sign}(f(X) - c(X))]
\]

be the loss function of cost-sensitive binary classification, which is derived from the utility-maximizing problem in (3.1). A sample of i.i.d. observations with sample size  \( n \) is denoted by  \( \mathcal{D}_n \equiv \{(Y_i, X_i)\}_{i=1}^n \). Given a prediction rule  \( f \) constructed based on  \( \mathcal{D}_n \), let  \( L(f) = \mathbb{E}[\ell(Y, X, f) \mid \mathcal{D}_n] \) and  \( L_n(f) = \frac{1}{n} \sum_{i=1}^{n} \ell(Y_i, X_i, f) \) be the expected loss and the empirical loss of the prediction rule  \( f \), respectively. Consider a hierarchy of classes \( \{\mathcal{F}_k\}_{k=1}^\infty \) of functions; i.e.,

\[
\mathcal{F}_1 \subset \mathcal{F}_2 \subset \cdots \subset \mathcal{F}_k \subset \cdots \text{ and } \mathcal{F} \equiv \bigcup_{k=1}^{\infty} \mathcal{F}_k.
\]

---

8 In this case, if  \( f(x) \) is further parameterized as  \( f(x) = \theta^\top x + \theta_0 \), the maximum utility estimation reduces to Manski’s (1975, 1985) maximum score estimation, which is well-known in econometrics.

9 The expectation involved in the definition of  \( L(f) \) is taken with respect to an observation  \( (Y, X) \), which is independent of  \( \mathcal{D}_n \). Note that  \( L(f) \) could be random because of the random sample  \( \mathcal{D}_n \). We suppress the dependence of  \( L(f) \) on  \( \mathcal{D}_n \) for convenience of exposition.
A maximum utility estimator selected from $\mathcal{F}_k$ is denoted by

$$\hat{f}_k \in \arg \max_{f \in \mathcal{F}_k} S_n(f) = \arg \min_{f \in \mathcal{F}_k} L_n(f),$$

where “arg” stands for the set of estimators in $\mathcal{F}_k$ that achieve the optimum.

Proposition 3.1 below shows that for sufficiently large sample size $n$, the difference between $L_n(\hat{f}_k)$ and $L(\hat{f}_k)$ is small with high probability if the growth function $\Pi_{k,c}()$ of the collection $\mathcal{F}_{k,c} \equiv \{ x \mapsto \text{sign}(f(x) - c(x)) : f \in \mathcal{F}_k \}$ is of polynomial order for each $k$.

**Proposition 3.1.** Suppose the data $D_n = \{(Y_i, X_i)\}_{i=1}^n$ are independent and identically distributed. Under Assumptions A1-A3, we have for any $n, k \in \mathbb{N}$ and $\varepsilon > 0$,

$$\mathbb{P} \left( \sup_{f \in \mathcal{F}_k} |L(f) - L_n(f)| > \varepsilon \right) \leq 4(2n + 1)\Pi_{k,c}(2n) \exp \left\{ -\frac{n\varepsilon^2}{16M^2} \right\},$$

where $\Pi_{k,c}()$ is the growth function of the collection $\mathcal{F}_{k,c}$.

The growth function $\Pi_{k,c}()$ is of polynomial order if $\mathcal{F}_k$ is a vector space of real-valued functions for each $k \in \mathbb{N}$. Indeed, by Theorems 3.5-3.7 of Anthony and Bartlett (1999), for $2n \geq \dim(\mathcal{F}_k)$,

$$\Pi_{k,c}(2n) \leq \sum_{i=0}^{\dim(\mathcal{F}_k)} \left( \begin{array}{c} 2n \\ i \end{array} \right) \leq \left( \frac{2ne}{\dim(\mathcal{F}_k)} \right)^{\dim(\mathcal{F}_k)},$$

where $\dim(\mathcal{F}_k)$ is the dimension of $\mathcal{F}_k$. To ensure the dimension of $\mathcal{F}_k$ is finite, we can choose (finite-dimensional) nondecreasing linear sieves of univariate functions and generate linear sieves of multivariate functions by tensor-product construction. In this case, if the conditional probability $p^*(\cdot)$ is smooth enough, then $\inf_{f \in \mathcal{F}_k} \sup_{x \in \mathcal{X}} |f(x) - p^*(x)| \to 0$ as $k \to \infty$, and different sieve approximation error rates can be found in Chen (2007). Elliott and Lieli (2013) illustrate some preferences and data generating processes of $(Y,X)$ in which finite order polynomial functions in $X$ can completely replicate the crossing points between $p^*(x)$ and $c(x)$; more precisely, there is some polynomial function $f_0$ with sufficient order such that $\text{sign}(f_0(x) - c(x)) = \text{sign}(p^*(x) - c(x))$. To put this in the language of machine learning, let $\mathcal{P}_k$ be the class of polynomial transformations on $X$.

