

Semiparametric Models for Dynamic Treatment Effects and Mediation Analyses with Observational Data

Sukjin Han

Sungwon Lee

School of Economics

Department of Economics

University of Bristol

Sogang University

sukjin.han@gmail.com

sungwonlee@sogang.ac.kr

August 26, 2023

Abstract

This paper proposes a semiparametric model that captures how a sequence of interventions interacts with a sequence of outcomes. In this setup, the outcome at the given period is affected by the history of treatments and outcomes, directly or indirectly through mediators. The main challenge in understanding various channels of dynamic effects is that, in observational settings, individuals make dynamically endogenous decisions whether to select into treatments. Using the approach of instrumental variables, this paper shows how the average and quantile dynamic treatment effects and mediation effects can be point identified and efficiently estimated in a class of semiparametric models under treatment endogeneity and flexible heterogeneity. Our procedure only requires binary instruments. As a byproduct of our semiparametric specification, we also identify and estimate parameters that reflect the degree of endogenous selection and time-invariant heterogeneity.

JEL Numbers: C14, C31, C32, C36

Keywords: dynamic treatment effects, dynamic mediation analysis, treatment endogeneity, instrumental variables, copula.

1 Introduction

This paper proposes a semiparametric model that captures how a sequence of interventions interacts with a sequence of outcomes over periods. Understanding the dynamic causal mechanism of treatments influencing outcomes is important in designing more informed policies. For example, a multi-year after-school program has influences on the time path of student performance. A performance at a given year is influenced by previous participation decisions and performances through multiple channels: the performance is affected (i) by the current intervention, (ii) directly by previous interventions, (iii) indirectly by previous interventions through previous performances, which in turn are associated with the current performance via (iv) state dependence and (v) time-invariant heterogeneity. Without understanding which channels are important in improving the performance, designing effective after-school programs will not be successful. Other examples of dynamic treatments and outcomes can be found in education (e.g., a household intervention program for disadvantaged children), health (e.g., a sequence of medical treatments), development (e.g., multi-stage field experiments), and online platforms (e.g., A/B testings).

The main challenge in understanding various channels of dynamic effects is that, in observational settings, individuals make dynamically endogenous decisions whether to select into treatments. Students decide to participate in the program based on their previous decisions and performance as well as their prospect of future decisions and performance. Even in experimental settings, possibly due to learning over time, individuals participating in a multi-stage experiment are likely to deviate from the random assignments. To address this challenge while remaining flexible in modeling dynamics and treatment heterogeneity, we use the approach of instrumental variables (IVs). We assume IVs are generated from a sequence of exogenous shocks or sequential experiments. We consider the most challenging

setting that IVs have minimal variation (i.e., binary variation). This setting incorporates wide range of interesting examples (e.g., multi-period/stage experiments, sequential fuzzy regression continuity). The results of this paper will immediately apply with IVs of richer variation.

This paper shows how the average and quantile dynamic treatment and mediation effects (exemplified in (i)–(v) above) can be point identified and efficiently estimated in a class of semiparametric models. We consider a sequence of outcomes that are either discrete or continuous and a sequence of binary treatments. Naturally, a sequence of nonparametric threshold-crossing models arises in the model construction. We remain fully nonparametric in the structure of outcome and treatment-selection equations. We introduce a semiparametric structure for the joint distribution of the unobservables that determine outcomes and selection decisions of all time periods. Specifically, we introduce a parametric copula to model the dependence among the unobservables while letting the marginal distributions fully nonparametric. The motivation for the semiparametric specification of the joint distribution is twofold. First, the identification in nonparametric models for dynamic treatment effects is deemed challenging in the literature. The existing results either consider irreversible treatments, rely on IVs with large support or extra exogenous variables, or resort to partial identification; see below for references. We show how the semiparametric approach is sufficiently flexible while lends us a tractable point identification strategy with minimal exogenous variation. Second, a fully nonparametric joint distribution may cause the curse of dimensionality in the current multi-period setting. We show, on the other hand, how the semiparametric approach achieves efficiency in estimation and leads to a simple estimation procedure. As a byproduct of our specification, we identify the parametrized dependence structure of the joint distribution, in addition to dynamic treatment and mediation effects. These dependence parameters capture the degree of endogenous selection and serial correlation (that reflects time-invariant heterogeneity), which are by themselves important policy-relevant parameters. We make sure that the marginal distributions of unobservables are fully nonparametric, which is crucial to avoid misspecification because the effects we want to identify are direct functions

of these marginals. For the reasons described, we believe that the semiparametric compromise may have great appeal to practitioners, for whom practically useful and easy-to-implement methods have been scarcely available to estimate dynamic treatment and mediation effects under endogeneity and flexible heterogeneity.

The main idea for identification is to model the joint distribution of unobservables using a multi-variate copula that are generated from vine copulas. We assume that each dependence parameter (between outcome and treatment unobservables or between different periods) captures certain *pairwise* stochastic ordering. The idea of using copula for identification and estimation builds on [Han and Vytlacil \(2017\)](#) and [Han and Lee \(2019\)](#). However, the current work differs from its predecessors in several important ways. First, this paper considers multi-period models, which produce a wide range of interesting parameters that have not been considered in the previous studies on static models. Second, [Han and Vytlacil \(2017\)](#) and [Han and Lee \(2019\)](#) only consider a binary outcome while the current work considers (a sequence of) binary or continuous outcomes. We show how the identification with continuous outcomes remains tractable without imposing additional restrictions. Third, we allow that each counterfactual outcome is generated by a distinct unobservable depending on the treatment status. This effectively assumes that each observed outcome is generated by a *vector* of unobservables, which is crucial in allowing for flexible treatment heterogeneity. On the other hand, the previous papers implicitly assume a scalar unobservable, or equivalently, rank invariance ([Chernozhukov and Hansen \(2005\)](#)), which significantly limits heterogeneity. Finally, it is not a priori obvious that the useful property of bivariate copula would continue to hold with multi-variate copulas. We show that this is in fact the case but only within a class of multi-variate copulas that is newly proposed in this paper. We show that this class includes the multi-variate Gaussian copula, which implies that identification is achieved in the dynamic and multi-variate extension of the popular bivariate probit model.

Under the copula specification and with a sequence of binary IVs, we identify the dynamic treatment effects, treatment effects mediated by previous outcomes, nonparametric state-dependence, treatment-status-specific endogenous selection parameters, and serial correlation

parameters, both the average and quantile effects and all conditional (or unconditional) on covariates. When outcomes are continuous, we identify the distribution of counterfactual outcomes and thus quantile treatment and mediation effects and quantile state-dependence. With binary outcomes, the effects of lagged outcomes as mediators can be simply defined and identified as if treatment effects. With continuous outcomes, we propose a framework to maintain this nice aspect. In addition, we identify the responses of treatments to previous treatments and outcomes, which capture habit and learning. Given the rich set of identified parameters, we show how they can be combined to answer further policy questions, such as dynamic complementarity.

We propose to use a sieve maximum likelihood (ML) to estimate the parameters. The sieve methods provide a flexible and tractable way to estimate semi-/non-parametric models (Chen (2007)). We develop the asymptotic theory for the sieve ML estimators of the dynamic effects and dependence parameters, including consistency, convergence rates, and \sqrt{n} -asymptotic normality. We also establish the asymptotic theory for the sieve likelihood ratio test statistic, which helps perform inference on the parameters without estimating asymptotic variances. Interestingly, our model becomes saturated with discrete covariates, in which case one can use the standard parametric ML estimation.

This paper mainly contributes to the literature on treatment effects and policy evaluation (e.g., Abbring and Heckman (2007)). Heckman and Navarro (2007) and Heckman et al. (2016) consider identification of dynamic effects of treatment timing (i.e., irreversible treatments). They allow the joint distribution to be unknown while requiring IVs to have large support. For the identification of the joint distribution of counterfactual outcomes, they introduce a factor structure for the unobservables. Han (2021) considers a fully nonparametric model for dynamic treatment effects with a general behavior of treatment sequence which nests treatment timing. Due to the flexibility and allowing for binary IVs, he relies on additional exogenous variables with specific support restrictions. Han (2022) embraces partial identification and characterizes bounds on dynamic treatment effect and the identified set for

optimal dynamic treatment allocation rules.¹ Another related line of work concerns multiple or multi-valued treatments that are ordered or unordered where Imbens and Angrist (1994)-type monotonicity assumptions fail to hold (Heckman and Pinto (2018); Lee and Salanié (2018); Balat and Han (forthcoming)).

This paper also relates to the literature on dynamic discrete choice models, although the approach is very different. Models in this literature typically include lagged dependent variables, whose effects are interpreted as state-dependence, and time-invariant unobserved individual heterogeneity.² For example, Honoré and Kyriazidou (2000) study identification and estimation of the parameters in dynamic discrete choice models focusing on state-dependence. Relatedly, Kyriazidou (2001) considers a dynamic sample selection model with lagged dependent variables. This model may be generalized to a dynamic switching regression model for treatment effects. Our approach complements this literature in several ways. First, the main focus of this literature is to identify and estimate state-dependence parameters, whereas our main purpose of using dynamic two-stage model is to identify the dynamic treatment effects in addition to state-dependence as part of mediation effects. Second, we consider nonparametric specifications for the structural functions instead of linear specifications. We follow the approach of the treatment effect literature and write all the effects as nonparametric marginal effects instead of coefficients in a linear specification that are sometimes less interpretable (i.e., in a discrete choice model with linear index). Individual heterogeneity is subsumed in our marginal effects while explicitly specifying it is crucial for identification in the literature of dynamic discrete choice model. Finally, with the semiparametric structure, we achieve \sqrt{n} -asymptotic normality for many interesting functionals of the structural functions, whereas the estimator of Honoré and Kyriazidou (2000) converges at a slower rate than \sqrt{n} when the time-varying covariate vector contains a continuous random variable³ and the

¹Murphy et al. (2001) and subsequent work in the biostatistics literature consider the problem of optimal dynamic allocation, but mostly under sequential unconfoundedness assumptions with a few exceptions (Cui and Tchetgen Tchetgen (2021); Qiu et al. (2021)). This line of work can be adapted to identify and estimate dynamic treatment effects.

²State-dependence and individual heterogeneity as the sources of observed serial dependence have different policy implications (Heckman (1981); Arellano and Honoré (2001); Abbring and Heckman (2007)).

³Honoré and Weidner (2021) recently propose a different identification strategy based on Bonhomme

estimator of Kyriazidou (2001) at a slower rate than \sqrt{n} due to kernel estimation. As a price for these gains, our approach requires a sequence of excluded variables, and thus is more suitable for a relatively short time horizon. Given that many studies on dynamic discrete choice models require a sufficient number of periods for identification, the two approaches are complementary in this way as well.⁴ Another strand of the literature on dynamic discrete choice models adopts the partial identification approach (e.g., Honoré and Tamer (2006); Torgovitsky (2019)), but our focus is point identification in a more parsimonious model.

The rest of the paper is organized as follows. Section 2 introduces the model and identifying assumptions for the leading case of $T = 2$ and binary outcomes. Section 3 defines the parameters of interest, and Section 4 shows the identifiability of the parameters. Section 5 discusses identification with general T . Section 6 extends the previous analysis by considering continuous outcomes. Section 7 considers semiparametric estimation and develops the asymptotic theory. Most proofs are contained in Section A.

2 Model and Identifying Assumptions

As a leading case, we consider a two-period model for dynamic treatment effects with binary outcomes. Even with this simple model, we can capture many interesting dynamic effects that are not available in a static model. In Section 5, we consider a general T -period model.

The extension to continuous outcomes is considered in Section 6. Let $D \equiv (D_1, D_2)$ and (2012) for the same model as Honoré and Kyriazidou (2000), and their estimator is shown to be \sqrt{n} -asymptotically normal.

⁴Honoré and Kyriazidou (2000) and Honoré and Weidner (2021) require more than four periods for identification in a dynamic discrete choice model.

$d \equiv (d_1, d_2)$. We posit the following model:

$$Y_2 = 1[\mu_2(Y_1, D, X) \geq U_2(Y_1, D)], \quad (2.1)$$

$$D_2 = 1[\pi_2(Y_1, D_1, Z_2, X) \geq V_2], \quad (2.2)$$

$$Y_1 = 1[\mu_1(D_1, X) \geq U_1(D_1)], \quad (2.3)$$

$$D_1 = 1[\pi_1(Z_1, X) \geq V_1]. \quad (2.4)$$

In this model, $U_1(d_1)$ and $U_2(y_1, d)$ are introduced to allow for rich heterogeneity in treatment and mediation effects. To see this, consider counterfactual outcomes $Y_1(d, z_1)$ and $Y_2(y_1, d, z)$ and counterfactual treatments $D_1(d_2, z_1)$ and $D_2(y_1, d_1, z)$ and assume the following:

Assumption 2.1. (i) (No Anticipation) $Y_1(d, z_1) = Y_1(d_1, z_1)$ and $D_1(d_2, z_1) = D_1(z_1)$; (ii) (Exclusion) $Y_1(d_1, z_1) = Y_1(d_1)$ and $Y_2(y_1, d, z) = Y_2(y_1, d)$.

Under this assumption, $Y_1(d_1)$ and $Y_2(y_1, d)$ are the counterfactual outcomes that define treatment and mediation effects; see the next section for details of the parameters of interest. The observed outcomes relate to the counterfactual outcomes via $Y_1 = D_1 Y_1(1) + (1 - D_1) Y_1(0)$ and $Y_2 = \sum_{y_1, d \in \{1, 0\}^3} 1[Y_1 = y_1, D = d] Y_2(y_1, d)$. Since $Y_2(y_1, d)$ is a function of $U_2(y_1, d)$, the observed Y_2 is effectively a function of the entire vector of $(U_2(1, 1, 1), U_2(1, 1, 0), U_2(1, 0, 1), U_2(1, 0, 0), U_2(0, 1, 1), U_2(0, 1, 0), U_2(0, 0, 1), U_2(0, 0, 0))$. Similarly, Y_1 is a function of $(U_1(1), U_1(0))$. Therefore the equations for outcomes contain *vector unobservables*. This aspect is in contrast to models that assume a scalar unobservable (e.g., $Y = 1[\mu(D, X) \geq U]$) as in [Vytlacil and Yildiz \(2007\)](#) and [Shaikh and Vytlacil \(2011\)](#) or models that assume rank invariance ([Chernozhukov and Hansen \(2005\)](#)). The following example illustrates the role of vector unobservables.

Example 1 (Dynamic Roy Models). *Our model nests dynamic Roy models as a special case. In this scenario, the agent may select into treatment when he benefit $(Y_2(1) - Y_2(0))$ exceeds the cost $(-\delta_2(Z_2))$, namely based on $D_2 = 1\{\delta_2(Z_2) + Y_2(1) - Y_2(0) \geq 0\}$, where*

$Y_2(d_2) \equiv Y_2(Y_1, D_1, d_2)$ and

$$\begin{aligned} \delta_2(Z_2) + Y_2(1) - Y_2(0) &\equiv \delta_2(Z_2) + \mu_2(Y_1, D_1, 1) - \mu_2(Y_1, D_1, 0) \\ &\quad - (U_2(Y_1, D_1, 1) - U_2(Y_1, D_1, 0)) \\ &\equiv \pi_2(Y_1, D_1, Z_2) - V_2. \end{aligned}$$

Therefore, if $U_2(y_1, d) = U_2$, the selection model ends up not having any unobservable. Similarly, $D_1 = 1\{\delta_1(Z_1) + Y_1(1) - Y_1(0) \geq 0\}$ where

$$\begin{aligned} \delta_1(Z_1) + Y_1(1) - Y_1(0) &= \delta_1(Z_1) + \mu_1(1) - \mu_1(0) - (U_1(1) - U_1(0)) \\ &\equiv \pi_1(Z_1) - V_1 \end{aligned}$$

contains no unobservable if $U_1(d_1) = U_1$.

We normalize $(V_1, U_1(d_1), V_2, U_2(y_1, d))|X = x$ to be uniform random variables on $[0, 1]$ ^{4,5}. Under this normalization, we have $\pi_1(Z_1, X) = \Pr[D_1 = 1|Z_1, X]$ and $\pi_2(Y_1, D_1, Z_2, X) = \Pr[D_2 = 1|Y_1, D_1, Z_2, X]$. We make the following assumptions:

Assumption 2.2 (Independence). $(Z_1, Z_2) \perp (V_1, V_2, U_1(d_1), U_2(y_1, d))|X$ for $(y_1, d) \in \{0, 1\}^3$.

Assumption 2.3 (Relevance). π_1 and π_2 are non-trivial functions of Z_1 and Z_2 , respectively, and $(Z_1, Z_2)|X$ are non-degenerate.

Assumption 2.3 assumes that instruments are relevant conditional on X .

Assumption 2.4 (Copula). For each $(y_1, d) \in \{1, 0\}^3$, the unobservables are jointly distributed as

$$(V_1, V_2, U_1(d_1), U_2(y_1, d))|_{X=x} \sim C(v_1, v_2, u_1, u_2; \Sigma(y_1, d, x)),$$

⁵This normalization needs caution in this semiparametric setting. It does *not* necessarily impose exogeneity of X , although it may seem so.

where $C(v_1, v_2, u_1, u_2; \Sigma)$ is a 4-copula with dependence matrix Σ .