---

10 For any collection $\mathcal{H}$ of functions from $\mathcal{X}$ to $\{-1, 1\}$, the growth function $\Pi : \mathbb{N} \to \mathbb{N}$ of $\mathcal{H}$ is

$$\Pi(m) = \max_{(x_1, \ldots, x_m) \in \mathcal{X}^m} |\{(h(x_1), \ldots, h(x_m)) : h \in \mathcal{H}\}|.$$

That is, the growth function $\Pi(m)$ is the maximum number of distinct ways in which $m$ points $(x_1, \ldots, x_m)$ can be classified using functions in $\mathcal{H}$.
\( \mathcal{X} \subseteq \mathbb{R}^d \) of order at most \( k \).\(^{11}\) It follows that

\[
L^* \equiv \inf_f L(f) = \inf_{f \in \mathcal{P}_k} L(f) \quad \text{and} \quad S^* \equiv \sup_f S(f) = \sup_{f \in \mathcal{P}_k} S(f)
\]

for some \( k \in \mathbb{N} \). In addition, the dimension of \( \mathcal{P}_k \) is \( D(d, k) = \binom{d + k}{k} \) by Theorem 11.8 of Anthony and Bartlett (1999).

We can also consider the logit specification of \( p^*(x) \), which has been used in many applications. Consider the standard logistic function \( \Lambda(u) = (1 + \exp\{-u\})^{-1} \) for all \( u \in \mathbb{R} \). Let \( \Lambda(\mathcal{P}_k) \equiv \{ x \mapsto \Lambda(f(x)) : f \in \mathcal{P}_k \} \) for each \( k \in \mathbb{N} \). Note that if the cutoff function \( c(\cdot) \) is the composition of a sigmoid function \( s_1(\cdot) \) over a polynomial transformation \( f_1(\cdot) \); i.e., \( c(x) = s_1(f_1(x)) \), then any function in \( \mathcal{F}_{k,c} = \{ x \mapsto \text{sign}(f(x) - c(x)) : f \in \Lambda(\mathcal{P}_k) \} \) is a two-layer sigmoid neural network.\(^{12}\) It follows by Theorem 8.13 of Anthony and Bartlett (1999) that for \( 2n \geq D(d, k) \),

\[
\Pi_{k,c}(2n) \leq 2^{3D(d, k)^2/2} [162D(d, k)]^{15D(d, k)} \left( \frac{2n \epsilon}{D(d, k)} \right)^{D(d, k)}.
\]

So, the growth function of \( \mathcal{F}_{k,c} \) in this case is of polynomial order.

For each \( k \in \mathbb{N} \), if the growth function \( \Pi_{k,c}(\cdot) \) in Proposition 3.1 is of polynomial order, then both \( \sup_{f \in \mathcal{F}_k} |L(f) - L_n(f)| \) and \( \sup_{f \in \mathcal{F}_k} |S(f) - S_n(f)| \) converge almost surely to 0;\(^{13}\) hence, both \( |L(\hat{f}_k) - L_n(\hat{f}_k)| \) and \( |S(\hat{f}_k) - S_n(\hat{f}_k)| \) converge almost surely to 0. Note that given i.i.d. observations, such almost sure convergence can be guaranteed whenever the growth function \( \Pi_{k,c}(\cdot) \) is of polynomial order; in fact, technical conditions, imposed by Proposition 2 of Elliott and Lieli (2013) such as compactness of parameter space and lipschitz continuity of functions with respect to the parameter, can be relaxed.

Moreover, Proposition 3.1 shows that for any \( n, k \in \mathbb{N} \) and \( \delta \in (0, 1) \), with probability at least \( 1 - \delta \),

\[
L(f) \leq L_n(f) + 4M \sqrt{\frac{\log \left( \frac{4(2n + 1)\Pi_{k,c}(2n)}{\delta} \right)}{n}}
\]

A polynomial transformation on \( X \subseteq \mathbb{R}^d \) of degree at most \( k \) is a function of the form \( f(x) = c_0 + \sum_{j=1}^q c_j \varphi_j(x) \), where \((c_0, c_1, \ldots, c_q) \in \mathbb{R}^{d+1} \) and \( \varphi_j(x) = \prod_{i=1}^d x_{p_{j,i}} \) with \( \sum_{i=1}^d p_{j,i} \leq k \) for each \( j \) and \( q \in \mathbb{N} \).

A sigmoid function \( s : \mathbb{R} \to \mathbb{R} \) is a nondecreasing function that satisfies \( \lim_{u \to -\infty} s(u) = 0 \) and \( \lim_{u \to \infty} s(u) = 1 \). A multi-layer sigmoid neural network is a cascade of perceptrons. See Anthony and Bartlett (1999) for the rigorous definition of a multi-layer sigmoid neural network and its applications in machine learning.

\(^{11}\) In fact, we have for any \( n, k \in \mathbb{N} \) and \( \varepsilon, \eta > 0 \),

\[
P \left( \sup_{f \in \mathcal{F}_k} |S(f) - S_n(f)| > \varepsilon \right) \leq 2 \exp \left\{ -\frac{n(1 - \eta)^2 \varepsilon^2}{8M^2} \right\} + 4(2n + 1)\Pi_{k,c}(2n) \exp \left\{ -\frac{n\eta^2 \varepsilon^2}{64M^2} \right\}
\]

by Proposition 3.1 and Hoeffding’s inequality. The desired almost sure convergence follows by Borel-Cantelli Lemma.
for all \( f \in \mathcal{F}_k \). Thus, given sample size \( n \), an increase in model complexity (say \( k \)) tends to decrease empirical loss \( L_n \); however, it may meanwhile increase \( \Pi_{k,c}(2n) \). In other words, small empirical loss arising from sophisticated models does not guarantee small expected loss.

For each \( k \), we consider an estimate of expected loss \( L(\hat{f}_k) \) to be

\[
R_{n,k} \equiv L_n(\hat{f}_k) + 4M \sqrt{\frac{\log \{4(2n+1)\psi(k,n)\}}{n}}
\]

for some (easily computable) upper bound \( \psi(k,n) \) of \( \Pi_{k,c}(2n) \). It follows by Proposition 3.1 that we obtain an upper bound on the tail probability for \( L(\hat{f}_k) - R_{n,k} \) for each \( k \). To be specific, we have for any \( n,k \in \mathbb{N} \) and \( \varepsilon > 0 \),

\[
\begin{align*}
\mathbb{P}(L(\hat{f}_k) > R_{n,k} + \varepsilon) & \leq \mathbb{P}\left( \sup_{f \in \mathcal{F}_k} |L(f) - L_n(f)| > 4M \sqrt{\frac{\log \{4(2n+1)\psi(k,n)\}}{n}} + \varepsilon \right) \\
& \leq 4(2n+1)\Pi_{k,c}(2n) \exp \left\{ -n \left[ 4M \sqrt{\frac{\log \{4(2n+1)\psi(k,n)\}}{n}} + \varepsilon \right]^2 / 16M^2 \right\} \\
& \leq \exp \left\{ -\frac{n\varepsilon^2}{16M^2} \right\};
\end{align*}
\]

(3.9)

where the second inequality holds by Proposition 3.1. Following the suggestion in Bartlett et al. (2002), we consider the distribution-free complexity penalty

\[
C_n(k;\alpha,\psi) \equiv R_{n,k} - L_n(\hat{f}_k) + 4M \sqrt{\frac{(1+\alpha)\log\{k+1\}}{n}} = 4M \left( \sqrt{\frac{\log \{4(2n+1)\psi(k,n)\}}{n}} + \sqrt{\frac{(1+\alpha)\log\{k+1\}}{n}} \right)
\]

and the associated complexity penalized empirical loss for \( f \in \mathcal{F}_k \) to be

\[
\tilde{L}_n(f;k) \equiv L_n(f) + C_n(k;\alpha,\psi)
\]

for some \( \alpha > 0 \).

We define a utility-maximizing prediction rule (UMPR) as a maximum utility estimator \( \hat{f}_k \) in (3.8) that minimizes \( \tilde{L}_n(\hat{f}_k;k) \); specifically,

\[
\tilde{f}_n = \hat{f}_{k^*} \quad \text{where} \quad k^* \in \arg \min_{k=1,2,\ldots} \tilde{L}_n(\hat{f}_k;k).
\]

(3.10)

Equivalently, this utility-maximizing prediction rule \( \tilde{f}_n \) is a maximum utility estimator \( \hat{f}_{k^*} \) that achieves the maximal complexity penalized empirical utility \( \tilde{S}_n(\hat{f}_{k^*};k^*) \equiv \)
\( G_n - 2\tilde{L}_n(\hat{f}_k^*; k^*) \), where \( G_n \) is the sample analog of \( G \) in (3.5). We abbreviate by writing \( \tilde{L}_n(\hat{f}_n) \equiv \tilde{L}_n(\hat{f}_k^*; k^*) \) and \( \tilde{S}_n(\hat{f}_n) \equiv \tilde{S}_n(\hat{f}_k^*; k^*) \). The following theorem presents properties of the utility-maximizing prediction rule \( \hat{f}_n \).

**Theorem 3.1.** Suppose that (i) the data \( \mathscr{D}_n = \{(Y_i, X_i)\}_{i=1}^n \) are independent and identically distributed, and (ii) Assumptions A1-A3 hold.

1. For any \( n \in \mathbb{N} \) and \( \alpha, \varepsilon > 0 \),
\[
P \left( L(\hat{f}_n) - \tilde{L}_n(\hat{f}_n) > \varepsilon \right) \leq \left( \frac{2 + \alpha}{2\alpha} \right) \exp \left\{ -\frac{n\varepsilon^2}{16M^2} \right\}.
\]

2. For any \( n \in \mathbb{N} \) and \( \alpha > 0 \),
\[
E \left[ L(\hat{f}_n) \right] - L^* \\
\leq \min_k \left\{ C_n(k; \alpha, \psi) + \left( \inf_{f \in \mathcal{F}_k} L(f) - L^* \right) \right\} + 4M \sqrt{\frac{1 + \log\{(2 + \alpha)/2\alpha\}}{n}},
\]

where \( L^* = \inf_f L(f) \) is the minimal expected loss, and the expectation is taken with respect to the sample \( \mathscr{D}_n \) of i.i.d. observations.