In Assumption 2.4, $\Sigma(y_1, d, x)$ captures all the dependences among $(V_1, V_2, U_1(d_1), U_2(y_1, d))$ conditional on $X = x$. Notable elements in $\Sigma(y_1, d, x)$ are $\rho_{V_1, U_1(d_1), x}$ and $\rho_{V_t, U_2(y_1, d), x}$ (for $t = 1, 2$ and $(y_1, d) \in \{0, 1\}^3$), which capture the treatment-state- and covariate- specific selection, which can be economically meaningful. The rank similarity or rank invariance (Chernozhukov and Hansen (2005)) will impose restrictions such as $\rho_{V_1, U_1(1), x} = \rho_{V_1, U_1(0), x} \equiv \rho_{V_1, U_1, x}$, which rules out state-specific selection. Although we can also allow the form of the copula to depend on x and d , we do not pursue this specification for succinctness. In the next assumption, $C(\cdot|\cdot; \rho)$ denotes the conditional copula of $C(\cdot, \cdot; \rho)$ and the stochastic ordering “ \prec_S ” is defined as follows.

Definition 2.1 (Strictly More SI). *Let $F(w_1|w_2)$ and $\tilde{F}(w_1|w_2)$ be conditional distributions. Suppose that $F(w_1|w_2)$ and $\tilde{F}(w_1|w_2)$ are continuous in w_1 for all w_2 . Then \tilde{F} is strictly more stochastically increasing than F if $\psi(w_1, w_2) \equiv \tilde{F}^{-1}(F(w_1|w_2)|w_2)$ is strictly increasing in w_2 , which is denoted as $F(\cdot|\cdot) \prec_S \tilde{F}(\cdot|\cdot)$.*

In this definition, the ordering is defined in terms of the degree of a particular positive dependence between two random variables.

Assumption 2.5 (Vine Structure). *The copula $C(v_1, v_2, u_1, u_2; \Sigma)$ in Assumption 2.4 and its margins satisfy the following conditions:*

- (i) $C(v_1, u_1; \rho_{v_1 u_1})$ satisfies that $C(v_1|u_1; \rho_{v_1 u_1}) \prec_S C(v_1|u_1; \tilde{\rho}_{v_1 u_1})$ for any $\rho_{v_1 u_1} < \tilde{\rho}_{v_1 u_1}$;
- (ii) $C(v_1, v_2, u_1; \rho_{v_1 v_2}, \rho_{v_1 u_1}, \rho_{v_2 u_1})$ and $C(v_1, v_2, u_1, u_2; \Sigma)$ are represented by

$$C(v_1, v_2, u_1; \rho_{v_1 v_2}, \rho_{v_1 u_1}, \rho_{v_2 u_1}) = \int^{v_1} C(C(v_2|\tilde{v}_1), C(u_1|\tilde{v}_1); \rho(\rho_{v_1 v_2}, \rho_{v_1 u_1}, \rho_{v_2 u_1})) d\tilde{v}_1,$$

$$C(v_1, v_2, u_1, u_2; \Sigma) = \int^{v_1, v_2} C(C(u_1|\tilde{v}_1, \tilde{v}_2), C(u_2|\tilde{v}_1, \tilde{v}_2); \rho(\Sigma)) dC(\tilde{v}_1, \tilde{v}_2),$$

where the outer copula $C(\cdot, \cdot; \rho)$ on the r.h.s. satisfies $C(\cdot|\cdot; \rho) \prec_S C(\cdot|\cdot; \tilde{\rho})$ for $\rho < \tilde{\rho}$;

- (iii) $\rho(\rho_{v_1 v_2}, \rho_{v_1 u_1}, \rho_{v_2 u_1})$ and $\rho(\Sigma)$ are strictly increasing in $\rho_{v_2 u_1}$ and $\rho_{u_1 u_2}$, respectively.

This assumption is crucial later for the global identifiability of the treatment parameters in our semiparametric model. The parameters are defined in the next section. Assumption 2.5(i) naturally holds for many well-known bivariate copulas; e.g., see Han and Vytlačil (2017). The vine copula structure in Assumption 2.5(ii) builds a multivariate copula from a mixture of conditional copulas. It is a simple and effective way to impose semiparametric structure for joint distributions in multi-period multi-stage models of this paper. By characterizing Assumption 2.5(iii) as the additional requirement, we show how the ordering property of a bivariate copula does not automatically extend to a multi-variate setup. We show that Gaussian copulas satisfies Assumption 2.5.

Example 2 (Gaussian Copulas). *First, Assumption 2.5(i) holds with Gaussian copula (Han and Vytlačil (2017)). Let $(U_1, U_2, U_3) \sim C(\cdot, \cdot, \cdot; \rho_{12}, \rho_{23}, \rho_{13})$, where C is a trivariate Gaussian copula. Define $Z_j \equiv \Phi^{-1}(U_j)$ for $j \in \{1, 2, 3\}$. Then, $Z_j \sim N(0, 1)$. Observe that from Example 4.4 in Joe (1997, p.113), we have*

$$\begin{aligned} C(u_1, u_2, u_3; \rho_{12}, \rho_{23}, \rho_{13}) &= \Phi(z_1, z_2, z_3; \rho_{12}, \rho_{23}, \rho_{13}) \\ &= \int_0^{u_1} C(C_{2|1}(u_2|u), C_{3|1}(u_3|u); \rho_{23;1}) du, \end{aligned}$$

where all the copulas on the r.h.s. are Gaussian and $\rho_{23;1} = \frac{\rho_{23} - \rho_{12}\rho_{13}}{\sqrt{(1-\rho_{12}^2)(1-\rho_{13}^2)}}$ is the partial correlation between Z_2 and Z_3 given Z_1 . Then, the outer copula satisfies \prec_S -ordering and thus Assumption 2.5(ii). Also, the trivariate Gaussian copula satisfies Assumption 2.5(iii) with $\rho(\rho_{12}, \rho_{13}, \rho_{23}) = \rho_{23;1}$. Similarly, we can construct a 4-variate Gaussian copula: From Joe (2014, p.120), we have that

$$C(u_1, u_2, u_3, u_4; \Sigma) = \int_0^{u_1} \int_0^{u_2} C(C_{3|12}(u_3|u, v), C_{4|12}(u_4|u, v); \rho_{34;12}) dC(u, v),$$

where all the copulas on the r.h.s. are Gaussian and $\rho_{34;12} = \frac{\rho_{34;1} - \rho_{23;1}\rho_{24;1}}{\sqrt{(1-\rho_{23;1}^2)(1-\rho_{24;1}^2)}}$, which then satisfies Assumption 2.5(ii)-(iii) with $\rho(\Sigma) = \rho_{34;12}$.

3 Dynamic Treatment and Mediation Effects

In this section, we define the dynamic treatment and mediation effects and show that they can be expressed as known functions of the model primitives: $\mu_2(y_1, d, x)$, $\mu_1(d_1, x)$, $\pi_2(y_1, d_1, z_2)$ in (2.1)–(2.4) and the dependence parameters $\Sigma(y_1, d, x)$ in Assumption 2.4. Our goal in the subsequent section is to identify the model primitives.

Let $Y_2(y_1, d)$, $Y_2(d)$, $Y_2(d_2)$, $Y_2(d_1)$ and $Y_1(d_1)$ be the potential outcomes. Note that $Y_2(Y_1, D_1, d_2) = Y_2(d_2)$ by implicitly assuming no anticipation, but $Y_2(Y_1(d_1), d) = Y_2(d)$ and $Y_2(d_1, D_2(d_1)) = Y_2(d_1)$ where $D_2(d_1)$ is the counterfactual treatment given d_1 .⁶ First, we define the basic causal objects that serve as building blocks to construct treatment and mediation parameters and show they are expressed in terms of the primitives. We focus on the case of binary Y_t here; the expressions with continuous Y_t are shown in Section 6. Consider

$$\begin{aligned} E[Y_2(y_1, d)|X = x] &= \Pr[U_2(y_1, d) \leq \mu_2(y_1, d, x)|X = x] = \mu_2(y_1, d, x), \\ E[Y_1(d_1)|X = x] &= \Pr[U_1(d_1) \leq \mu_1(d_1, x)|X = x] = \mu_1(d_1, x), \end{aligned}$$

because $U_2(y_1, d)$ and $U_1(d_1)$ are uniform conditional on $X = x$. Also consider

$$\begin{aligned} E[Y_2(Y_1(\tilde{d}_1), d)|X = x] &= E[E[Y_2(Y_1(\tilde{d}_1), d)|Y_1(\tilde{d}_1)]|X = x] \\ &= \sum_{y_1 \in \{0,1\}} \Pr[Y_1(\tilde{d}_1) = y_1|X = x] \Pr[Y_2(y_1, d) = 1|Y_1(\tilde{d}_1) = y_1, X = x] \\ &= \sum_{y_1 \in \{0,1\}} \Pr[Y_1(\tilde{d}_1) = y_1, Y_2(y_1, d) = 1|X = x] \\ &= C(\mu_1(\tilde{d}_1, x), \mu_2(1, d, x); \rho_{U_1(\tilde{d}_1), U_2(1, d), x}) \\ &\quad + \mu_2(0, d, x) - C(\mu_1(\tilde{d}_1, x), \mu_2(0, d, x); \rho_{U_1(\tilde{d}_1), U_2(0, d), x}) \\ E[Y_2(d_1)|X = x] &= \sum_{y_1, d_2 \in \{0,1\}^2} \Pr[Y_1(d_1) = y_1, D_2(d_1) = d_2, Y_2(y_1, d) = 1|X = x], \end{aligned}$$

⁶Note that $Y_2(Y_1, d)$ and $Y_2(d_1, D_2)$ are counterfactual objects with different interpretations and $Y_2(Y_1, d) \neq Y_2(d)$ and $Y_2(d_1, D_2) \neq Y_2(d_1)$.

where the last expression entails a formula with relevant copulas similar to the equation one above. Then, examples of (conditional) dynamic treatment effects can be identified as follows:

$$E[Y_2(\tilde{y}_1, \tilde{d}) - Y_2(y_1, d)|X = x] = \mu_2(\tilde{y}_1, \tilde{d}, x) - \mu_2(y_1, d, x)$$

and

$$\begin{aligned} E[Y_2(\tilde{d}_1, d_2) - Y_2(d_1, d_2)|X = x] &= C(\mu_1(\tilde{d}_1, x), \mu_2(1, \tilde{d}_1, d_2, x); \rho_{U_1(\tilde{d}_1), U_2(1, \tilde{d}_1, d_2), x}) \\ &\quad + \mu_2(0, \tilde{d}_1, d_2, x) - C(\mu_1(\tilde{d}_1, x), \mu_2(0, \tilde{d}_1, d_2, x); \rho_{U_1(\tilde{d}_1), U_2(0, \tilde{d}_1, d_2), x}) \\ &\quad - \{C(\mu_1(d_1, x), \mu_2(1, d, x); \rho_{U_1(d_1), U_2(1, d), x}) \\ &\quad + \mu_2(0, d, x) - C(\mu_1(d_1, x), \mu_2(0, d, x); \rho_{U_1(d_1), U_2(0, d), x})\}. \end{aligned}$$

Note that in the first example, we can learn dynamic complementarity by setting $\tilde{y}_1 = y_1$ and comparing $\tilde{d} = (0, 1)$ and $d = (0, 0)$ with $\tilde{d} = (1, 1)$ and $d = (1, 0)$:

$$E[Y_2(y_1, 0, 1) - Y_2(y_1, 0, 0)] \quad \text{vs.} \quad E[Y_2(y_1, 1, 1) - Y_2(y_1, 1, 0)].$$

Also, we can learn state dependence by setting $\tilde{d} = d$ and $\tilde{y}_1 = 1$ and $y_1 = 0$:

$$E[Y_2(1, d) - Y_2(0, d)].$$

In the second example, we can decompose the parameter into the direct effect and indirect effect mediated by Y_1 (i.e., the mediation effect) as follows. Note that $Y_2(d) = Y_2(Y_1(d_1), d)$.

Therefore,

$$\begin{aligned}
E[Y_2(1, d_2) - Y_2(0, d_2)|X = x] &= E[Y_2(Y_1(0), 1, d_2) - Y_2(Y_1(0), 0, d_2)|X = x] \\
&\quad + E[Y_2(Y_1(1), 1, d_2) - Y_2(Y_1(0), 1, d_2)|X = x] \\
&= E[Y_2(Y_1(1), 1, d_2) - Y_2(Y_1(1), 0, d_2)|X = x] \\
&\quad + E[Y_2(Y_1(1), 0, d_2) - Y_2(Y_1(0), 0, d_2)|X = x],
\end{aligned}$$

where the expressions for $E[Y_2(Y_1(d_1), d_1, d_2)|X = x]$ are given above and $E[Y_2(Y_1(d_1), d'_1, d_2)|X = x]$ ($d_1 \neq d'_1$) can be recovered by using similar expressions as that for $E[Y_2(d)]$ above. Note that the mediation effect is path-dependent, so we define two different versions of direct and indirect effects. The unconditional versions of all the effects above can be recovered by taking expectations over X . Next, the time-, “sector-” and covariate- specific selection can be measured by $(\rho_{V_1, U_1(d_1), x}, \rho_{V_2, U_2(y_1, d), x})$, which are also identified in the next section. Finally, let $D_2(y_1, d_1)$ be the counterfactual treatment given (y_1, d_1) and let $p_{Z_2}(x) \equiv \Pr[Z_2 = 1|X = x]$. If we identify $\pi_2(y_1, d_1, z_2, x)$, then we identify

$$E[D_2(y_1, d_1)|X = x] = p_{Z_2}(x)\pi_2(y_1, d_1, 1, x) + (1 - p_{Z_2}(x))\pi_2(y_1, d_1, 0, x).$$

This counterfactual object can be used to measure habit and learning in the treatment decision. The habit of decisions can be captured by

$$E[D_2(y_1, 1) - D_2(y_1, 0)]$$

and the learning from the previous experience can be reflected in the complementarity of Y_1 and D_1 in forming D_2 :

$$E[D_2(1, 1) - D_2(0, 1)] - E[D_2(1, 0) - D_2(0, 0)].$$

Remark 3.1. As seen in $\Pr[U_2(y_1, d) \leq \mu_2(y_1, d, x)|X = x] = \mu_2(y_1, d, x)$, the marginal dis-

tribution of $U_2(y_1, d)$ is absorbed in $\mu_2(y_1, d, x)$ due to the normalization that $U_2(y_1, d)|X = x$ is uniform regardless of (y_1, d) . Nonetheless, the misspecification of $U_2(y_1, d) = U_2(y'_1, d') \equiv U_2((y_1, d) \neq (y'_1, d'))$ will have consequences in identifying and consistently estimating $\mu_2(y_1, d, x)$.

4 Identification Analysis

We conduct the identification analysis in the model (2.1)–(2.4). We show the identifiability of $\mu_2(y_1, d, x)$, $\mu_1(d_1, x)$, $\pi_2(y_1, d_1, z_2)$ and $\Sigma(y_1, d, x)$ in three steps. First, consider

$$\begin{aligned} Y_1 &= 1[\mu_1(D_1, X) \geq U_1(D_1)], \\ D_1 &= 1[\pi_1(Z_1, X) \geq V_1]. \end{aligned}$$

Fix $x \in \mathcal{X}$. Note that $\pi_1(z_1, x)$ is trivially identified as $\pi_1(z_1, x) = \Pr[D_1 = 1|Z_1 = z_1, X = x]$ by our normalization. We list $\Pr[D_1 = d, Y_1 = y|Z_1 = z, X = x]$ for $(d, y, z) \in \{0, 1\}^3$. For example, by Assumptions 2.2 and 2.4,

$$\begin{aligned} \Pr[D_1 = 1, Y_1 = 1|Z_1 = 0, X = x] &= \Pr[V_1 \leq \pi_1(0, x), U_1(1) \leq \mu_1(1, x)|X = x] \\ &= C(\pi_1(0, x), \mu_1(1, x); \rho_{V_1, U_1(1), x}). \end{aligned}$$

The six (non-redundant) fitted probabilities can be written as follows:

$$\Pr[D_1 = 1, Y_1 = 1|Z_1 = 0, X = x] = C(\pi_1(0, x), \mu_1(1, x); \rho_{V_1, U_1(1), x}), \quad (4.1)$$

$$\Pr[D_1 = 1, Y_1 = 1|Z_1 = 1, X = x] = C(\pi_1(1, x), \mu_1(1, x); \rho_{V_1, U_1(1), x}), \quad (4.2)$$

$$\Pr[D_1 = 0, Y_1 = 1|Z_1 = 0, X = x] = \mu_1(0, x) - C(\pi_1(0, x), \mu_1(0, x); \rho_{V_1, U_1(0), x}), \quad (4.3)$$

$$\Pr[D_1 = 0, Y_1 = 1|Z_1 = 1, X = x] = \mu_1(0, x) - C(\pi_1(1, x), \mu_1(0, x); \rho_{V_1, U_1(0), x}), \quad (4.4)$$

Under Assumptions 2.3 and 2.5(i), by applying Lemma 4.1 of Han and Vytlačil (2017), it is easy to see that the Jacobian matrix of (4.1)–(4.2)

$$J_1(x) = \begin{bmatrix} C_2(\pi_1(0, x), \mu_1(1, x)) & C_{\rho_{V_1, U_1(1), x}}(\pi_1(0, x), \mu_1(1, x)) \\ C_2(\pi_1(1, x), \mu_1(1, x)) & C_{\rho_{V_1, U_1(1), x}}(\pi_1(1, x), \mu_1(1, x)) \end{bmatrix}$$

is a P-matrix (except at the boundary of the parameter space). Similarly, the Jacobian of (4.3)–(4.4)

$$J_2(x) = \begin{bmatrix} 1 - C_2(\pi_1(0, x), \mu_1(0, x)) & C_{\rho_{V_1, U_1(0), x}}(\pi_1(0, x), \mu_1(0, x)) \\ 1 - C_2(\pi_1(1, x), \mu_1(0, x)) & C_{\rho_{V_1, U_1(0), x}}(\pi_1(1, x), \mu_1(0, x)) \end{bmatrix}$$

is a P-matrix. Therefore, we can apply the global univalence theorem by Gale and Nikaido (1965) and identify $(\mu_1(d_1, x), \rho_{V_1, U_1(d_1), x})$ for all $d_1 \in \{0, 1\}$ and $x \in \mathcal{X}$. This step is closely related to Han and Vytlačil (2017).⁷

For the rest of the proof, we suppress X for simplicity but the idea of incorporating X is the same as above. Next, we consider identification of $(\pi_2(y_1, d_1, z_2), \rho_{V_1, V_2})$ for each (y_1, d_1, z_2) using

$$D_2 = 1[\pi_2(Y_1, D_1, Z_2) \geq V_2],$$

$$Y_1 = 1[\mu_1(D_1) \geq U_1(D_1)],$$

$$D_1 = 1[\pi_1(Z_1) \geq V_1].$$

⁷Note that Gale and Nikaido (1965)'s theorem does not require the technical assumptions on the parameter space used for Hadamard's global inverse function theorem in Han and Vytlačil (2017). The latter uses all the fitted probabilities to calculate a larger Jacobian matrix, which is not a P-matrix, and thus Gale and Nikaido (1965)'s theorem is not applicable.

Fix $z_2 \in \{0, 1\}$. For $z_1 \in \{0, 1\}$, consider

$$\begin{aligned} & \Pr[D_1 = 1, D_2 = 1, Y_1 = 1 | Z_1 = z_1, Z_2 = z_2] \\ &= \Pr[V_1 \leq \pi_1(z_1), V_2 \leq \pi_2(1, 1, z_2), U_1(1) \leq \mu_1(1)] \\ &= C(\pi_1(z_1), \pi_2(1, 1, z_2), \mu_1(1); \rho_{V_1, V_2}, \rho_{V_1, U_1(1)}). \end{aligned}$$

Note that we write the copula with two dependence parameters (instead of three) without loss of generality, and similarly for 4-copula below. There are alternative copula representations with alternative pairs of dependence parameters. The dependence parameters in those models can be recovered from the parameters in the current model. For example, $\rho_{V_1, U_2(d)}$ can be recovered from $(\rho_{V_1, V_2}, \rho_{V_2, U_2(d)})$; see Example 4.4 in [Darsow et al. \(1992\)](#). Then, the Jacobian for $(\pi_2(1, 1, z_2), \rho_{V_1, V_2})$ is

$$J_3 = \begin{bmatrix} C_2(\pi_1(0), \pi_2(1, 1, z_2), \mu_1(1)) & C_{\rho_{V_1, V_2}}(\pi_1(0), \pi_2(1, 1, z_2), \mu_1(1)) \\ C_2(\pi_1(1), \pi_2(1, 1, z_2), \mu_1(1)) & C_{\rho_{V_1, V_2}}(\pi_1(1), \pi_2(1, 1, z_2), \mu_1(1)) \end{bmatrix},$$

which is a P-matrix if and only if

$$\frac{C_2(\pi_1(0), \pi_2(1, 1, z_2), \mu_1(1))}{C_{\rho_{V_1, V_2}}(\pi_1(0), \pi_2(1, 1, z_2), \mu_1(1))} \neq \frac{C_2(\pi_1(1), \pi_2(1, 1, z_2), \mu_1(1))}{C_{\rho_{V_1, V_2}}(\pi_1(1), \pi_2(1, 1, z_2), \mu_1(1))}. \quad (4.5)$$

The latter is guaranteed by Assumptions [2.3](#) and [2.5](#) and the following lemma:

Lemma 4.1. *Suppose $C(v_1, v_2, u_1, u_2; \rho_{v_1 v_2}, \rho_{v_1 u_1}, \rho_{v_2 u_2})$ and its margins (where only relevant dependence parameters are shown) satisfy Assumption [2.5\(ii\)–\(iii\)](#). Then for any $\rho_{v_1 u_1}, \rho_{v_1 v_2}, \rho_{v_2 u_2} \in (-1, 1)$ and $v_1, u_1, u_2 \in (0, 1)$, the copula and its semi-survival functions and all their margins satisfy*

$$\frac{H_2(v_1, v_2, u_1; \rho_{v_1 v_2})}{H_{\rho_{v_1 v_2}}(v_1, v_2, u_1; \rho_{v_1 v_2})} \quad (4.6)$$

is strictly decreasing in v_1 , and

$$\frac{H_4(v_1, v_2, u_1, u_2; \rho_{v_2 u_2})}{H_{\rho_{v_2 u_2}}(v_1, v_2, u_1, u_2; \rho_{v_2 u_2})} \quad (4.7)$$

is strictly decreasing in v_2 .

This lemma exhibits one of the theoretical contributions of this paper. To prove that the vine copula in Assumption 2.5 implies (4.6)–(4.7), we employ the functional derivative (Gelfand et al. (2000)) of the vine copula with respect to the conditional copula and exploit the ordering property of the outer copula to restore the ordering property of the multivariate *unconditional* copulas. Then, we prove that the latter guarantees the monotonicity of the ratios in the lemma. For details, see Section A in the Appendix.

Returning to (4.5), by Lemma 4.1, we identify $(\pi_2(1, 1, z_2), \rho_{V_1, V_2})$. In what follows, parallel arguments are made for $\pi_2(y_1, d_1, z_2)$ with other values of (y_1, d_1) by applying Lemma 4.1 with corresponding H functions. We identify $\pi_2(0, 1, z_2)$ from the conditional probabilities with $Y_1 = 0$: For $z_1 \in \{0, 1\}$, consider

$$\begin{aligned} & \Pr[D_1 = 1, D_2 = 1, Y_1 = 0 | Z_1 = z_1, Z_2 = z_2] \\ &= \Pr[V_1 \leq \pi_1(z_1), V_2 \leq \pi_2(0, 1, z_2), U_1(1) > \mu_1(1)] \\ &= C(\pi_1(z_1), \pi_2(0, 1, z_2); \rho_{V_1, V_2}) - C(\pi_1(z_1), \pi_2(0, 1, z_2), \mu_1(1); \rho_{V_1, V_2}, \rho_{V_1, U_1(1)}). \end{aligned}$$

Note that the r.h.s. expression is strictly increasing in $\pi_2(0, 1, z_2)$, which is clear to see from the second line above. Also note that the expression is a *known* function $\pi_2(0, 1, z_2)$ because the other components in the copulas are all identified earlier. Therefore, by inverting this

function, we identify $\pi_2(0, 1, z_2)$. We repeat this proof for $D_1 = 0$. For $z_1 \in \{0, 1\}$, consider

$$\begin{aligned} & \Pr[D_1 = 0, D_2 = 1, Y_1 = 1 | Z_1 = z_1, Z_2 = z_2] \\ &= \Pr[V_1 > \pi_1(z_1), V_2 \leq \pi_2(1, 0, z_2), U_1(0) \leq \mu_1(0)] \\ &= C(\pi_2(1, 0, z_2), \mu_1(0); \rho_{V_2, U_1(0)}) - C(\pi_1(z_1), \pi_2(1, 0, z_2), \mu_1(0); \rho_{V_1, V_2}, \rho_{V_1, U_1(0)}). \end{aligned}$$

Similarly as before, this expression is a known strictly increasing function of $\pi_2(1, 0, z_2)$ because all other components are identified earlier; note $\rho_{V_2, U_1(0)}$ can be recovered from $(\rho_{V_1, V_2}, \rho_{V_1, U_1(0)})$. Consider

$$\begin{aligned} & \Pr[D_1 = 0, D_2 = 1, Y_1 = 0 | Z_1 = z_1, Z_2 = z_2] \\ &= \Pr[V_1 > \pi_1(z_1), V_2 \leq \pi_2(0, 0, z_2), U_1(0) > \mu_1(0)] \\ &= \pi_2(0, 0, z_2) - C(\pi_1(z_1), \pi_2(0, 0, z_2); \rho_{V_1, V_2}) \\ &\quad - \{C(\pi_2(0, 0, z_2), \mu_1(0); \rho_{V_2, U_1(0)}) - C(\pi_1(z_1), \pi_2(0, 0, z_2), \mu_1(0); \rho_{V_1, V_2}, \rho_{V_1, U_1(0)})\}, \end{aligned}$$

which is a known strictly increasing function of $\pi_2(0, 0, z_2)$. This identifies $\pi_2(0, 0, z_2)$. In sum, this step identifies $\pi_2(y_1, d_1, z_2)$ for all $(y_1, d_1, z_2) \in \{0, 1\}^3$ and ρ_{V_1, V_2} .

Finally, consider

$$\begin{aligned} Y_2 &= 1[\mu_2(Y_1, D) \geq U_2(Y_1, D)], \\ D_2 &= 1[\pi_2(Y_1, D_1, Z_2) \geq V_2], \\ Y_1 &= 1[\mu_1(D_1) \geq U_1(D_1)], \\ D_1 &= 1[\pi_1(Z_1) \geq V_1], \end{aligned}$$

where the remaining parameters to identify are $(\mu_2(y_1, d), \rho_{V_2, U_2(y_1, d)})$ for $(y_1, d) \in \{0, 1\}^2$.

Fix $z_1 \in \{0, 1\}$. First, consider

$$\begin{aligned}
& \Pr[D_1 = 1, D_2 = 1, Y_1 = 1, Y_2 = 1 | Z_1 = z_1, Z_2 = z_2] \\
&= \Pr[V_1 \leq \pi_1(z_1), V_2 \leq \pi_2(1, 1, z_2), U_1(1) \leq \mu_1(1), U_2(1, 1, 1) \leq \mu_2(1, 1, 1)] \\
&= C(\pi_1(z_1), \pi_2(1, 1, z_2), \mu_1(1), \mu_2(1, 1, 1); \rho_{V_1, V_2}, \rho_{V_1, U_1(1)}, \rho_{V_2, U_2(1, 1, 1)}).
\end{aligned}$$

By varying $z_2 \in \{0, 1\}$ we can identify $(\mu_2(1, 1, 1), \rho_{V_2, U_2(1, 1, 1)})$ from a relevant Jacobian matrix, which is again a P-matrix by Assumptions 2.3 and 2.5 and Lemma 4.1. Similarly, consider

$$\begin{aligned}
& \Pr[D_1 = 1, D_2 = 0, Y_1 = 1, Y_2 = 1 | Z_1 = z_1, Z_2 = z_2] \\
&= \Pr[V_1 \leq \pi_1(z_1), V_2 > \pi_2(1, 1, z_2), U_1(1) \leq \mu_1(1), U_2(1, 1, 0) \leq \mu_2(1, 1, 0)] \\
&= C(\pi_1(z_1), \mu_1(1), \mu_2(1, 1, 0); \rho_{V_1, U_1(1)}, \rho_{V_2, U_2(1, 1, 0)}) \\
&\quad - C(\pi_1(z_1), \pi_2(1, 1, z_2), \mu_1(1), \mu_2(1, 1, 0); \rho_{V_1, V_2}, \rho_{V_1, U_1(1)}, \rho_{V_2, U_2(1, 1, 0)}).
\end{aligned}$$

By varying $z_2 \in \{0, 1\}$ we can identify $(\mu_2(1, 1, 0), \rho_{V_2, U_2(1, 1, 0)})$ from a relevant Jacobian matrix, which is again a P-matrix by Assumptions 2.3 and 2.5 and Lemma 4.1. By changing the possible remaining values of (D_1, D_2, Y_1) , the remainder of the proof is analogous to the two cases above, which identifies $(\mu_2(y_1, d), \rho_{V_2, U_2(y_1, d)})$ for the remaining values of (y_1, d) . The following theorem summarizes the identification results for the case of $T = 2$. Let \mathcal{X} be the support of X .

Theorem 4.1. *Under Assumptions 2.2–2.5, the parameters*

$$(\pi_1(z_1, x), \mu_1(d_1, x), \pi_2(y_1, d_1, z_2, x), \mu_2(y_1, d, x), \Sigma(y_1, d, x))$$

as functions of x are globally identified for all $(y_1, z, d, x) \in \{0, 1\}^5 \times \mathcal{X}$.

Remark 4.1. *One may be curious whether the three step approach is necessarily in the proof of identification. The is in fact the case because, with a two-step approach of the following,*

ρ_{12} is not identified:

$$\begin{aligned}
& \Pr[D_2 = 1, Y_2 = 1 | D_1 = 1, Y_1 = 1, Z_1 = z_1, Z_2 = z_2] \\
&= \Pr[V_2 \leq \pi_2(1, 1, z_2), U_2(1, 1, 1) \leq \mu_2(1, 1, 1) | V_1 \leq \pi_1(z_1), U_1(1) \leq \mu_1(1)] \\
&= \frac{1}{\pi_1(z_1)\mu_1(1)} \int^{\pi_1(z_1)} \int^{\mu_1(1)} C(\pi_2(1, 1, z_2), \mu_2(1, 1, 1) | v_1, u_1; \rho_{V_1, V_2}, \rho_{V_1, U_1(1)}, \rho_{V_2, U_2(1, 1, 1)}) dv_1 du_1.
\end{aligned}$$

5 Identification with General T

We now give an overview of a model with general T and related identification results. For any random variable W_t , let $W^t \equiv (W_1, \dots, W_t)$ and $W \equiv W^T$. We also use the convention that $W^0 = W_0 = \emptyset$. For $t = 1, \dots, T$, consider

$$Y_t = 1[\mu_t(Y_{t-1}, D^t, X) \geq U_t(Y_{t-1}, D^t)], \quad (5.1)$$

$$D_t = 1[\pi_t(Y_{t-1}, D_{t-1}, Z_t, X) \geq V_t]. \quad (5.2)$$

Let $U^t(y^{t-1}, d^t) \equiv (U_1(d_1), U_2(y_1, d^2), \dots, U_t(y_{t-1}, d^t))$.

Assumption 5.1. (i) (No Anticipation) $Y_t(y^{t-1}, d^T, z^t) = Y_t(y^{t-1}, d^t, z^t)$ and $D_t(d_{-t}^T, z^t) = D_t(d^{t-1}, z^t)$ for $t = 1, \dots, T$; (ii) (Exclusion) $Y_t(y^{t-1}, d^t, z^t) = Y_t(y^{t-1}, d^t)$ for $t = 1, \dots, T$.

Assumption 5.2. $Z \perp (V, U(y^{T-1}, d)) | X$ for $(y^{T-1}, d^T) \in \{0, 1\}^{2T-1}$.

Assumption 5.3. For $t = 1, \dots, T$, π_t is a non-trivial function of Z_t and $Z | X$ is non-degenerate.

Assumption 5.4. For each $(y^{T-1}, d^T) \in \{0, 1\}^{2T-1}$, the unobservables are jointly distributed as

$$(V, U(y^{T-1}, d)) |_{X=x} \sim C(v, u; \Sigma(y^{T-1}, d, x)),$$

where $C(v, u; \Sigma)$ is a $2T$ -copula with dependence matrix Σ .

Assumption 5.5. *The copula $C(v, u; \Sigma)$ in Assumption 5.4 and all its margins satisfy pairwise “ \prec_{SJ} ” with respect to the associated dependence parameter.*

Assumption 5.3*. *The following conditions hold:*

- (i) *Condition (i) of Assumption 2.5 holds;*
- (ii) *the conditional versions of $C(v, u; \Sigma)$ and its margins are represented by*

$$C(v_{t-1}, v_t | u^{t-1}, v^{t-2}; \Sigma_{v^t u^{t-1}}) = C(C(v_{t-1} | u^{t-1}, v^{t-2}), C(v_t | u^{t-1}, v^{t-2}); \rho(\Sigma_{v^t u^{t-1}})),$$

$$\frac{C(v_t, u_t | v^{t-1}, u^{t-1}; \Sigma_{v^t u^t})}{c(v^{t-1}, u^{t-1}; \Sigma_{v^{t-1} u^{t-1}})} = C(C(v_t | v^{t-1}, u^{t-1}), C(u_t | v^{t-1}, u^{t-1}); \rho(\Sigma_{v^t u^t}))$$

for $t = 2, \dots, T$, where the outer copula $C(\cdot, \cdot; \rho)$ on the r.h.s. satisfies $C(\cdot, \cdot; \rho) \prec_S C(\cdot, \cdot; \tilde{\rho})$ for $\rho < \tilde{\rho}$;

- (iii) $\rho(\Sigma_{v^t u^{t-1}})$ and $\rho(\Sigma_{v^t u^t})$ are strictly increasing in $\rho_{v_{t-1} v_t}$ and $\rho_{v_t u_t}$, respectively.