We immediately have the following corollary by Hoeffding’s inequality and simple algebraic rearrangements.

**Corollary 3.1.** Suppose the assumptions of Theorem 3.1 hold.

1. For any \( n \in \mathbb{N} \), \( \eta \in (0, 1) \), and \( \alpha, \varepsilon > 0 \),
\[
P \left( \tilde{S}_n(\hat{f}_n) - S(\hat{f}_n) > \varepsilon \right) \\
\leq \exp \left\{ -\frac{n(1 - \eta)^2\varepsilon^2}{8M^2} \right\} + \left( \frac{2 + \alpha}{2\alpha} \right) \exp \left\{ -\frac{n\eta^2\varepsilon^2}{64M^2} \right\}.
\]

2. For any \( n \in \mathbb{N} \) and \( \alpha > 0 \),
\[
S^* - E[S(\hat{f}_n)] \\
\leq \min_k \left\{ 2C_n(k; \alpha, \psi) + \left( S^* - \sup_{f \in \mathcal{F}_k} S(f) \right) \right\} + 8M \sqrt{\frac{1 + \log\{(2 + \alpha)/2\alpha\}}{n}},
\]

where \( S^* = \sup_f S(f) \) is the maximal expected utility, and the expectation is taken with respect to the sample \( \mathscr{D}_n \) of i.i.d. observations.
Substituting $\eta = 3/4$ in Corollary 3.1 yields that for any $\alpha, \varepsilon > 0,$
\[
\mathbb{P}\left( \hat{S}_n(f_n) - S(f_n) > \varepsilon \right) \leq \exp \left\{ -\frac{n\varepsilon^2}{128M^2} \right\} + \left( \frac{2 + \alpha}{2\alpha} \right) \exp \left\{ -\left( \frac{9}{8} \right) \frac{n\varepsilon^2}{128M^2} \right\}
\leq \left( \frac{2 + 3\alpha}{2\alpha} \right) \exp \left\{ -\frac{n\varepsilon^2}{128M^2} \right\}.
\]
Setting the right-hand side to be equal to $\delta \in (0,1),$ we obtain
\[
S(f_n) \geq \tilde{S}_n(f_n) - 8M \sqrt{\frac{2\log\{3(\alpha + 2)/2\alpha\delta\}}{n}}
\]
with probability at least $1 - \delta$. In addition to the probabilistic lower bound of the expected utility $S(f_n),$ Corollary 3.1 also shows an upper bound for the difference between the maximal expected utility $S^*$ and the generalized expected utility $\mathbb{E}[S(f_n)]$. This upper bound takes into account the trade-off between the complexity penalty $2C_n(k; \alpha, \psi)$ and the approximation error $S^* - \sup_{f \in \mathcal{F}_k} S(f)$. Hence, if the approximation error is equal to zero for some $k$, then this upper bound shrinks to zero as the sample size tends to infinity.

### 3.4 Simulation

To study the finite-sample performance of the utility-maximizing prediction rule in (3.10), we carried out a Monte Carlo experiment. The simulation designs are those in Elliott and Lieli (2013). Specifically, we consider two data generating processes:

**DGP 1** The covariate $X$ follows the distribution $5 \cdot \text{beta}(1,1.3) - 2.5$ and $p^*(X) = \Lambda(-0.5X + 0.2X^3);$

**DGP 2** Both covariates $X_1$ and $X_2$ are independent and uniformly distributed on $[-3.5,3.5]$ and $p^*(X_1, X_2) = \Lambda(Q(1.5X_1 + 1.5X_2)),$ where $Q(u) = (1.5 - 0.1u) \exp\{-(0.25u + 0.1u^2 - 0.04u^3)\}.$

In addition, we consider four preferences:

**Pref 1.** $b(X) = 20$ and $c(X) = 0.5;$

**Pref 2.** $b(X) = 20$ and $c(X) = 0.5 + 0.025X;$

**Pref 3.** $b(X_1, X_2) = 20$ and $c(X_1, X_2) = 0.75;$

**Pref 4.** $b(X_1, X_2) = 20 + 40 \cdot 1_{|X_1+X_2|<1.5}$ and $c(X_1, X_2) = 0.75.$
The first two preferences are associated with DGP 1, whereas the last two preferences are associated with DGP 2. For DGP 1 together with either preference 1 or 2, not only the cubic ML-logit but also the cubic MU are correctly specified because there are three crossing points between the conditional probability $p^*(\cdot)$ and the cutoff $c(\cdot)$ in the support of $X$.\textsuperscript{14} Although any logit is clearly misspecified for DGP 2, Elliott and Lieli (2013) demonstrate that the cubic MU is correctly specified.

For UMPR and MU estimators, we specify the hierarchy $\mathcal{F}_k = \mathcal{P}_k$ for $k \in \{1, 2\}$ and $\mathcal{F}_k = \mathcal{P}_3$ for all $k \geq 3$; additionally, we set the upper bound $\psi(k,n) = (2ne/\dim(\mathcal{F}_k))^{\dim(\mathcal{F}_k)}$. To evaluate the performance of a prediction rule $f^\dagger_n$, we aim to compute its relative (generalized) expected utility

$$\text{REU}(f^\dagger_n) \equiv \frac{\mathbb{E}[S(f^\dagger_n)]}{S^*},$$

which can be approximated via simulation because

$$\text{REU}(f^\dagger_n) = \mathbb{E} \left[ \frac{S(f^\dagger_n)}{S(p^*)} \right] \approx \frac{1}{S} \sum_{j=1}^{S} \frac{S_{m,j}(f^\dagger_n|\mathcal{D}_{n,j})}{S_{m,j}(p^*)},$$

where $S_{m,j}(f^\dagger_n|\mathcal{D}_{n,j})$ is the $j$-th (out-of-sample) empirical utility with size $m$ of $f^\dagger_n$, constructed by the $j$-th training data $\mathcal{D}_{n,j}$ with size $n$, $S_{m,j}(p^*)$ is the $j$-th (out-of-sample) empirical utility with size $m$ of $p^*$, and $S$ is the number of simulation replications. We set $n \in \{500, 1000, 1500, 2000\}$, $m = 5000$ and $S = 500$ in the simulation experiment.

Table 3.1 shows the relative (generalized) expected utility of ML-logit, MU, and UMPR under different designs with $n = 500$. As expected, a correctly specified ML-logit achieves the largest relative expected utility among these estimators for DGP 1. However, a misspecified ML-logit, compared to MU and UMPR, usually has the worst performance. As $\alpha$ decreases (i.e., the complexity penalty is smaller), UMPR may have larger relative expected utility because it tends to select the cubic MU, which is correctly specified in these designs. However, the selection depends on the underlying data generating process and preference. In the last design, UMPR and linear MU have the same relative expected utility, which is a caveat that the correctly specified cubic MU is never selected out of 500 simulation replications. This phenomenon arises probably because the distribution-free complexity penalty used to construct the UMPR is too conservative.\textsuperscript{15}

---

\textsuperscript{14} By the cubic MU, we mean that the MU optimization is taken over the class $\mathcal{P}_3$ of polynomial transformations of order at most 3. Similarly, we refer to the cubic ML-logit as the maximum likelihood estimation with optimization taken over the class $\Lambda(\mathcal{P}_3) \equiv \{x \mapsto \Lambda(f(x)) : f \in \mathcal{P}_3\}$.

\textsuperscript{15} It is possible to use data-dependent penalties. See for example Bartlett et al. (2002).
Table 3.1: Relative Expected Utility of ML-logit, MU, and UMPR

<table>
<thead>
<tr>
<th>DGP 1</th>
<th>( p(x) = \Lambda(-0.5x + 0.2x^3) )</th>
<th>( p(x) = 20 ) and ( c(x) = 0.5 )</th>
<th>( p(x) = 20 ) and ( c(x) = 0.5 + 0.025x )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pref.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( k = 1 )</td>
<td>( k = 2 )</td>
<td>( k = 3 )</td>
<td>( k = 1 )</td>
</tr>
<tr>
<td>ML-logit</td>
<td>0.3469</td>
<td>0.3172</td>
<td>0.9393</td>
</tr>
<tr>
<td>MU</td>
<td>0.5093</td>
<td>0.5417</td>
<td>0.6650</td>
</tr>
<tr>
<td>UMPR</td>
<td>0.5786</td>
<td>0.5890</td>
<td>0.5921</td>
</tr>
</tbody>
</table>

| DGP 2 | \( p(x_1, x_2) = \Lambda(Q(1.5x_1 + 1.5x_2)) \) where \( Q(u) = \frac{1.5 - 0.1u}{\exp(0.25u + 0.1u^2 - 0.04u^3)} \) | \( b(x_1, x_2) = 20 \) and \( c(x_1, x_2) = 0.75 \) | \( b(x_1, x_2) = 20 + 40 \cdot \mathbb{I}_{||x_1 + x_2|| < 1.5} \) and \( c(x_1, x_2) = 0.75 \) |
|---|---|---|---|
| Pref. | | | |
| \( k = 1 \) | \( k = 2 \) | \( k = 3 \) | \( k = 1 \) | \( k = 2 \) | \( k = 3 \) |
| ML-logit | 0.6026 | 0.5941 | 0.6009 | 0.3086 | 0.2919 | 0.3460 |
| MU | 0.6705 | 0.5038 | 0.6811 | 0.4880 | 0.3083 | 0.5013 |
| UMPR | 0.6725 | 0.6725 | 0.6725 | 0.6725 | 0.4880 | 0.4880 |

Note: We use \texttt{glmfit} and \texttt{simulannealbnd} algorithms in MATLAB\textsuperscript{\textregistered} to compute ML-logit and MU, respectively. All tuning parameters in both algorithms are set by default.