In Assumption 5.5*(ii), for example, $\Sigma_{v^2 u_1} = (\rho_{v_1 v_2}, \rho_{v_1 u_1}, \rho_{v_2 u_1})$ and

$$\Sigma_{v^2 u^2} = (\rho_{v_2 u_2}, \rho_{v_1 v_2}, \rho_{v_2 u_1}, \rho_{v_1 u_2}, \rho_{u_1 u_2}, \rho_{v_1 u_1}).$$

To state the generalized version of Lemma 4.1, let

$$C(v_{t-1}, v_t, v^{t-2}, u^{t-1}; \rho_{v_{t-1} v_t}) \equiv C(v_{t-1}, v_t, v^{t-2}, u^{t-1}; \rho_{v_{t-1} v_t}, \Sigma_{v^{t-2} u^{t-1}}),$$

$$C(v_t, u_t, v^{t-1}, u^{t-1}; \rho_{v_t u_t}) \equiv C(v_t, u_t, v^{t-1}, u^{t-1}; \rho_{v_t u_t}, \Sigma_{v^{t-1} u^{t-1}}).$$

Note $\Sigma_{v^{t-2} u^{t-1}}$ and $\Sigma_{v^{t-1} u^{t-1}}$ are identified in previous steps. Let conditioning variables v^0 and u^0 mean no conditioning.

Lemma 5.1. *Suppose $C(v, u; \Sigma)$ and its margin satisfy Assumption 5.5. Then for $t = 2, \dots, T$, the copula and its semi-survival functions and all their margins satisfy*

$$\frac{H_{\rho_{v_{t-1} v_t}}(v_{t-1}, v_t, v^{t-2}, u^{t-1}; \rho_{v_{t-1} v_t})}{H_2(v_{t-1}, v_t, v^{t-2}, u^{t-1}; \rho_{v_{t-1} v_t})}$$

is strictly monotonic in v_{t-1} , and for $t = 1, \dots, T$

$$\frac{H_{\rho_{v_t u_t}}(v_t, u_t, v^{t-1}, u^{t-1}; \rho_{v_t u_t})}{H_2(v_t, u_t, v^{t-1}, u^{t-1}; \rho_{v_t u_t})}$$

are strictly monotonic in v_t .

Based on this lemma, we can follow identification arguments analogous to those in Section 4. The key observation for the identification is that, regardless of T , each step only involves a 2×2 Jacobian matrix, which is easy to show to be a P-matrix under Assumptions 5.3– 5.5.

Theorem 5.1. *Under Assumptions 5.2–5.5, the parameters*

$$(\pi_t(y_{t-1}, d_{t-1}, z_t, x), \mu_t(y_{t-1}, d^t, x), \Sigma(y^{T-1}, d, x)) \quad \text{for all } t = 1, \dots, T$$

as functions of x are globally identified for all $(y^{T-1}, z, d, x) \in \{0, 1\}^{3T-1} \times \mathcal{X}$.

6 Extension: Continuous Outcome Variables

We extend the identification results of this paper to the case of continuous outcome variables. With continuous outcomes, we can recover parameters defined in Section 3 for *both* average and quantile effects.

Let $Y_t \in \mathcal{Y}_t \subseteq \mathbb{R}$ for $t = 1, \dots, T$. We consider $T = 2$ for simplicity. Now, define the (continuous) counterfactual outcomes $Y_1(d_1)$ and $Y_2(y_1, d)$ and the counterfactual treatments $D_1(z_1)$ and $D_2(y_1, d_1, z_2)$. We maintain the same treatment selection model (2.2) and (2.4) for D_2 and D_1 , respectively. However, we do not specify the mechanism under which the outcome is formed, which is similar to the spirit of (2.1) and (2.4) as discussed in the next section. This framework allows us to maintain the same set of dynamic treatment and mediation parameters as in the discrete case.

We introduce a new set of identifying assumptions. At the same time, we maintain Assumptions 2.1 (no anticipation and exclusion), 2.3 (IV relevance) and 2.5 (vine copula).

Assumption 6.1. $Z \perp (V_1, V_2, Y_1(d_1), Y_2(y_1, d)) | X$ for $(d, \tilde{y}_1) \in \{0, 1\}^3$.

Assumption 6.2. For each $(d, \tilde{y}_1) \in \{0, 1\}^3$, the unobservables are jointly distributed as

$$(V_1, V_2, Y_1(d_1), Y_2(y_1, d)) |_{X=x} \sim C(v_1, v_2, F_{Y_1(d_1)}(y_1), F_{Y_2(y_1, d)}(y_2); \Sigma(d, y_1, x)),$$

where $C(v_1, v_2, u_1, u_2; \Sigma)$ is a 4-copula with dependence matrix Σ .

In Assumption 6.2, $\Sigma(d, y_1, x)$ captures all the dependences among $(V_1, V_2, Y_1(d_1), Y_2(y_1, d))$ conditional on $X = x$. Notable elements in $\Sigma(d, x)$ are $\rho_{V_1, Y_1(d_1), x}$ and $\rho_{V_t, Y_2(y_1, d), x}$ (for $t = 1, 2$ and $d \in \{0, 1\}^2$). Below, we use $\rho_{V_1, U_1(d_1), x}$ and $\rho_{V_t, U_2(y_1, d), x}$ interchangeably, where $U_1(d_1)$ and $U_2(y_1, d)$ are the CDF transformations of $Y_1(d_1)$ and $Y_2(y_1, d)$.

We briefly outline the identification strategy. We suppress X for simplicity. For given $y \in \mathcal{Y}$, consider

$$\begin{aligned} F_{Y_1|D_1, Z_1}(y|D_1 = 1, Z_1 = z_1)\pi_1(z_1) &= \Pr[Y_1 \leq y, V_1 \leq \pi_1(z_1)] \\ &= \Pr[U_1(1) \leq F_{Y_1(1)}(y), V_1 \leq \pi_1(z_1)] \\ &= C(\pi_1(z_1), F_{Y_1(1)}(y); \rho_{V_1, U_1(1)}) \end{aligned}$$

and similarly for $D_1 = 0$. From these equations, we identify $(\pi_1(z_1), F_{Y_1(d_1)}(y), \rho_{V_1, U_1(d_1)})$ by the same argument as the first step of identification in Section 4. Next, we want to identify $\pi_2(y_1, z_2)$ and ρ_{V_1, V_2} . For given $y \in \mathcal{Y}$, consider

$$\begin{aligned} &\Pr[D_1 = 1, D_2 = 1, Y_1 \leq y | Z = z] \\ &= \Pr[Y_1 \leq y | D_1 = 1, D_2 = 1, Z = z] \Pr[D_1 = 1, D_2 = 1 | Z = z] \\ &= \Pr[D_1(z_1) = 1, D_2(z_2) = 1 | U_1(1) \leq F_{Y_1(1)}(y)] \Pr[Y_1(1) \leq y] \\ &= \int^{F_{Y_1(1)}(y)} \Pr[V_1 \leq \pi_1(z_1), V_2 \leq \pi_2(F_{Y_1(1)}^{-1}(t), 1, z_2) | U_1(1) = t] dt \\ &= \int^{F_{Y_1(1)}(y)} C(\pi_1(z_1), \pi_2(F_{Y_1(1)}^{-1}(t), 1, z_2) | t) dt. \end{aligned}$$

Then, differentiating both sides w.r.t y yields

$$\begin{aligned} & f_{Y_1}(y|D_1 = 1, D_2 = 1, Z = z) \Pr[D_1 = 1, D_2 = 1|Z = z] \\ &= C(\pi_1(z_1), \pi_2(y, 1, z_2)|F_{Y_1(1)}(y); \rho_{V_1, V_2}) f_{Y_1(1)}(y). \end{aligned}$$

Since $F_{Y_1(1)}(y)$ is already identified, by varying the value of Z_1 , we can produce two equations, whose Jacobian is a P-matrix as in Section 4. Therefore we can identify $\pi_2(y, 1, z_2)$ and ρ_{V_1, V_2} by showing Jacobian is P-matrix. By using $D_1 = 0$ instead, we can also identify $\pi_2(y, 0, z_2)$. The remaining proof can be followed analogous to the discrete case.

Theorem 6.1. *Under Assumptions 2.1, 2.3, 2.5, 6.1 and 6.2, for each $y \in \mathcal{Y}$, the parameters*

$$(\pi_1(z_1, x), F_{Y_1(d_1)|X}(y|x), \pi_2(y_1, d_1, z_2, x), F_{Y_2(y_1, d)|X}(y|x), \Sigma(y_1, d, x))$$

are globally identified for all $(y_1, z, d, x) \in \{0, 1\}^5 \times \mathcal{X}$.

The identified parameters can be used to construct the dynamic treatment effect parameters introduced in Section 3. For example,

$$\begin{aligned} E[D_2(Y_1(\tilde{d}_1), d_1, z_2)] &= E[E[D_2(Y_1(\tilde{d}_1), d_1, z_2)|Y_1(\tilde{d}_1)]] \\ &= \int \Pr[V_2 \leq \pi_2(y_1, d_1, z_2)|Y_1(\tilde{d}_1) = y_1] dF_{Y_1(\tilde{d}_1)}(y_1) \end{aligned}$$

but

$$\begin{aligned} \Pr[V_2 \leq \pi_2(y_1, d_1, z_2)|Y_1(\tilde{d}_1) = y_1] &= \Pr[V_2 \leq \pi_2(y_1, d_1, z_2)|U_1(\tilde{d}_1) = F_{Y_1(\tilde{d}_1)}(y_1)] \\ &= C_1(F_{Y_1(\tilde{d}_1)}(y_1), \pi_2(y_1, d_1, z_2); \rho_{U_1(\tilde{d}_1), V_2}), \end{aligned}$$

where all the components in the last expression are identified in Theorem 6.1. Similarly,

$$\begin{aligned}
E[Y_2(Y_1(\tilde{d}_1), d)] &= E[E[Y_2(Y_1(\tilde{d}_1), d)|Y_1(\tilde{d}_1)]] \\
&= \int E[Y_2(y_1, d)|Y_1(\tilde{d}_1) = y_1]dF_{Y_1(\tilde{d}_1)}(y_1) \\
&= \int \int 1 - F_{Y_2(y_1, d)|Y_1(\tilde{d}_1)}(y_2|y_1)dy_2dF_{Y_1(\tilde{d}_1)}(y_1)
\end{aligned}$$

but

$$\begin{aligned}
F_{Y_2(y_1, d)|Y_1(\tilde{d}_1)}(y_2|y_1) &= P[Y_2(y_1, d) \leq y_2|Y_1(\tilde{d}_1) = y_1] \\
&= P[U_2(y_1, d) \leq F_{Y_2(y_1, d)}(y_2)|U_1(\tilde{d}_1) = F_{Y_1(\tilde{d}_1)}(y_1)] \\
&= C_1(F_{Y_1(\tilde{d}_1)}(y_1), F_{Y_2(y_1, d)}(y_2); \rho_{U_1(\tilde{d}_1), U_2(y_1, d)}).
\end{aligned}$$

Moreover, since we directly identify the CDFs of $Y_1(d_1)$ and $Y_2(y_1, d)$ and the other copula components, we can also identify the quantile versions of dynamic treatment and mediation effects as well as the quantile versions of state dependence and learning and habit effects.

7 Estimation and Inference

7.1 Sieve Maximum Likelihood Estimation

We now consider estimation of the parameters in the semiparametric model. We focus on binary Y_t for simplicity. Let $W \equiv (Y, D, Z, X')' \equiv (Y^T, D^T, Z^T, X')'$ and $\{W_i : i = 1, 2, \dots, n\}$ be a random sample of size n drawn from W . Define the infinite-dimensional parameters

$$h(\cdot) \equiv (\pi_1(z_1, \cdot), \mu_1(d_1, \cdot), \dots, \pi_T(y_{T-1}, d_{T-1}, z_T, \cdot), \mu_T(y_{T-1}, d, \cdot))_{(y^{T-1}, d, z) \in \{0,1\}^{3T-1}}$$

and $\rho(\cdot) \equiv (\Sigma(y^{T-1}, d, \cdot))_{y^{T-1}, d \in \{0,1\}^{2T-1}}$ as functions of $x \in \mathcal{X}$. We denote the vector of the parameters by α (i.e., $\alpha \equiv (h, \rho)'$) and let α_0 be the true parameter value. Let \mathcal{H}^1 and \mathcal{H}^2 be the parameter spaces for h and ρ , respectively, and let $\mathcal{A} \equiv \mathcal{H}^1 \times \mathcal{H}^2$ be the parameter

space for α .

Let $p_{ydz,x,i}(\alpha)$ denote the (normalized) copula function corresponding to $\Pr[Y_i = y, D_i = d, Z_i = z | X_i = x]$. For example, $p_{ydz,x,i}(\alpha)$'s are the r.h.s. objects in (4.4) multiplied by $\Pr[Z_i = z | X_i = x]$. Then, the log-likelihood function is written as

$$L_n(\alpha) = \frac{1}{n} \sum_i^n l(W_i, \alpha), \quad (7.1)$$

where $l(W_i, \alpha) \equiv \sum_{(y,d,z) \in \{0,1\}^{3T}} \mathbf{1}(Y_i = y, D_i = d, Z_i = z) \cdot \log(p_{ydz,x,i}(\alpha))$. Then, a ML estimator of α is obtained by solving

$$\sup_{\alpha \in \mathcal{A}} L_n(\alpha).$$

Since the parameter space \mathcal{A} is infinite-dimensional, it is not feasible to solve the maximization problem over \mathcal{A} . In this paper, we propose to use sieve (ML) estimation. The method of sieves provides a flexible but tractable way to estimate the infinite-dimensional parameters.

A sieve ML estimator $\hat{\alpha}_n$ of α_0 is defined as follows:

$$\hat{\alpha}_n \equiv \arg \max_{\alpha \in \mathcal{A}_n} L_n(\alpha),$$

where \mathcal{A}_n is a sieve space for \mathcal{A} .

We introduce a class of functions of $x \in \mathcal{X}$. Let $g : \mathbb{D} \rightarrow \mathbb{R}$ where $\mathbb{D} \subseteq \mathbb{R}^{d_x}$ for some integer $d_x \geq 1$. For d_x -tuple of nonnegative integers, $\omega = (\omega_1, \dots, \omega_{d_x})$, we define the differential operator as $\nabla^\omega g \equiv \frac{\partial^{|\omega|}}{\partial x_1^{\omega_1} \partial x_2^{\omega_2} \dots \partial x_{d_x}^{\omega_{d_x}}} g(x)$, where $x = (x_1, x_2, \dots, x_{d_x}) \in \mathbb{D}$ and $|\omega| \equiv \sum_{i=1}^{d_x} \omega_i$. Let $p = m + \nu$ be a nonnegative real number with m being a nonnegative integer and $\nu \in (0, 1]$. We call a function $g : \mathcal{X} \rightarrow \mathbb{R}$ p -smooth if it is m times continuously differentiable on \mathcal{X} and for all ω such that $|\omega| = m$ and there exists a constant $c > 0$ such that $|\nabla^\omega g(x) - \nabla^\omega g(y)| \leq c \cdot \|x - y\|_E^\nu$ for all $x, y \in \mathcal{X}$, where $\|\cdot\|_E$ is the Euclidean norm. Let $\mathcal{C}^m(\mathcal{X})$ denote the space of all m -times continuously differentiable real-valued functions on \mathcal{X} . A Hölder ball

with smoothness p and radius $C > 0$ is defined as

$$\Lambda_C^p(\mathcal{X}) \equiv \left\{ g \in \mathcal{C}^m(\mathcal{X}) : \sup_{|\omega| \leq m} \sup_{x \in \mathcal{X}} |\nabla^\omega g(x)| \leq C, \sup_{|\omega|=m} \sup_{x, y \in \mathcal{X}, x \neq y} \frac{|\nabla^\omega g(x) - \nabla^\omega g(y)|}{\|x - y\|_E^\nu} \leq C \right\}.$$

The choice of sieve space \mathcal{A}_n depends on the parameter space \mathcal{A} and the support \mathcal{X} of X . When the parameters belong to some class of smooth functions (e.g., Hölder space, Sobolev space) and \mathcal{X} is compact, one can use polynomial, trigonometric, spline, or wavelet sieve spaces. When \mathcal{X} is unbounded, one can use Hermite polynomial sieve spaces. One can refer to [Chen \(2007\)](#) for the detail on the choice of sieve spaces.