Since the UMPR is constructed to alleviate the in-sample overfitting, we compare its performance to that of the tenfold cross-validation cubic Lasso-logit (i.e., cubic ML-logit with an \( \ell_1 \) penalty). Table 3.2 shows that in terms of relative (generalized) expected utility, the Lasso-logit outweighs the UMPR in DGP 1 but is dominated by the UMPR in DGP 2. This implies that the Lasso-logit may have better performance if the hierarchy of classes \( \{\Lambda(F_k)\}_{k=1}^\infty \) includes the correct model specification for ML-logit. More importantly, the relative expected utility of the UMPR increases in \( n \), the number of in-sample observations, for both DGPs. This phenomenon is guaranteed by Corollary 3.1 because the approximation error \( S^* - \sup_{f \in F_3} S(f) \) is equal to zero. In contrast, for DGP 2 in which any ML-logit is misspecified, the Lasso-logit even has smaller relative expected utility as \( n \) increases.

### 3.5 Conclusion

The maximum utility estimation can be viewed as the binary classification with a decision-based loss function. Despite its possible improvement in utility over traditional maximum likelihood methods, maximum utility estimation has inherited the in-sample overfitting from the perceptron learning.

To alleviate the in-sample overfitting, we adopt the structural risk minimization approach to construct a utility-maximizing prediction rule. We show that the difference
Table 3.2: Relative Expected Utility of Lasso-logit and UMPR

<table>
<thead>
<tr>
<th>DGP 1</th>
<th>( p(x) = \Lambda(-0.5x + 0.2x^3) )</th>
<th>( p(x) = 0.5 + 0.025x )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pref.</td>
<td>( b(x) = 20 ) and ( c(x) = 0.5 )</td>
<td>( b(x) = 20 ) and ( c(x) = 0.5 + 0.025x )</td>
</tr>
<tr>
<td></td>
<td>( n = 500 )</td>
<td>( n = 500 )</td>
</tr>
<tr>
<td></td>
<td>( n = 1000 )</td>
<td>( n = 1000 )</td>
</tr>
<tr>
<td></td>
<td>( n = 1500 )</td>
<td>( n = 1500 )</td>
</tr>
<tr>
<td></td>
<td>( n = 2000 )</td>
<td>( n = 2000 )</td>
</tr>
<tr>
<td>Lasso-logit</td>
<td>0.6052</td>
<td>0.6496</td>
</tr>
<tr>
<td>UMPR</td>
<td>0.5921</td>
<td>0.5428</td>
</tr>
</tbody>
</table>

| DGP 2 | \( p(x_1, x_2) = \Lambda(Q(1.5x_1 + 1.5x_2)) \) where \( Q(u) = \frac{1.5 - 0.14}{\exp(0.25u + 0.1u^2 - 0.04u^3)} \) | \( b(x_1, x_2) = 20 + 40 \cdot \mathbb{I}_{|x_1 + x_2| < 1.5} \) and \( c(x_1, x_2) = 0.75 \) |
|-------|---------------------------------|------------------|
| Pref. | \( b(x_1, x_2) = 20 \) and \( c(x_1, x_2) = 0.75 \) | \( b(x_1, x_2) = 20 + 40 \cdot \mathbb{I}_{|x_1 + x_2| < 1.5} \) and \( c(x_1, x_2) = 0.75 \) |
|       | \( n = 500 \) | \( n = 500 \) |
|       | \( n = 1000 \) | \( n = 1000 \) |
|       | \( n = 1500 \) | \( n = 1500 \) |
|       | \( n = 2000 \) | \( n = 2000 \) |
| Lasso-logit | 0.5950 | 0.3307 |
| UMPR | 0.6737 | 0.2418 |

Note: (i) The tuning parameter \( \alpha \) of the complexity penalty for UMPR is 0.1. (ii) We use \texttt{lassoglm} and \texttt{simulannealbnd} algorithms in MATLAB\textsuperscript{R} to compute Lasso-Logit and UMPR, respectively. All tuning parameters in both algorithms are set by default.

between the maximal expected utility and the generalized expected utility based on this utility-maximizing prediction rule is bounded. In addition, this upper bound is close to zero for a large sample if the approximation error is equal to zero for some pre-specified classes of functions. Simulation results show that this utility-maximizing prediction rule, compared to the Lasso-logit, may have larger relative expected utility if the conditional probability of the binary outcome is misspecified. The utility-maximizing prediction rule is thus important for the decision-making based on the binary prediction.

Appendix 3.A Technical proofs

3.A.1 Proof of Proposition 3.1

Proof. By Assumptions A2 and A3, \( 0 \leq \ell(y, x, f) \leq 4M \) for all \( (y, x, f) \in \{-1, 1\} \times \mathcal{X} \times \mathcal{F} \). For any \( \beta \in (0, 4M) \) and \( f \in \mathcal{F} \), let \( g(y, x; f, \beta) = \mathbb{I}_{[\ell(y, x, f) \geq \beta]} \). It follows by Equation (3.10) of Vapnik (1995) that for any \( n, k \in \mathbb{N} \) and \( \varepsilon > 0 \),

\[
\mathbb{P} \left( \sup_{f \in \mathcal{F}_k} |L(f) - L_n(f)| > \varepsilon \right) \leq 4\bar{\Pi}_k(2n) \exp \left\{ -\frac{n\varepsilon^2}{16M^2} \right\},
\]

where \( \bar{\Pi}_k(\cdot) \) is the growth function of

\( \Psi_k \equiv \{(x, y) \mapsto g(y, x; f, \beta) : f \in \mathcal{F}_k, \beta \in (0, 4M)\} \).

Since \( \Psi_k \) can be rewritten as

\[
\{(x, y) \mapsto \mathbb{I}_{[b(x)(y(1-2c(x))+1) \geq \beta]}\mathbb{I}_{[\text{sign}(f(x)-c(x)) \neq y]} : f \in \mathcal{F}_k, \beta \in (0, 4M)\},
\]

112
we have $\Pi_k(2n) \leq \Pi_{k,1}^\dagger(2n)\Pi_{k,2}^\dagger(2n)$, where $\Pi_{k,1}^\dagger(\cdot)$ and $\Pi_{k,2}^\dagger(\cdot)$ are the growth function of $\{(x, y) \mapsto 1_{b(x)(1-2c(x))+1} : \beta \in (0, 4M)\}$ and $\{(x, y) \mapsto 1_{\text{sign}(f(x)-c(x))\neq y} : f \in \mathcal{F}_k\}$, respectively. The result follows by recognizing that $\Pi_{k,1}^\dagger(2n) \leq (2n+1)$ and $\Pi_{k,2}^\dagger(2n) = \Pi_{k,c}\kappa(2n)$.

3.A.2 Proof of Theorem 3.1

Proof. 1. By construction, we have

$$\tilde{L}_n(\tilde{f}_j; j) = R_{n,j} + 4M\sqrt{\frac{(1+\alpha) \log \{j+1\}}{n}}$$

for each $j \in \mathbb{N}$. So, for any $n \in \mathbb{N}$ and $\alpha, \varepsilon > 0$,

$$\mathbb{P}\left( L(\tilde{f}_n) - \tilde{L}_n(\tilde{f}_n) > \varepsilon \right)$$

$$\leq \mathbb{P}\left( \sup_j L(\hat{f}_j) - \tilde{L}_n(\hat{f}_j; j) > \varepsilon \right)$$

$$\leq \sum_{j=1}^\infty \mathbb{P}\left( L(\hat{f}_j) - \tilde{L}_n(\hat{f}_j; j) > \varepsilon \right)$$

$$\leq \sum_{j=1}^\infty \mathbb{P}\left( L(\hat{f}_j) - R_{n,j} > \varepsilon + 4M\sqrt{\frac{(1+\alpha) \log \{j+1\}}{n}} \right)$$

$$\leq \sum_{j=1}^\infty \exp \left\{ -\frac{n}{(4M)^2} \left[ \varepsilon + 4M\sqrt{\frac{(1+\alpha) \log \{j+1\}}{n}} \right]^2 \right\}$$

$$\leq \sum_{j=1}^\infty \exp \left\{ -\frac{n}{(4M)^2} \left[ \varepsilon^2 + (4M)^2 \frac{(1+\alpha) \log \{j+1\}}{n} \right] \right\}$$

$$\leq \exp \left\{ -\frac{n\varepsilon^2}{16M^2} \right\} \sum_{j=1}^\infty \exp \left\{ -(1+\alpha) \log \{j+1\} \right\}$$

$$\leq \left( \frac{2+\alpha}{2\alpha} \right) \exp \left\{ -\frac{n\varepsilon^2}{16M^2} \right\}.$$

The last inequality holds because

$$\sum_{j=1}^n (j+1)^{-(1+\alpha)} \leq \frac{1}{2^{1+\alpha}} + \int_1^{\infty} (x+1)^{-(1+\alpha)} \, dx$$

$$= \frac{1}{2^{1+\alpha}} + \frac{1}{\alpha 2^\alpha}$$

$$= \frac{1}{2^\alpha} \left( \frac{1}{2} + \frac{1}{\alpha} \right) \leq \frac{1}{2} + \frac{1}{\alpha} = \frac{2+\alpha}{2\alpha}.$$
2. For ease of notation, set \( L_k^* \equiv \inf_{f \in F_k} L(f) \) for each \( k \in \mathbb{N} \). By part 1 and Lemma 3.1 below, we have

\[
\mathbb{E} \left[ L(\tilde{f}_n) - \tilde{L}_n(\tilde{f}_n) \right] \leq 4M \sqrt{\frac{1 + \log \{ (2 + \alpha)/2\alpha \}}{n}}.
\]

In addition, for each \( k \in \mathbb{N} \),

\[
\mathbb{E}[\tilde{L}_n(\tilde{f}_n)] - L_k^* \leq \mathbb{E}[\tilde{L}_n(\tilde{f}_k; k)] - L_k^* \\
= \mathbb{E}[L_n(\hat{f}_k)] - L_k^* + C_n(k; \alpha, \psi) \\
\leq C_n(k; \alpha, \psi).
\]