Remark 7.1 (Saturated Semiparametric Models). *It is worth noting that when X is discrete or when there is no X , the semiparametric model we propose is fully saturated. In this case, the estimation problem becomes the standard parametric ML estimation. Given the flexibility we allow for in the model (e.g., heterogeneity), we view this saturation as an appealing feature of our framework. We omitted the standard asymptotic theory for the parametric ML estimation.*

7.2 Asymptotic Theory

We develop the asymptotic theory for the sieve estimator $\hat{\alpha}_n$. To this end, we introduce several norms on \mathcal{A} . For given $\alpha \in \mathcal{A}$, we denote the supremum and L_2 norms of α by $\|\alpha\|_\infty$ and $\|\alpha\|_2$, respectively, where the supremum and integration are taken over \mathcal{X} . We denote the range of the dependence parameters by \mathcal{R} for a given copula function. Define

$$\mathcal{H}_c^{p,1}(\mathcal{X}) \equiv \{g \in \Lambda_c^p(\mathcal{X}) : 0 \leq g(x) \leq 1 \text{ for all } x \in \mathcal{X}\},$$

$$\mathcal{H}_c^{p,2}(\mathcal{X}) \equiv \{g \in \Lambda_c^p(\mathcal{X}) : g(x) \in \mathcal{R} \text{ for all } x \in \mathcal{X}\}.$$

In this paper, we consider linear sieve spaces for \mathcal{A} . Let $\{p_j(\cdot)\}_{j=1}^\infty$ be a sequence of some basis functions and $p^{k_n}(x) \equiv (p_1(x), p_2(x), \dots, p_{k_n}(x))'$. We impose the following assumptions.

Assumption 7.1. (i) The data $\{W_i : i = 1, 2, \dots, n\}$ are i.i.d.; (ii) $E[\|X\|_E^2] < \infty$; (iii) \mathcal{X} is a compact subset of \mathbb{R}^{d_x} .

Assumption 7.2. (i) $\mathcal{H}^1 = \mathcal{H}_c^{p,1}(\mathcal{X})$ and $\mathcal{H}^2 = \mathcal{H}_c^{p,2}$ for some $c > 0$ and $p > 1/2$, and thus, $\mathcal{A} = \mathcal{H}_c^{p,1} \times \dots \times \mathcal{H}_c^{p,1} \times \mathcal{H}_c^{p,2} \times \dots \times \mathcal{H}_c^{p,2}$; (ii) there exists a measurable function $\bar{p}(\cdot)$ on \mathcal{X} such that for any $\alpha \in \mathcal{A}$ and for all $x \in \mathcal{X}$, $p_{y_1 d_1 y_2 d_2 z_2 z_2, x}(\alpha) \geq \bar{p}(x)$ and $E[\bar{p}(X)^{-2}] < \infty$.

Let

$$\begin{aligned}\mathcal{H}_n^1 &\equiv \left\{ p^{k_n}(x)' \beta_n : 0 \leq p^{k_n}(x)' \beta_n \leq 1 \text{ for all } x \in \mathcal{X} \right\}, \\ \mathcal{H}_n^2 &\equiv \left\{ p^{k_n}(x)' \beta_n : p^{k_n}(x)' \beta_n \in \mathcal{R} \text{ for all } x \in \mathcal{X} \right\}.\end{aligned}$$

Assumption 7.3. The following conditions hold: (i) $\mathcal{A}_n = \mathcal{H}_n^1 \times \dots \times \mathcal{H}_n^1 \times \mathcal{H}_n^2 \times \dots \times \mathcal{H}_n^2$, where $k_n/n \rightarrow 0$; (ii) the smallest eigenvalue of $E[p^{k_n}(X) \cdot p^{k_n}(X)']$ is bounded away from zero uniformly in k_n ; (iii) there exists $(\pi_n \alpha_0)_n$ such that $\|\alpha_0 - \pi_n \alpha_0\|_\infty = O(k_n^{-\gamma})$ for some $\gamma > 0$.

Assumption 7.4. The pathwise derivative of the copula function with respect to each dependence parameter is uniformly bounded and continuous.

Assumption 7.2 defines the parameter space. The degrees of smoothness can be different across the parameter spaces, and it is assumed to be identical for simplicity. We may need to impose additional restrictions on the parameter space, especially for the dependence parameters. The range of the dependence parameters, \mathcal{R} , varies across copula functions. For example, when we use the Gaussian copula, we need to impose that the dependence parameters lie in $[-1, 1]$. Assumption 7.2(ii) holds if we observe the fitted probabilities for all possible combinations of the values of (y, d, z) for each $x \in \mathcal{X}$.

Assumption 7.3 defines the sieve space for \mathcal{A} . We consider linear sieve spaces. Assumption 7.3(iii) holds under Assumption 7.2 if we choose polynomial, trigonometric, or spline sieve spaces. For example, if $(p_j(\cdot))_{j=1}^\infty$ is a sequence of polynomial or spline functions, then $\gamma = \frac{p}{d_x}$ by Newey (1997).

Assumption 7.4 imposes some smoothness of the copula function, which holds with many copula functions, including the Gaussian copula.

Under these assumptions, we show the sieve ML estimator is consistent with respect to $\|\cdot\|_\infty$.

Theorem 7.1. *Suppose that Assumptions 2.2–2.5 hold. If Assumptions 7.1–7.4 are satisfied, then,*

$$\|\hat{\alpha}_n - \alpha_0\|_\infty \xrightarrow{p} 0.$$

We now establish the convergence rate of the sieve estimator with respect to $\|\cdot\|_2$. For given $\epsilon > 0$, define an ϵ -neighborhood of α_0 with respect to the consistency norm $\|\cdot\|_\infty$ as $\mathcal{A}_n(\epsilon) \equiv \mathcal{A}_n \cap \mathcal{A}(\epsilon)$, where $\mathcal{A}(\epsilon) \equiv \{\alpha \in \mathcal{A} : \|\alpha - \alpha_0\|_\infty < \epsilon\}$.

Assumption 7.5. $\|\alpha - \alpha_0\|_2^2 \asymp E[l(W, \alpha_0) - l(W, \alpha)]$ for all $\alpha \in \mathcal{A}_n(\epsilon)$.

Note that Assumption 7.5 is not restrictive when focusing on a neighborhood of α_0 . Since we show that the sieve estimator $\hat{\alpha}_n$ is consistent, it suffices to consider a neighborhood of α_0 . Assumption 7.5 is standard in the literature on M-estimation (see, for example, Section 12.3 of [van de Geer \(2000\)](#)).

The following theorem establishes the convergence rate of

Theorem 7.2. *Suppose that Assumptions 2.2–2.5 and Assumptions 7.1–7.5 hold. Then,*

$$\|\hat{\alpha}_n - \alpha_0\|_2 = O_p \left(\max \left\{ \sqrt{\frac{k_n}{n}}, \|\pi_n \alpha_0 - \alpha_0\|_2 \right\} \right).$$

Let $\xi_n \equiv \sup_{x \in \mathcal{X}} \|p_j^{k_n}(x)\|_E$. If we additionally assume $\xi_n^2 k_n / n \rightarrow 0$, then

$$\|\hat{\alpha}_n - \alpha_0\|_\infty = O_p \left(\max \left\{ \xi_n \sqrt{\frac{k_n}{n}}, \|\pi_n \alpha_0 - \alpha_0\|_\infty \right\} \right).$$

Now, we develop the asymptotic normality of functionals of the sieve estimator. While asymptotic normality is useful enough to perform statistical inference on functionals, the

main practical challenge is to consistently estimate the asymptotic variance. To address this, we show that sieve likelihood ratio (LR) test statistics converge in distribution of a χ^2 distribution. We adopt the results of [Chen and Liao \(2014\)](#), who develop a sieve inference method that is valid regardless of whether a functional of interest is regular or irregular.⁸ Since one does not need to verify whether a functional of interest is regular or not, the proposed inferential method has great practicality for empirical research.

Let

$$\Delta(W, \alpha_0) \equiv \lim_{\tau \rightarrow 0} \frac{l(W, \alpha_0 + \tau[\alpha - \alpha_0]) - l(W, \alpha_0)}{\tau}$$

be the pathwise derivative of $l(W, \alpha)$ at α_0 in the direction $[\alpha - \alpha_0]$. Then, for any $\alpha \in \mathcal{A}(\epsilon)$,

$$\|\alpha - \alpha_0\|^2 \equiv - \left. \frac{\partial E [\Delta(W, \alpha_0 + \tau[\alpha - \alpha_0])][\alpha - \alpha_0]]}{\partial \tau} \right|_{\tau=0}$$

defines a norm on $\mathcal{A}(\epsilon)$ by the fact that α_0 is the unique maximizer of $L_0(\alpha)$ over \mathcal{A} . Let \mathcal{V} be the closed linear span of $\mathcal{A}(\epsilon) - \{\alpha_0\}$ under $\|\cdot\|$. Then, \mathcal{V} is a Hilbert space under $\|\cdot\|$, and its inner product is defined as

$$\langle v_1, v_2 \rangle \equiv - \left. \frac{\partial E [\Delta(W, \alpha_0 + \tau[v_2])][v_1]]}{\partial \tau} \right|_{\tau=0}$$

for any $v_1, v_2 \in \mathcal{V}$. Let $\alpha_{0,n} \equiv \arg \min_{\alpha \in \mathcal{A}_n(\epsilon)} \|\alpha - \alpha_0\|$ and \mathcal{V}_n be the closed linear span of $\mathcal{A}_n(\epsilon) - \{\alpha_{0,n}\}$ under $\|\cdot\|$. Note that \mathcal{V}_n is a finite-dimensional Hilbert space under $\|\cdot\|$. Let $f(\cdot) : \mathcal{A} \rightarrow \mathbb{R}$ be a functional on \mathcal{A} and define the pathwise derivative of $f(\cdot)$ at α_0 in the direction of $v = \alpha - \alpha_0 \in \mathcal{V}$ as

$$\frac{\partial f(\alpha_0)}{\partial \alpha}[v] \equiv \left. \frac{\partial f(\alpha_0 + \tau v)}{\partial \tau} \right|_{\tau=0}$$

for $v \in \mathcal{V}$. We assume that $\frac{\partial f(\alpha_0)}{\partial \alpha}[\cdot]$ is linear functional on \mathcal{V} . Since \mathcal{V}_n is a finite-dimensional

⁸A functional is irregular if it is not \sqrt{n} -estimable.

Hilbert space under $\|\cdot\|$, there exists $v_n^* \in \mathcal{V}_n$ such that

$$\frac{\partial f(\alpha_0)}{\partial \alpha}[v] = \langle v_n^*, v \rangle$$

for all $v \in \mathcal{V}_n$ and that

$$\frac{\partial f(\alpha_0)}{\partial \alpha}[v_n^*] = \|v_n^*\|^2 = \sup_{v \in \mathcal{V}_n: \|v\| \neq 0} \frac{\left| \frac{\partial f(\alpha_0)}{\partial \alpha}[v] \right|^2}{\|v\|^2} < \infty$$

by the Riesz Representation Theorem. v_n^* is called the sieve Riesz representer of the linear functional $\frac{\partial f(\alpha_0)}{\partial \alpha}[\cdot]$. For any $v \in \mathcal{V}$, define

$$\|v\|_{sd} \equiv \sqrt{\text{Var}(\Delta(W, \alpha_0)[v])}$$

as a pseudo-norm, provided it is finite. The scaled sieve Riesz representer for functional $f(\cdot)$ is defined as $u_n^* \equiv \frac{v_n^*}{\|v_n^*\|_{sd}}$.

For $\delta_{2,n}^* \equiv \max \left\{ \sqrt{\frac{k_n}{n}}, \|\pi_n \alpha_0 - \alpha_0\|_2 \right\}$ and $\delta_{\infty,n}^* \equiv \max \left\{ \xi_n \sqrt{\frac{k_n}{n}}, \|\pi_n \alpha_0 - \alpha_0\|_\infty \right\}$, let $\delta_{2,n} \equiv \delta_{2,n}^* \cdot \gamma_n$ and $\delta_{\infty,n} \equiv \delta_{\infty,n}^* \cdot \gamma_n$, where $\gamma_n = \log(\log n)$. We assume that $\delta_{\infty,n} = o(1)$. Define shrinking neighborhoods of α_0 as follows: $\mathcal{N}_0 \equiv \{\alpha \in \mathcal{A} : \|\alpha - \alpha_0\|_2 \leq \delta_n, \|\alpha - \alpha_0\|_\infty \leq \delta_{\infty,n}\}$ and $\mathcal{N}_n \equiv \mathcal{N}_0 \cap \mathcal{A}_n$.

Assumption 7.6. *The following conditions hold:*

- (i) $\sup_{\alpha \in \mathcal{N}_n} \frac{\left| f(\alpha) - f(\alpha_0) - \frac{\partial f(\alpha_0)}{\partial \alpha}[\alpha - \alpha_0] \right|}{\|v_n^*\|} = o(n^{-1/2})$;
- (ii) either (a) or (b) holds:
 - (a) $\|v_n^*\| \nearrow \infty$ and $\frac{\left| \frac{\partial f(\alpha_0)}{\partial \alpha}[\alpha_{0,n} - \alpha_0] \right|}{\|v_n^*\|} = o(n^{-1/2})$;
 - (b) $\|v_n^*\| \nearrow \|v^*\| < \infty$ and $\|v^* - v_n^*\| \times \|\alpha_{0,n} - \alpha_0\| = o(n^{-1/2})$.

Assumption 7.7. *The copula function is twice pathwise continuously differentiable, and all second-order partial derivatives with respect to its arguments and dependence parameters are uniformly bounded.*

Assumption 7.8. *The second-order partial derivatives of the copula function is Hölder continuous with exponent $\kappa \geq 1$ uniformly over \mathcal{N}_n with respect to the supremum norm and $\delta_{\infty,n}^\kappa \cdot \delta_{2,n}^2 = o(n^{-1})$.*

Assumption 7.9. *There exists $\kappa_2 > 0$ such that $\lim_{n \rightarrow \infty} n^{-\kappa_2/2} E \left[\left| \Delta(W, \alpha_0)[u_n^*] \right|^{2+\kappa_2} \right] \rightarrow 0$.*

Assumption 7.6 imposes conditions on the functional of interest, $f(\cdot)$. Assumption 7.6(i) requires that the functional is well approximated by $\frac{\partial f(\alpha_0)}{\partial \alpha}[\cdot]$. When f is a linear functional, there is no need to verify this condition. To see this, consider, for given $y_1 \in \{0, 1\}$, $d \in \{0, 1\}^2$, functional $f(\alpha) = \int_{\mathcal{X}} \frac{\partial \mu_2(y_1, d, x)}{\partial x} w(x) dx$, where $w(\cdot)$ is a nonnegative weighting function over \mathcal{X} such that $\int_{\mathcal{X}} w(x) dx = 1$. Under a set of regularity conditions (e.g., Newey (1997)), one can show that $f(\alpha) = - \int \mu_2(y_1, d, x) \frac{\partial w(x)}{\partial x} dx$ and that $\left| f(\alpha) - f(\alpha_0) - \frac{\partial f(\alpha_0)}{\partial \alpha}[\alpha - \alpha_0] \right| = 0$. When f is a nonlinear functional, one can verify Assumption 7.6(i) by using the convergence rate of the sieve estimator (Chen et al. (2014)). Assumption 7.6(ii) restricts the bias part of the pathwise derivative of f that is reflected by $\alpha_{0,n} - \alpha_0$. It is worth pointing out that Assumption 7.6(ii) allows the functional of interest to be either regular or irregular. When the functional is irregular, then $\|v_n^*\| \nearrow \infty$. There are several functionals of practical interest, and one example is $f(\alpha) = \mu_2(y_1, d, x)$ evaluated at some $y_1 \in \{0, 1\}$, $d \in \{0, 1\}^2$, $x \in \mathcal{X}$. Condition (b) of Assumption 7.6(ii) considers regular functionals (i.e., $\|v_n^*\| \nearrow \|v^*\| < \infty$) and is identical to Assumption 4 of Chen et al. (2006) and Condition 4.1(iii) of Chen (2007).

Assumption 7.7 imposes a smoothness condition on the copula function. It is usually satisfied with various copula functions, including the Gaussian copula. Assumption 7.8 is similar to Assumptions 5 and 6 in Chen et al. (2006). We need to control the second-order terms in the Taylor expansion of $L_n(\cdot)$ and impose some condition on the rate of k_n . Specifically, when the convergence rates of the sieve estimator are those in Theorem 7.2, one can choose k_n such that $\xi_n^2 \frac{k_n^2}{n} = o(1)$ and $\sqrt{n} \|\pi_n \alpha_0 - \alpha_0\|_\infty^2 = o(1)$. When α_0 belongs to a Hölder ball and we use spline sieve spaces, then we have $\xi_n = O(\sqrt{k_n})$ and $\|\pi_n \alpha_0 - \alpha_0\|_\infty = k_n^{-\gamma}$ for some $\gamma > 0$ that depends on the dimension and smoothness of the nonparametric function. Therefore, the latter condition in Assumption 7.8 holds if $k_n^3/n =$

$o(1)$ and $\sqrt{n}k_n^{-2\gamma} = o(1)$.

Assumption 7.9 is a sufficient condition for the Lyapounov's central limit theorem. When this assumption holds, we have $\frac{1}{\sqrt{n}} \sum_{i=1}^n (\Delta(W_i, \alpha_0)[u_n^*] - E[\Delta(W_i, \alpha_0)[u_n^*]]) \xrightarrow{d} N(0, 1)$. Chen and Liao (2014) and Chen et al. (2014) show that $\sqrt{n} \frac{f(\hat{\alpha}_n) - f(\alpha_0)}{\|v_n^*\|_{sd}}$ is identical to that empirical process up to $o_p(1)$ term under a set of conditions implied by the assumptions in this paper. Therefore, we need Assumption 7.9 to establish the asymptotic normality of the sieve plug-in estimator of the functional.