So, for each \( k \in \mathbb{N} \),

\[
\mathbb{E}[L(\tilde{f}_n)] - L^* \\
= \mathbb{E} \left[ L(\tilde{f}_n) \right] - \mathbb{E} \left[ \tilde{L}_n(\tilde{f}_n) \right] + \mathbb{E} \left[ \tilde{L}_n(\tilde{f}_n) \right] - L_k^* + L_k^* - L^* \\
\leq 4M \sqrt{\frac{1 + \log \{ (2 + \alpha)/2\alpha \}}{n}} + C_n(k; \alpha, \psi) + L_k^* - L^*.
\]

It follows that

\[
\mathbb{E} \left[ L(\tilde{f}_n) \right] - L^* \\
\leq 4M \sqrt{\frac{1 + \log \{ (2 + \alpha)/2\alpha \}}{n}} + \min_k \{ C_n(k; \alpha, \psi) + L_k^* - L^* \}.
\]

3.A.3 Proof of Corollary 3.1

Proof. 1. By definition, we have

\[
\tilde{S}_n(\tilde{f}_n) - S(\tilde{f}_n) = [G_n - G] + 2[L(\tilde{f}_n) - \tilde{L}_n(\tilde{f}_n)].
\]

It follows that for any \( \eta \in (0, 1) \),

\[
\mathbb{P} \left( \tilde{S}_n(\tilde{f}_n) - S(\tilde{f}_n) > \varepsilon \right) \\
\leq \mathbb{P} (G_n - G > (1 - \eta)\varepsilon) + \mathbb{P} \left( L(\tilde{f}_n) - \tilde{L}_n(\tilde{f}_n) > \eta\varepsilon/2 \right).
\]

The first term is bounded above by \( \exp \left\{ -\frac{n(1-\eta)^2\varepsilon^2}{8M^2} \right\} \) by Hoeffding’s inequality. Applying Theorem 1 yields the upper bound of the second term.
2. Note that
\[ E[S(\hat{f}_n)] - S^* \]
\[ = E[G - \triangle L(\hat{f}_n)] - [G - 2L^*] \]
\[ = -2\{E[L(\hat{f}_n)] - L^*\} \]
\[ \geq -2 \min_k \left\{ C_n(k; \alpha, \psi) + \left( \inf_{f \in \mathcal{F}_k} L(f) - L^* \right) \right\} - 8M \sqrt{1 + \log \left\{ \frac{(2 + \alpha)}{2\alpha} \right\} \frac{1}{n}} \]
by part 2 of Theorem 1. After rearrangement, we have
\[ S^* - E[S(\hat{f}_n)] \]
\[ \leq \min_k \left\{ 2C_n(k; \alpha, \psi) + \left( \inf_{f \in \mathcal{F}_k} 2L(f) - 2L^* \right) \right\} + 8M \sqrt{1 + \log \left\{ \frac{(2 + \alpha)}{2\alpha} \right\} \frac{1}{n}} \]
\[ = \min_k \left\{ 2C_n(k; \alpha, \psi) + \left( G - 2L^* - \sup_{f \in \mathcal{F}_k} (G - 2L(f)) \right) \right\} \]
\[ + 8M \sqrt{1 + \log \left\{ \frac{(2 + \alpha)}{2\alpha} \right\} \frac{1}{n}} \]
\[ = \min_k \left\{ 2C_n(k; \alpha, \psi) + \left( S^* - \sup_{f \in \mathcal{F}_k} S(f) \right) \right\} + 8M \sqrt{1 + \log \left\{ \frac{(2 + \alpha)}{2\alpha} \right\} \frac{1}{n}}. \]
Proof. For any \( u > 0 \), we have

\[
E[Z^2 1_{Z \geq 0}] = \int_0^\infty P(Z^2 1_{Z \geq 0} > t) \, dt
\]

\[
= \int_0^u P(Z^2 1_{Z \geq 0} > t) \, dt + \int_u^\infty P(Z^2 1_{Z \geq 0} > t) \, dt
\]

\[
\leq u + \int_u^\infty P(Z 1_{Z \geq 0} > \sqrt{t}) \, dt
\]

\[
= u + \int_u^\infty P(Z 1_{Z \geq 0} > \sqrt{t}, Z \geq 0) \, dt + \int_u^\infty P(Z 1_{Z \geq 0} > \sqrt{t}, Z < 0) \, dt
\]

\[
= u + \int_u^\infty P(Z > \sqrt{t}) \, dt
\]

\[
\leq u + c \int_u^\infty \exp\{-2nt\} \, dt
\]

\[
= u + \frac{c}{2n} \exp\{-2nu\}.
\]

Taking \( u = \frac{\log c}{2n} \) yields

\[
E[Z] = E[Z 1_{Z \geq 0}] + E[Z 1_{Z < 0}] \leq E[Z 1_{Z \geq 0}] \leq \sqrt{E[Z^2 1_{Z \geq 0}]} \leq \sqrt{\frac{\log ce}{2n}}.
\]
Bibliography