Based on these assumptions, next theorems establish limiting distributions for the functional of the sieve plug-in estimator (Theorem 7.3) and the sieve LR test statistic (Theorem 7.4).

Theorem 7.3. *Suppose that Assumptions 2.2–2.5 and 7.1–7.5 hold. If Assumptions 7.6 – 7.9 are also satisfied, then,*

$$\sqrt{n} \frac{f(\hat{\alpha}_n) - f(\alpha_0)}{\|v_n^*\|_{sd}} \xrightarrow{d} N(0, 1).$$

Remark 7.2. *When Assumption 7.6(ii) holds with (b), the functional is regular (i.e., \sqrt{n} -estimable). An example of this functional is the (unconditional) average dynamic treatment effects. The plug-in estimator of $f(\alpha_0)$, $f(\hat{\alpha}_n)$, may be semiparametrically efficient in this case, based on the result in Chen et al. (2006).*

We consider testing $H_0 : f(\alpha_0) = 0$ and define the constrained sieve ML estimator $\tilde{\alpha}_n$ defined as

$$\tilde{\alpha}_n \equiv \arg \max_{\{\alpha \in \mathcal{A}_n : f(\alpha) = 0\}} L_n(\alpha).$$

Theorem 7.4. *Suppose that the identification conditions and Assumptions 7.1–7.5 hold. If Assumptions 7.6 – 7.9 are also satisfied and $\|\tilde{\alpha}_n - \alpha_0\|_2 = O_p(\delta_{2,n}^*)$, then, under $H_0 : f(\alpha_0) = 0$,*

$$2n[L_n(\hat{\alpha}_n) - L_n(\tilde{\alpha}_n)] \xrightarrow{d} \chi^2(1).$$

A Proofs for the Sections on Identification (Sections 4–6)

A.1 Preliminary Lemmas and Their Proofs

For $m \geq 2$, let $H : \mathcal{W}^m \rightarrow [0, 1]$ be a m -variate continuously differentiable function that is equal to either a m -variate CDF or “semi-survival” functions. For example, $H : \mathcal{W}^2 \rightarrow [0, 1]$ represents either $F(w_1, w_2)$, $\bar{F}^1(w_1, w_2) \equiv F(w_1) - F(w_1, w_2)$, or $\bar{F}^2(w_1, w_2) \equiv F(w_2) - F(w_1, w_2)$, and $H : \mathcal{W}^3 \rightarrow [0, 1]$ represents either one of the following:

$$F(w_1, w_2, w_3) \tag{A.1}$$

$$F(w_1, w_2) - F(w_1, w_2, w_3) \tag{A.2}$$

$$F(w_2, w_3) - F(w_1, w_2, w_3) \tag{A.3}$$

$$F(w_1, w_3) - F(w_1, w_2, w_3) \tag{A.4}$$

$$F(w_2) - F(w_1, w_2) - \{F(w_2, w_3) - F(w_1, w_2, w_3)\} \tag{A.5}$$

$$F(w_3) - F(w_2, w_3) - \{F(w_1, w_3) - F(w_1, w_2, w_3)\} \tag{A.6}$$

$$F(w_1) - F(w_1, w_3) - \{F(w_1, w_2) - F(w_1, w_2, w_3)\} \tag{A.7}$$

The functions relevant to our identification analysis are $H : [0, 1]^m \rightarrow [0, 1]$ that are either m -variate copulas or “semi-survival” copulas. Specifically, when $m = 2$, $H(u_1, u_2)$ is one of the following:

$$C(u_1, u_2), \tag{A.8}$$

$$u_2 - C(u_1, u_2), \tag{A.9}$$

and when $m = 3$, $H(u_1, u_2, u_3)$ is either one of the following:

$$C(u_1, u_2, u_3), \tag{A.10}$$

$$C(u_1, u_2) - C(u_1, u_2, u_3), \tag{A.11}$$

$$C(u_2, u_3) - C(u_1, u_2, u_3), \tag{A.12}$$

$$u_2 - C(u_1, u_2) - \{C(u_2, u_3) - C(u_1, u_2, u_3)\}. \tag{A.13}$$

We introduce a notion of stochastic ordering that is closely related to \prec_S -ordering introduced in Definition 2.1. We state the definition for $m = 2$ but a similar definition can be introduced for any multivariate distributions with $m \geq 3$ by focusing on a pair of arguments of the distribution.

Definition A.1 (Strictly More SI in “Joint Distribution”). *We say \tilde{H} is strictly more stochastically increasing in “joint distribution” than H if $w_2^*(w_1, w_2) \equiv \tilde{H}^{-1}(w_1, H(w_1, w_2))$ is strictly increasing in w_1 , which is denoted as $H(\cdot, \cdot) \prec_{SJ} \tilde{H}(\cdot, \cdot)$.*

The stochastic ordering is defined between two joint distributions or semi-survival functions rather than conditional distributions of Definition 2.1. In the next lemma, “ \prec_{S^*} ” refers to the stochastic ordering where the root is either strictly increasing or decreasing.

Lemma A.1. *Under Assumption 2.4(ii), for any $H(w_1, w_2, w_3; \rho_{13}, \rho_{23}, \rho_{13})$ of $F(w_1, w_2, w_3; \rho_{13}, \rho_{23}, \rho_{13})$ in (A.2)–(A.7), there exists an associated copula $C_H(\cdot, \cdot; \rho)$ of $C(\cdot, \cdot; \rho)$ (including itself) such that $C_H(\cdot; \rho) \prec_{S^*} C_H(\cdot; \tilde{\rho})$ for $\rho < \tilde{\rho}$ and*

$$H(w_1, w_2, w_3; \rho_{13}, \rho_{23}, \rho_{13}) = \int_{I_H(w_3)} C_H(H(w_1|\tilde{w}_3), H(w_2|\tilde{w}_3); \rho(\rho_{13}, \rho_{23}, \rho_{13})) dF(\tilde{w}_3)$$

for some interval $I_H(w_3) \subseteq \mathcal{W}$ that depends on w_3 and some $H(w_1|w_3)$ and $H(w_2|w_3)$, each of which is either a conditional CDF or survival function and

$$H(w_1, w_2, w_3, w_4; \Sigma) = \int_{I_H(w_3, w_4)} C_H(H(w_1|\tilde{w}_3, \tilde{w}_4), H(w_2|\tilde{w}_3, \tilde{w}_4); \rho(\Sigma)) dF(\tilde{w}_3, \tilde{w}_4)$$

for some interval $I_H(w_3, w_4) \subseteq \mathcal{W}$ that depends on (w_3, w_4) and some $H(w_1|w_3, w_4)$ and $H(w_2|w_3, w_4)$, each of which is either a conditional CDF or survival function.

An associated copula is introduced in Joe (1997, p. 15). In the lemma, we include the original copula as an example of associated copulas. Other examples can be found in the proof of this lemma.

Proof. Let $\mathcal{W} = \mathbb{R}$. Consider (A.2):

$$\begin{aligned} H(w_1, w_2, w_3) &\equiv F(w_1, w_2) - F(w_1, w_2, w_3) \\ &= \int_{-\infty}^{\infty} C(F(w_1|\tilde{w}_3), F(w_2|\tilde{w}_3); \rho(\rho_{13}, \rho_{23}, \rho_{13})) dF(\tilde{w}_3) \\ &\quad - \int_{-\infty}^{w_3} C(F(w_1|\tilde{w}_3), F(w_2|\tilde{w}_3); \rho(\rho_{13}, \rho_{23}, \rho_{13})) dF(\tilde{w}_3) \\ &= \int_{w_3}^{\infty} C(F(w_1|\tilde{w}_3), F(w_2|\tilde{w}_3); \rho(\rho_{13}, \rho_{23}, \rho_{13})) dF(\tilde{w}_3), \end{aligned}$$

where the second equality is by Assumption 2.4(ii) with $w_3 \rightarrow \infty$.

Next, consider (A.3):

$$\begin{aligned} H(w_1, w_2, w_3) &\equiv F(w_2, w_3) - F(w_1, w_2, w_3) \\ &= \int_{-\infty}^{w_3} F(w_2|\tilde{w}_3) dF(\tilde{w}_3) \\ &\quad - \int_{-\infty}^{w_3} C(F(w_1|\tilde{w}_3), F(w_2|\tilde{w}_3); \rho(\rho_{13}, \rho_{23}, \rho_{13})) dF(\tilde{w}_3) \\ &= \int_{-\infty}^{w_3} \bar{C}(H(w_1|\tilde{w}_3), F(w_2|\tilde{w}_3); \rho(\rho_{13}, \rho_{23}, \rho_{13})) dF(\tilde{w}_3), \end{aligned}$$

where the second equality is by Assumption 2.4(ii) and

$$\bar{C}(u, v) = v - C(1 - u, v)$$

is an associated copula of $C(\cdot, \cdot)$ and $H(w_1|w_3) \equiv 1 - F(w_1|w_3)$ is the survival function. Note that $\bar{C}(u|v) = 1 - C(1 - u|v)$. Let $\bar{C}(\psi(u, v)|v; \tilde{\rho}) = \bar{C}(u|v; \rho)$ ($\tilde{\rho} > \rho$) or equivalently, $C(1 - \psi(u, v)|v; \tilde{\rho}) = C(1 - u|v; \rho)$. Then, $1 - \psi(u, v)$ is strictly increasing in v as $C(1 - u|v; \rho)$

satisfies \prec_S -ordering in ρ , or $\psi(u, v)$ is strictly decreasing in v .

Symmetrically, consider (A.4):

$$\begin{aligned}
H(w_1, w_2, w_3) &\equiv F(w_1, w_3) - F(w_1, w_2, w_3) \\
&= \int^{w_3} F(w_1|\tilde{w}_3) dF(\tilde{w}_3) \\
&\quad - \int^{w_3} C(F(w_1|\tilde{w}_3), F(w_2|\tilde{w}_3); \rho(\rho_{13}, \rho_{23}, \rho_{13})) dF(\tilde{w}_3) \\
&= \int^{w_3} \bar{C}(F(w_1|\tilde{w}_3), H(w_2|\tilde{w}_3); \rho(\rho_{13}, \rho_{23}, \rho_{13})) dF(\tilde{w}_3),
\end{aligned}$$

where the second equality is by Assumption 2.4(ii) and

$$\bar{\bar{C}}(u, v) = u - C(u, 1 - v)$$

is another associated copula of $C(\cdot, \cdot)$ and $H(w_2|w_3) \equiv 1 - F(w_2|w_3)$ is the survival function. Note that $\bar{\bar{C}}(u|v) = C(u|1 - v)$. Let $\bar{C}(\psi(u, v)|v; \tilde{\rho}) = \bar{\bar{C}}(u|v; \rho)$ ($\tilde{\rho} > \rho$) or equivalently, $C(\psi(u, v)|1 - v; \tilde{\rho}) = C(u|1 - v; \rho)$. Then, $\psi(u, v)$ is strictly increasing in $1 - v$ as $C(u|1 - v; \rho)$ satisfies \prec_S -ordering in ρ , or $\psi(u, v)$ is strictly decreasing in v .

Finally, consider (A.5)

$$\begin{aligned}
H(w_1, w_2, w_3) &\equiv F(w_2) - F(w_1, w_2) - \{F(w_2, w_3) - F(w_1, w_2, w_3)\} \\
&= \int^{w_3} F(w_2|\tilde{w}_3) dF(\tilde{w}_3) - \int^{w_3} C(F(w_1|\tilde{w}_3), F(w_2|\tilde{w}_3); \rho(\rho_{13}, \rho_{23}, \rho_{13})) dF(\tilde{w}_3) \\
&\quad - \int^{w_3} \bar{C}(H(w_1|\tilde{w}_3), F(w_2|\tilde{w}_3); \rho(\rho_{13}, \rho_{23}, \rho_{13})) dF(\tilde{w}_3) \\
&= \int^{w_3} \bar{C}(H(w_1|\tilde{w}_3), F(w_2|\tilde{w}_3); \rho(\rho_{13}, \rho_{23}, \rho_{13})) dF(\tilde{w}_3) \\
&\quad - \int^{w_3} \bar{C}(H(w_1|\tilde{w}_3), F(w_2|\tilde{w}_3); \rho(\rho_{13}, \rho_{23}, \rho_{13})) dF(\tilde{w}_3) \\
&= \int_{w_3}^{\infty} \bar{C}(H(w_1|\tilde{w}_3), F(w_2|\tilde{w}_3); \rho(\rho_{13}, \rho_{23}, \rho_{13})) dF(\tilde{w}_3).
\end{aligned}$$

The remaining (A.6)–(A.7) are symmetric to (A.5), so omitted. We omit the proof for the 4-variate CDFs. \square

To prove Lemma 4.1, we state a necessary lemma and prove it first. The following condition is implied by Assumption 2.5.

Assumption 2.4*. *The copula $C(v_1, v_2, u_1, u_2; \Sigma)$ in Assumption 2.4 and its semi-survival functions and all their margins (i.e., (A.8)–(A.13)) satisfy the following conditions:*

- (i) $H(v_1, u_1; \rho_{v_1 u_1}) \prec_{SJ} H(v_1, u_1; \tilde{\rho}_{v_1 u_1})$ for any $\rho_{v_1 u_1} < \tilde{\rho}_{v_1 u_1}$;
- (ii) $H(v_1, v_2, u_1; \rho_{v_1 v_2}, \rho_{v_1 u_1}, \rho_{v_2 u_1}) \prec_{SJ} H(v_1, v_2, u_1; \tilde{\rho}_{v_1 v_2}, \rho_{v_1 u_1}, \rho_{v_2 u_1})$ for any $\rho_{v_1 v_2} < \tilde{\rho}_{v_1 v_2}$;
- (iii) $H(v_1, v_2, u_1, u_2; \Sigma) \prec_{SJ} H(v_1, v_2, u_1, u_2; \tilde{\Sigma})$ for any $\rho_{v_2 u_2} < \tilde{\rho}_{v_2 u_2}$ where $\rho_{v_2 u_2}$ and $\tilde{\rho}_{v_2 u_2}$ belong to Σ and $\tilde{\Sigma}$, respectively.

Lemma A.2. *Assumption 2.5 implies Assumption 2.4*.*

Proof. The sufficiency of Assumption 2.5(i) for Assumption 2.4*(i) is shown in Lemma A.1 of Han and Vytlačil (2017). We prove the rest of the claims. Note that Lemma A.1 of Han and Vytlačil (2017) cannot be extended to multivariate cases and thus we need a different proof strategy.

We prove the sufficiency of Assumption 2.5(ii)–(iii) for Assumption 2.4*(ii)–(iii). First, by Assumption 2.5(ii) and Lemma 0.1, any semi-survival function of $C(u_1, u_2, u_3; \rho_{12}, \rho_{13}, \rho_{23})$ (i.e., each of (A.9)–(A.13)) satisfies

$$H(u_1, u_2, u_3; \rho_{12}, \rho_{13}, \rho_{23}) = \int_{I_H(u_3)} C_H(H(u_1|u), H(u_2|u); \rho(\rho_{12}, \rho_{13}, \rho_{23})) du$$

for some interval $I_H(u_3) \subseteq [0, 1]$ and $H(u_1|u_3)$ and $H(u_2|u_3)$, each of which is either a conditional copula or a survival function.

Henceforth in the proof, we suppress ρ_{23} and ρ_{13} for simplicity. For given u_3 and $\tilde{\rho}_{12} > \rho_{12}$ being the dependence parameters for (u_1, u_2) , consider

$$H(u_1^*, u_2, u_3; \tilde{\rho}_{12}) = H(u_1, u_2, u_3; \rho_{12}), \quad (\text{A.14})$$

where $u_1^* \equiv u_1^*(u_1, u_2, u_3; \tilde{\rho}_{12}, \rho_{12})$ is the root of the equation. To prove Assumption 2.4*(ii), we want to show that u_1^* is strictly increasing in u_2 . Under Assumption 2.5(ii), (A.14) can

be expressed as

$$\int_{I_H(u_3)} C_H (H_{1|3}(u_1^*|u), H_{2|3}(u_2|u); \tilde{\rho}) du = \int_{I_H(u_3)} C_H (H_{1|3}(u_1|u), H_{2|3}(u_2|u); \rho) du, \quad (\text{A.15})$$

where the outer copula $C_H(\cdot, \cdot)$ satisfies the \prec_{S^*} -ordering and, under Assumption 2.5(iii), $\tilde{\rho} \equiv \rho(\tilde{\rho}_{12})$ and $\rho \equiv \rho(\rho_{12})$ are strictly increasing in ρ_{12} and thus $\tilde{\rho} > \rho$. Using the notion of functional differentiation (Gelfand et al. (2000)), differentiating (A.15) w.r.t. $H_{2|3}(u_2|\bar{u})$ for fixed \bar{u} yields

$$C_{H,2} (H_{1|3}(u_1^*|\bar{u}), H_{2|3}(u_2|\bar{u}); \tilde{\rho}) = C_{H,2} (H_{1|3}(u_1|\bar{u}), H_{2|3}(u_2|\bar{u}); \rho)$$

or equivalently,

$$C_H (H_{1|3}(u_1^*|\bar{u}) | H_{2|3}(u_2|\bar{u}); \tilde{\rho}) = C_H (H_{1|3}(u_1|\bar{u}) | H_{2|3}(u_2|\bar{u}); \rho). \quad (\text{A.16})$$

The functional derivatives are well-defined because the copulas are continuously differentiable and they are bounded functions on the compact support of $[0, 1]$. Let

$$C_H (u_1^\dagger | H_{2|3}(u_2|\bar{u}); \tilde{\rho}) = C_H (H_{1|3}(u_1|\bar{u}) | H_{2|3}(u_2|\bar{u}); \rho), \quad (\text{A.17})$$

where $u_1^\dagger \equiv u_1^\dagger(H_{1|3}(u_1|\bar{u}), H_{2|3}(u_2|\bar{u}); \tilde{\rho}, \rho)$ is the root of the equation. This root is strictly monotonic in $H_{2|3}$ because $\tilde{\rho} > \rho$ and the outer conditional copula $C_H(\cdot | u_2; \rho)$ satisfies the \prec_{S^*} -ordering as mentioned above. Since the outer conditional copula $C_H(\cdot | u_2; \rho)$ is strictly increasing for any ρ , from (A.16) and (A.17), we then have

$$H_{1|3}(u_1^*|\bar{u}) = u_1^\dagger(H_{1|3}(u_1|\bar{u}), H_{2|3}(u_2|\bar{u}); \tilde{\rho}, \rho).$$

By differentiating both sides w.r.t. u_2 yields

$$\begin{aligned} h_{13}(u_1^*, \bar{u}) \frac{\partial u_1^*}{\partial u_2} &= \frac{\partial u_1^\dagger}{\partial H_{2|3}} \cdot \frac{\partial H_{2|3}(u_2|\bar{u})}{\partial u_2} \\ &= \frac{\partial u_1^\dagger}{\partial H_{2|3}} \cdot h_{23}(u_2, \bar{u}), \end{aligned}$$

where h_{ij} is the cross derivative of $H(u_i, u_j)$. In the case of $H(u_1, u_2, u_3) = C(u_1, u_2, u_3)$, h_{ij} is the copula density with $H(u_i|u_j) = C(u_i|u_j)$. In this case, $\frac{\partial u_1^\dagger}{\partial H_{2|3}} > 0$ and h_{ij} 's are non-negative, and therefore we have $\frac{\partial u_1^*}{\partial u_2} > 0$. In other cases, h_{13} and h_{23} has opposite signs and $\frac{\partial u_1^\dagger}{\partial H_{2|3}} < 0$ (e.g., in the case of (A.3), $H_{1|3}(u_1|u_3) = 1 - C(u_1|u_3)$ and thus $h_{13} = -c_{13}$, while $H_{2|3} = C_{2|3}$ and thus $h_{23} = c_{23}$), and therefore again, we have $\frac{\partial u_1^*}{\partial u_2} > 0$. This proves the claim for trivariate copulas. We omit the proof for 4-variate copulas. \square

A.2 Proof of Lemma 4.1

We prove the case where H is the original copula. Let $\rho' < \rho''$ and $v_1^* \equiv v_1^*(v_1, v_2, u_1; \rho'', \rho', \rho_{v_1 u_1})$ be the root of

$$H(v_1^*, v_2, u_1; \rho'', \rho_{v_1 u_1}) = H(v_1, v_2, u_1; \rho', \rho_{v_1 u_1}). \quad (\text{A.18})$$

Note that $\frac{\partial v_1^*}{\partial v_2} > 0$ by Assumption 2.4*(ii), which in turn holds by Assumption 2.5 and Lemma A.2. For notational simplicity, we henceforth drop the argument $\rho_{v_1 u_1}$ from v_1^* and the copulas. Differentiating (A.18) w.r.t. v_2 yields

$$H_1(v_1^*, v_2, u_1; \rho'') \frac{\partial v_1^*}{\partial v_2} + H_2(v_1^*, v_2, u_1; \rho'') = H_2(v_1, v_2, u_1; \rho').$$

Therefore, $\frac{\partial v_1^*}{\partial v_2} > 0$ is equivalent to that

$$H_2(v_1, v_2, u_1; \rho') - H_2(v_1^*, v_2, u_1; \rho'') > 0, \quad (\text{A.19})$$

because $C_1(v_1^*, v_2, u_1; \rho'') = C(v_2, u_1 | v_1^*; \rho'') > 0$. From equation (A.18), $v_1^* = v_1^*(v_1, v_2, u_1; \rho'', \rho') \rightarrow v_1$ as $\rho' \rightarrow \rho''$ (while $\rho' < \rho''$). Let $v_1^*(\rho) \equiv v_1^*(v_1, v_2, u_1; \rho, \rho')$. Then, (A.19) is also equivalent to

$$\frac{\partial}{\partial \rho} C_2(v_1^*(\rho), v_2, u_1; \rho) < 0,$$

or equivalently

$$C_{21}(v_1^*(\rho), v_2, u_1; \rho) \frac{\partial v_1^*(\rho)}{\partial \rho} + C_{2\rho}(v_1^*(\rho), v_2, u_1; \rho) < 0. \quad (\text{A.20})$$

Also, by differentiating (A.18) w.r.t. ρ'' and letting $\rho'' = \rho$,

$$\frac{\partial v_1^*(\rho)}{\partial \rho} = -\frac{C_\rho(v_1^*(\rho), v_2, u_1; \rho)}{C_1(v_1^*(\rho), v_2, u_1; \rho)}. \quad (\text{A.21})$$

By combining (A.20) and (A.21), we have

$$C_{2\rho}(v_1^*(\rho), v_2, u_1; \rho) C_1(v_1^*(\rho), v_2, u_1; \rho) < C_{21}(v_1^*(\rho), v_2, u_1; \rho) C_\rho(v_1^*, v_2, u_1; \rho). \quad (\text{A.22})$$

Finally, note that

$$\frac{\partial}{\partial v_2} \left(\frac{C_\rho(v_1, v_2, u_1; \rho)}{C_1(v_1, v_2, u_1; \rho)} \right) = \frac{C_{2\rho}(v_1, v_2, u_1; \rho) C_1(v_1, v_2, u_1; \rho) - C_\rho(v_1, v_2, u_1; \rho) C_{21}(v_1, v_2, u_1; \rho)}{C_1(v_1, v_2, u_1; \rho)^2},$$

which is negative by (A.22). This completes the proof when H is the original copula. The other cases can be symmetrically shown. The remaining proof with 4-variate copula is analogous so omitted.

B Proofs for the Section on Estimation (Section 7)

B.1 Proof of Theorem 7.1

Proof. We verify the sufficient conditions of Proposition B.1 in Han and Lee (2019). Condition (i) of Proposition B.1 is satisfied under the identification conditions and Assumption 7.2 with $Q_0(\alpha) = L_0(\alpha) \equiv E[L_n(\alpha)]$. Conditions (ii), (iii), and (iv) are satisfied by the same logic of the proof of Theorem 4.1 in Han and Lee (2019). We thus turn to verifying condition (v) in Proposition B.1. Let $\delta > 0$. For any $\alpha, \tilde{\alpha} \in \mathcal{A}_n$ such that $\|\alpha - \tilde{\alpha}\|_\infty \leq \delta$, we have, by the mean value theorem, Theorem 2.10.7 in Nelsen (2006), and Assumption 7.4,

$$|l(W, \alpha) - l(W, \tilde{\alpha})| \lesssim U(W) \|\alpha - \tilde{\alpha}\|_\infty \leq U(W) \delta, \quad (\text{B.1})$$

where $E[U(W)^2] < \infty$. Therefore, the second condition of Condition 3.5M in Chen (2007) is satisfied with $s = 1$. Finally, by Lemma 2.5 in van de Geer (2000), we have

$$\log N(\delta, \mathcal{A}_n, \|\cdot\|_\infty) = k_n \log \left(1 + \frac{C}{\delta} \right)$$

for some finite $C > 0$. By Assumption 7.3, $\log N(\delta, \mathcal{A}_n, \|\cdot\|_\infty) = o(n)$. In all, condition (v) of Proposition B.1 in Han and Lee (2019) is met, and thus, we have $\|\hat{\alpha}_n - \alpha_0\|_\infty = o_p(1)$. \square

B.2 Proof of Theorem 7.2

Proof. We verify the conditions of Theorem 3.2 in Chen (2007). For any $\alpha \in \mathcal{A}_n(\epsilon)$, we have

$$\begin{aligned} \text{Var}(l(W_i, \alpha) - l(W_i, \alpha_0)) &\leq E[(l(W_i, \alpha) - l(W_i, \alpha_0))^2] \\ &\lesssim \|\alpha - \alpha_0\|_2^2 \leq C\epsilon^2 \end{aligned}$$

for some $C > 0$ under the imposed assumptions. Therefore, Condition 3.7 in Chen (2007) is satisfied. Since $\|\alpha - \alpha_0\|_2 \leq \|\alpha - \alpha_0\|_\infty$, Condition 3.8 in Chen (2007) is also met with

$s = 1$ by equation (B.1). Lastly, we follow the proof of Theorem 4.2 in Han and Lee (2019) to calculate the bracketing number, and obtain that

$$\|\hat{\alpha}_n - \alpha_0\|_2 = O_p \left(\max \left\{ \sqrt{\frac{k_n}{n}}, \|\pi_n \alpha_0 - \alpha_0\|_2 \right\} \right)$$

by applying Theorem 3.2 in Chen (2007).

The proof of the convergence rate with respect to $\|\cdot\|_\infty$ relies on equation (2.4) in Chen and Liao (2014). Specifically, we have

$$\begin{aligned} \|\hat{\alpha}_n - \alpha_0\|_\infty &\leq \|\hat{\alpha}_n - \pi_n \alpha_0\|_\infty + \|\pi_n \alpha_0 - \alpha_0\|_\infty \\ &\leq \frac{\|\hat{\alpha}_n - \pi_n \alpha_0\|_\infty}{\|\hat{\alpha}_n - \pi_n \alpha_0\|_2} \cdot \|\hat{\alpha}_n - \pi_n \alpha_0\|_2 + \|\pi_n \alpha_0 - \alpha_0\|_\infty \\ &\leq \sup_{\{\alpha \in \mathcal{A}_n : \|\alpha - \pi_n \alpha_0\|_2 \neq 0\}} \frac{\|\alpha - \pi_n \alpha_0\|_\infty}{\|\alpha - \pi_n \alpha_0\|_2} \cdot O_p \left(\sqrt{\frac{k_n}{n}} \right) + \|\pi_n \alpha_0 - \alpha_0\|_\infty \\ &\lesssim \sup_{\{\beta \in \mathbb{R}^{k_n} : p^{k_n}(x)' \beta \in \mathcal{A}_n, \beta \neq \beta_{k_n}\}} \frac{\|p^{k_n}(x)\|_E \cdot \|\beta - \beta_{k_n}\|_E}{\|\beta - \beta_{k_n}\|_E} O_p \left(\sqrt{\frac{k_n}{n}} \right) + \|\pi_n \alpha_0 - \alpha_0\|_\infty \\ &\lesssim \xi_n \cdot O_p \left(\sqrt{\frac{k_n}{n}} \right) + \|\pi_n \alpha_0 - \alpha_0\|_\infty = O_p \left(\max \left\{ \xi_n \sqrt{\frac{k_n}{n}}, \|\pi_n \alpha_0 - \alpha_0\|_\infty \right\} \right). \end{aligned}$$

□

B.3 Proof of Theorem 7.3

Let $\mu_n(g(W)) \equiv \frac{1}{n} \sum_i^n (g(W_i) - E[g(Z_i)])$ be the centered empirical process indexed by function g . We also define $r(W, \alpha)[v_1, v_2] \equiv \lim_{\tau \rightarrow 0} \frac{\Delta(W, \alpha + \tau v_2)[v_1] - \Delta(W, \alpha)[v_1]}{\tau}$ for given $v_1, v_2 \in \mathcal{V}$ and $\alpha \in \mathcal{A}$.

Lemma B.1. *Under the conditions imposed in Theorem 7.3, Assumption 2.2 (ii) in Chen and Liao (2014) is satisfied.*

Proof. We verify Assumption 2.2 (ii)' in Chen and Liao (2014) by using Lemma 4.2 in Chen

(2007). Note that for any $\alpha, \tilde{\alpha} \in \mathcal{N}_n$, by Assumption 7.7 and the Cauchy-Schwarz inequality,

$$\begin{aligned} \left| \Delta(W, \alpha)[u_n^*] - \Delta(W, \tilde{\alpha})[u_n^*] \right|^2 &= \left| r(W, \bar{\alpha})[\alpha - \tilde{\alpha}, u_n^*] \right|^2 \\ &\leq C \cdot \langle \alpha - \tilde{\alpha}, u_n^* \rangle_E^2 \\ &\leq C \cdot \|\alpha - \tilde{\alpha}\|_E^2 \cdot \|u_n^*\|_E^2 \end{aligned}$$

for some $C > 0$, where $\bar{\alpha}$ lies between α and $\tilde{\alpha}$ w.p.a.1. Therefore,

$$\sup_{\alpha, \tilde{\alpha} \in \mathcal{N}_n} \left| \Delta(W, \alpha)[u_n^*] - \Delta(W, \tilde{\alpha})[u_n^*] \right|^2 \lesssim \delta_{\infty, n}^2 \cdot \|u_n^*\|_E^2,$$

which implies that

$$E \left[\sup_{\alpha, \tilde{\alpha} \in \mathcal{N}_n} \left| \Delta(W, \alpha)[u_n^*] - \Delta(W, \tilde{\alpha})[u_n^*] \right|^2 \right] \lesssim \delta_{\infty, n}^2$$

by that $\|u_n^*\|$ is bounded and that $\|u_n^*\|_2 \lesssim \|u_n^*\|$. Also, it is easy to show that

$$\int_0^\infty \sqrt{\log N(\epsilon, \mathcal{A}_n, \|\cdot\|_2)} d\epsilon < \infty$$

by Lemma 2.5 in van de Geer (2000). As a result, all conditions for Lemma 4.2 in Chen (2007) are met, and thus,

$$\sup_{\alpha \in \mathcal{N}_n} \mu_n(\Delta(W, \alpha)[u_n^*] - \Delta(Z, \alpha_0)[u_n^*]) = o_p(n^{-1/2})$$

by Lemma 4.2 in Chen (2007), which means that Assumption 2.2 (ii)' is satisfied. \square

Lemma B.2. *Under the conditions imposed in Theorem 7.3, Assumption 2.2 (iii) in Chen and Liao (2014) is satisfied.*

Proof. We verify Assumption 2.2(iii)" in Chen and Liao (2014). By the Taylor expansion,

we have

$$\begin{aligned}
& E[l(W, \alpha_0) - l(W, \alpha)] \\
&= E \left[- \left(\Delta(W, \alpha_0)[\alpha - \alpha_0] + \frac{1}{2} r(W, \tilde{\alpha})[\alpha - \alpha_0, \alpha - \alpha_0] \right) \right] \\
&= -\frac{1}{2} E[r(W, \alpha_0)[\alpha - \alpha_0, \alpha - \alpha_0]] + \frac{1}{2} E[r(W, \alpha_0)[\alpha - \alpha_0, \alpha - \alpha_0] - r(W, \tilde{\alpha})[\alpha - \alpha_0, \alpha - \alpha_0]] \\
&= \frac{1}{2} \|\alpha - \alpha_0\|^2 + \frac{1}{2} E[r(W, \alpha_0)[\alpha - \alpha_0, \alpha - \alpha_0] - r(W, \tilde{\alpha})[\alpha - \alpha_0, \alpha - \alpha_0]],
\end{aligned}$$

where $\tilde{\alpha}$ lies between α and α_0 . It follows that

$$\begin{aligned}
& \sup_{\alpha \in \mathcal{N}_n} \left| E[l(W, \alpha_0) - l(W, \alpha)] - \frac{\|\alpha - \alpha_0\|^2}{2} \right| \\
&= \sup_{\alpha \in \mathcal{N}_n} \left| \frac{1}{2} E[r(W, \alpha_0)[\alpha - \alpha_0, \alpha - \alpha_0] - r(W, \tilde{\alpha})[\alpha - \alpha_0, \alpha - \alpha_0]] \right| \\
&\leq \sup_{\alpha \in \mathcal{N}_n} \frac{1}{2} E \left[\left| r(W, \alpha_0)[\alpha - \alpha_0, \alpha - \alpha_0] - r(W, \tilde{\alpha})[\alpha - \alpha_0, \alpha - \alpha_0] \right| \right] \\
&\lesssim \delta_{2,n}^2 \cdot E \left[\sup_{\alpha \in \mathcal{N}_n} \left\| \frac{\partial^2 C(\alpha)}{\partial \alpha \partial \alpha'} - \frac{\partial^2 C(\alpha_0)}{\partial \alpha \partial \alpha'} \right\| \right] \lesssim \delta_{2,n}^2 \delta_{\infty,n}^\gamma = o(n^{-1})
\end{aligned}$$

by Assumption 7.8. Therefore, Assumption 2.2(iii)" in Chen and Liao (2014) is satisfied, which implies that Assumption 2.2(iii) in Chen and Liao (2014) holds. \square

Proof of Theorem 7.3

Proof. Assumption 7.6 is the same as Assumption 2.1 in Chen and Liao (2014). By Lemmas B.1 and B.2, Assumption 2.2 in Chen and Liao (2014) is satisfied. Assumption 7.9 is a sufficient condition for Lyapounov's central limit theorem, and thus, Assumption 2.3 in Chen and Liao (2014) is also met. By Lemma 2.1 in Chen and Liao (2014), we have

$$\sqrt{n} \frac{f(\hat{\alpha}_n) - f(\alpha_0)}{\|v_n^*\|_{sd}} \xrightarrow{d} N(0, 1).$$

\square

B.4 Proof of Theorem 7.4

Let $\alpha^*(\alpha) \equiv \alpha \pm \langle u_n^*, \hat{\alpha}_n - \alpha_0 \rangle u_n^*$ and $H(\alpha)$ be the matrix of the second-order partial derivatives of the copula function with respect to its arguments and dependence parameters, evaluated at α .

Lemma B.3. *Under the conditions imposed in Theorem 7.4, Assumption 4.1(ii)-(a) in Chen and Liao (2014) is satisfied.*

Proof. We use Theorem 2.14.2 in Van der Vaart and Wellner (1996) to verify Assumption 4.1(ii)-(a) in Chen and Liao (2014). Define

$$\mathcal{G}_n \equiv \{ \pm \langle \hat{\alpha}_n - \alpha_0, u_n^* \rangle \cdot (\Delta(W, \alpha)[u_n^*] - \Delta(W, \alpha_0)[u_n^*]) : \alpha \in \mathcal{N}_n \}.$$

Assumption 4.1(ii)-(a) in Chen and Liao (2014) is implied if $\sup_{g \in \mathcal{G}_n} \mu_n(g) = o_p(n^{-1})$. Note that $g \in \mathcal{G}_n$, we have

$$\begin{aligned} |g| &\leq \left| \langle \hat{\alpha}_n - \alpha_0, u_n^* \rangle \right| \cdot \left| (\Delta(W, \alpha)[u_n^*] - \Delta(W, \alpha_0)[u_n^*]) \right| \\ &\leq \|\hat{\alpha}_n - \alpha_0\| \cdot \|u_n^*\| \cdot \left| (\Delta(W, \alpha)[u_n^*] - \Delta(W, \alpha_0)[u_n^*]) \right| \\ &= \|\hat{\alpha}_n - \alpha_0\| \cdot \left| r(W, \tilde{\alpha})[\alpha - \alpha_0, u_n^*] \right| \\ &\leq C \cdot \|\hat{\alpha}_n - \alpha_0\| \cdot \|\alpha - \alpha_0\|_E \cdot \|u_n^*\|_E \\ &\leq C \cdot \|\hat{\alpha}_n - \alpha_0\| \cdot \|\alpha - \alpha_0\|_\infty \\ &\leq C \cdot \|\hat{\alpha}_n - \alpha_0\| \cdot \delta_{\infty, n} \equiv G_n \end{aligned}$$

where the second inequality holds by the Cauchy-Schwarz inequality, the third line holds for some $\tilde{\alpha}$ between α and α_0 by the Taylor expansion, the fourth line holds by Assumption 7.7, and the last line holds by that $\alpha \in \mathcal{N}_n$. Therefore, G_n is an envelope for \mathcal{G}_n .

We calculate the bracketing integral of \mathcal{G}_n , $J_n(1, \mathcal{G}_n, \|\cdot\|_2)$.⁹ Since the class of functions,

⁹ $J_n(\delta, \mathcal{G}_n, \|\cdot\|_2) \equiv \int_0^\delta \sqrt{1 + \log N_{[]}(\epsilon \|G_n\|_2, \mathcal{G}_n, \|\cdot\|_2)} d\epsilon$

\mathcal{G}_n , is Lipschitz in α , we have

$$N_{[]}(\epsilon \|G_n\|_2, \mathcal{G}_n, \|\cdot\|_2) \leq N(\epsilon/2, \mathcal{N}_n, \|\cdot\|_2)$$

by Theorem 2.7.11 in [Van der Vaart and Wellner \(1996\)](#). Applying Lemma 2.5 in [van de Geer \(2000\)](#) results in that

$$\log N(\epsilon/2, \mathcal{N}_n, \|\cdot\|_2) \leq k_n \cdot \log \left(1 + \frac{2 \cdot \delta_{2,n}}{\epsilon} \right).$$

Therefore,

$$\begin{aligned} \sqrt{1 + \log N_{[]}(\epsilon \|G_n\|_2, \mathcal{G}_n, \|\cdot\|_2)} &\lesssim \sqrt{k_n \cdot \log \left(1 + \frac{2 \cdot \delta_{2,n}}{\epsilon} \right)} \\ &\lesssim \sqrt{k_n \cdot \delta_{2,n}} \cdot \epsilon^{-1/2}, \end{aligned}$$

and we have

$$J_n(1, \mathcal{G}_n, \|\cdot\|_2) \lesssim \sqrt{k_n \cdot \delta_{2,n}}.$$

Theorem 2.14.2 in [Van der Vaart and Wellner \(1996\)](#) implies that

$$\begin{aligned} \sup_{\alpha \in \mathcal{N}_n} \mu_n(g(\alpha)) &\lesssim \frac{1}{\sqrt{n}} \sqrt{k_n \cdot \delta_{2,n}} \cdot \|G_n\|_2 \\ &\lesssim \sqrt{\frac{k_n}{n}} \delta_{2,n} \delta_{\infty,n}, \end{aligned}$$

and this is $o(n^{-1})$ under Assumption [7.8](#). Therefore, Assumption 4.1(ii)-(a) in [Chen and Liao \(2014\)](#) is satisfied. \square

Lemma B.4. *Under the conditions imposed in Theorem [7.4](#), Assumption 4.1(ii)-(b) in [Chen and Liao \(2014\)](#) is satisfied.*

Proof. We first note that for any $\alpha \in \mathcal{N}_n$, $\alpha^*(\alpha) \in \mathcal{N}_n$ w.p.a.1. Pick any $\alpha \in \mathcal{N}_n$. Then,

$$\begin{aligned}
\|\alpha^*(\alpha) - \alpha_0\|_2 &\leq \|\alpha - \alpha_0\|_2 + \|\langle \hat{\alpha}_n - \alpha_0, u_n^* \rangle u_n^*\|_2 \\
&\leq \delta_{2,n} + |\langle \hat{\alpha}_n - \alpha_0, u_n^* \rangle| \cdot \|u_n^*\|_2 \\
&\leq \delta_{2,n} + \|\hat{\alpha}_n - \alpha_0\| \cdot \|u_n^*\| \cdot \|u_n^*\|_2 \\
&= \delta_{2,n} + o_p(1).
\end{aligned}$$

We also have

$$\begin{aligned}
&E[l(W, \alpha) - l(W, \alpha^*(\alpha))] \\
&= \frac{1}{2}E[r(W, \tilde{\alpha})[\alpha - \alpha_0, \alpha - \alpha_0]] - \frac{1}{2}E[r(W, \tilde{\alpha}^*)[\alpha^*(\alpha) - \alpha_0, \alpha^*(\alpha) - \alpha_0]] \\
&= -\frac{1}{2}\|\alpha - \alpha_0\|^2 + \frac{1}{2}\|\alpha^*(\alpha) - \alpha_0\|^2 + \frac{1}{2}E[r(W, \tilde{\alpha})[\alpha - \alpha_0, \alpha - \alpha_0] - r(W, \alpha_0)[\alpha - \alpha_0, \alpha - \alpha_0]] \\
&\quad - \frac{1}{2}E[r(W, \tilde{\alpha}^*)[\alpha^*(\alpha) - \alpha_0, \alpha^*(\alpha) - \alpha_0] - r(W, \alpha_0)[\alpha^*(\alpha) - \alpha_0, \alpha^*(\alpha) - \alpha_0]],
\end{aligned}$$

and thus, it is enough to show that

$$\begin{aligned}
&E[r(W, \tilde{\alpha})[\alpha - \alpha_0, \alpha - \alpha_0] - r(W, \alpha_0)[\alpha - \alpha_0, \alpha - \alpha_0]] \\
&\quad - E[r(W, \tilde{\alpha}^*)[\alpha^*(\alpha) - \alpha_0, \alpha^*(\alpha) - \alpha_0] - r(W, \alpha_0)[\alpha^*(\alpha) - \alpha_0, \alpha^*(\alpha) - \alpha_0]] = o(n^{-1}).
\end{aligned}$$

Since we have

$$\begin{aligned}
&\left| r(W, \tilde{\alpha})[\alpha - \alpha_0, \alpha - \alpha_0] - r(W, \alpha_0)[\alpha - \alpha_0, \alpha - \alpha_0] \right| \\
&\leq \|\alpha - \alpha_0\|_E^2 \cdot \left\| H(\tilde{\alpha}) - H(\alpha_0) \right\| \\
&\lesssim \|\alpha - \alpha_0\|_E^2 \cdot \|\tilde{\alpha} - \alpha_0\|_E^\kappa \\
&\lesssim \|\alpha - \alpha_0\|_E^2 \cdot \|\tilde{\alpha} - \alpha_0\|_\infty^\kappa,
\end{aligned}$$

it follows that

$$E[r(W, \tilde{\alpha})[\alpha - \alpha_0, \alpha - \alpha_0] - r(W, \alpha_0)[\alpha - \alpha_0, \alpha - \alpha_0]] \lesssim \delta_{2,n}^2 \cdot \delta_{\infty,n}^\kappa.$$

By Assumption 7.8, $\delta_{2,n}^2 \cdot \delta_{\infty,n}^\kappa = o(n^{-1})$; and therefore, Assumption 4.1 (ii)-(b) in Chen and Liao (2014) holds. \square

Proof of Theorem 7.4

Proof. Since the objective function in (7.1) is a sample log-likelihood function, we have $\|v_n^*\| = \|v_n^*\|_{sd}$ by the information equality. By Lemmas (B.3) and (B.4) and Assumption (7.9), Assumption 4.1 in Chen and Liao (2014) is satisfied. In the proof of Theorem (7.3), we have already shown that under the conditions imposed in Theorem 7.4, Assumption 2.2 in Chen and Liao (2014) is satisfied. By Theorem 4.1 in Chen and Liao (2014), we have $2n[L_n(\hat{\alpha}_n) - L_n(\tilde{\alpha}_n)] \xrightarrow{d} \chi^2(1)$. \square

References

- ABBRING, J. H. AND J. J. HECKMAN (2007): “Econometric evaluation of social programs, part III: Distributional treatment effects, dynamic treatment effects, dynamic discrete choice, and general equilibrium policy evaluation,” *Handbook of Econometrics*, 6B, 5145–5303. 1, 2
- ARELLANO, M. AND B. HONORÉ (2001): “Panel data models: some recent developments,” in *Handbook of Econometrics*, Elsevier, vol. 5, 3229–3296. 2
- BALAT, J. F. AND S. HAN (forthcoming): “Multiple treatments with strategic interaction,” *Journal of Econometrics*. 1
- BONHOMME, S. (2012): “Functional differencing,” *Econometrica*, 80, 1337–1385. 3

- CHEN, X. (2007): “Large sample sieve estimation of semi-nonparametric models,” *Handbook of econometrics*, 6, 5549–5632. [1](#), [7.1](#), [7.2](#), [B.1](#), [B.2](#), [B.3](#)
- CHEN, X., Y. FAN, AND V. TSYRENNIKOV (2006): “Efficient estimation of semiparametric multivariate copula models,” *Journal of the American Statistical Association*, 101, 1228–1240. [7.2](#), [7.2](#)
- CHEN, X. AND Z. LIAO (2014): “Sieve M inference on irregular parameters,” *Journal of Econometrics*, 182, 70–86. [7.2](#), [7.2](#), [B.2](#), [B.1](#), [B.3](#), [B.2](#), [B.3](#), [B.3](#), [B.3](#), [B.4](#), [B.4](#), [B.4](#)
- CHEN, X., Z. LIAO, AND Y. SUN (2014): “Sieve inference on possibly misspecified semi-nonparametric time series models,” *Journal of Econometrics*, 178, 639–658. [7.2](#)
- CHERNOZHUKOV, V. AND C. HANSEN (2005): “An IV model of quantile treatment effects,” *Econometrica*, 73, 245–261. [1](#), [2](#), [2](#)
- CUI, Y. AND E. TCHETGEN TCHETGEN (2021): “A semiparametric instrumental variable approach to optimal treatment regimes under endogeneity,” *Journal of the American Statistical Association*, 116, 162–173. [1](#)
- DARSOW, W. F., B. NGUYEN, E. T. OLSEN, ET AL. (1992): “Copulas and Markov processes,” *Illinois journal of mathematics*, 36, 600–642. [4](#)
- GALE, D. AND H. NIKAIDO (1965): “The Jacobian matrix and global univalence of mappings,” *Mathematische Annalen*, 159, 81–93. [4](#), [7](#)
- GELFAND, I. M., R. A. SILVERMAN, ET AL. (2000): *Calculus of variations*, Courier Corporation. [4](#), [A.1](#)
- HAN, S. (2021): “Identification in nonparametric models for dynamic treatment effects,” *Journal of Econometrics*, 225, 132–147. [1](#)
- (2022): “Optimal Dynamic Treatment Regimes and Partial Welfare Ordering,” *arXiv preprint arXiv:1912.10014*. [1](#)

- HAN, S. AND S. LEE (2019): “Estimation in a generalization of bivariate probit models with dummy endogenous regressors,” *Journal of Applied Econometrics*, 34, 994–1015. [1](#), [B.1](#), [B.1](#), [B.2](#)
- HAN, S. AND E. J. VYTLACIL (2017): “Identification in a generalization of bivariate probit models with dummy endogenous regressors,” *Journal of Econometrics*, 199, 63–73. [1](#), [2](#), [2](#), [4](#), [7](#), [A.1](#)
- HECKMAN, J. AND R. PINTO (2018): “Unordered monotonicity,” *Econometrica*, 86, 1–35. [1](#)
- HECKMAN, J. J. (1981): “Heterogeneity and state dependence,” in *Studies in labor markets*, University of Chicago Press, 91–140. [2](#)
- HECKMAN, J. J., J. E. HUMPHRIES, AND G. VERAMENDI (2016): “Dynamic treatment effects,” *Journal of Econometrics*, 191, 276–292. [1](#)
- HECKMAN, J. J. AND S. NAVARRO (2007): “Dynamic discrete choice and dynamic treatment effects,” *Journal of Econometrics*, 136, 341–396. [1](#)
- HONORÉ, B. E. AND E. KYRIAZIDOU (2000): “Panel data discrete choice models with lagged dependent variables,” *Econometrica*, 68, 839–874. [1](#), [3](#), [4](#)
- HONORÉ, B. E. AND E. TAMER (2006): “Bounds on parameters in panel dynamic discrete choice models,” *Econometrica*, 74, 611–629. [1](#)
- HONORÉ, B. E. AND M. WEIDNER (2021): “Moment conditions for dynamic panel logit models with fixed effects,” *arXiv preprint arXiv:2005.05942*. [3](#), [4](#)
- IMBENS, G. W. AND J. D. ANGRIST (1994): “Identification and Estimation of Local Average Treatment Effects,” *Econometrica*, 62, 467–475. [1](#)
- JOE, H. (1997): *Multivariate models and multivariate dependence concepts*, CRC press. [2](#), [A.1](#)

- (2014): *Dependence modeling with copulas*, CRC press. [2](#)
- KYRIAZIDOU, E. (2001): “Estimation of dynamic panel data sample selection models,” *The Review of Economic Studies*, 68, 543–572. [1](#)
- LEE, S. AND B. SALANIÉ (2018): “Identifying effects of multivalued treatments,” *Econometrica*, 86, 1939–1963. [1](#)
- MURPHY, S. A., M. J. VAN DER LAAN, J. M. ROBINS, AND C. P. P. R. GROUP (2001): “Marginal mean models for dynamic regimes,” *Journal of the American Statistical Association*, 96, 1410–1423. [1](#)
- NELSEN, R. B. (2006): *An introduction to copulas*, Springer Science & Business Media. [B.1](#)
- NEWBY, W. K. (1997): “Convergence rates and asymptotic normality for series estimators,” *Journal of Econometrics*, 79, 147–168. [7.2](#), [7.2](#)
- QIU, H., M. CARONE, E. SADIKOVA, M. PETUKHOVA, R. C. KESSLER, AND A. LUEDTKE (2021): “Optimal individualized decision rules using instrumental variable methods,” *Journal of the American Statistical Association*, 116, 174–191. [1](#)
- SHAIKH, A. M. AND E. J. VYTLACIL (2011): “Partial identification in triangular systems of equations with binary dependent variables,” *Econometrica*, 79, 949–955. [2](#)
- TORGOVITSKY, A. (2019): “Nonparametric inference on state dependence in unemployment,” *Econometrica*, 87, 1475–1505. [1](#)
- VAN DE GEER, S. (2000): *Empirical Processes in M-estimation*, Cambridge university press. [7.2](#), [B.1](#), [B.3](#), [B.4](#)
- VAN DER VAART, A. W. AND J. A. WELLNER (1996): *Weak convergence and empirical processes*, Springer. [B.4](#)
- VYTLACIL, E. AND N. YILDIZ (2007): “Dummy endogenous variables in weakly separable models,” *Econometrica*, 75, 757–779. [2](#)