Estimating a Continuous Treatment Model with Spillovers: A Control Function Approach

Tadao Hoshino

School of Political Science and Economics, Waseda University, Tokyo, Japan

ABSTRACT
We study a continuous treatment effect model in the presence of treatment spillovers through social networks. We assume that one's outcome is affected not only by his/her own treatment but also by a (weighted) average of his/her neighbors' treatments, both of which are treated as endogenous variables. Using a control function approach with appropriate instrumental variables, we show that the conditional mean potential outcome can be nonparametrically identified. We also consider a more empirically tractable semiparametric model and develop a three-step estimation procedure for this model. As an empirical illustration, we investigate the causal effect of the regional unemployment rate on the crime rate.

1. Introduction

Treatment evaluation under cross-unit interference is currently one of the most studied topics in the causal inference literature (see e.g., Halloran and Hudgens 2016; Aronow et al. 2021 for recent reviews). These previous studies have highlighted the importance of accounting for treatment spillovers from other units via empirical applications in many fields, including political science, epidemiology, education, economics, etc. However, most of these studies have focused on the case of simple binary treatments, and estimating treatment spillover effects with continuous treatments has been rarely considered. Even in the context of binary treatments, they often assume that the treatments are (conditionally) randomly assigned with full compliance (with some exceptions including Kang and Imbens 2016; Imai, Jiang, and Malani 2020; DiTraglia et al. 2021; Hoshino and Yanagi 2021a; Vazquez-Bare 2021). To fill these gaps, this article considers the estimation of a continuous treatment model with treatment spillovers while allowing the subjects to self-select their own treatment levels.

For conventional continuous treatment models without spillovers, several approaches have been proposed to estimate the average dose-response function—the expected value of a potential outcome at particular treatment level. Among others, Hirano and Imbens (2004) and Imai and Van Dyk (2004) propose the generalized propensity score (GPS) approach under an unconfoundedness assumption. A further extension of the GPS approach to the case of multiple correlated treatments is discussed in Egger and Von Ehrlich (2013). Then, it would be a natural idea to apply these methods to estimate continuous treatment and spillover effects jointly, as in Del Prete, Forastiere, and Scibolazzia (2020), in which the authors use a GPS method to adjust for confounding variables for both own and others' treatments.

However, the above-mentioned methods are not successful in the presence of unobservable confounding variables, which might lead to endogeneity issues not only with respect to own treatment but also to others' treatments. For example, suppose the outcome is academic performance for high school students, and the treatment of interest is hours spent on in-home private tutoring. Since close school mates are often taught by the same teachers and have similar socioeconomic backgrounds, they share a variety of unobservable attributes that affect the treatment and potential outcomes simultaneously. In this situation, not only one's own treatment level but also the treatments of his/her friends should be treated as endogenous variables.

To deal with this complex endogeneity problem without relying on strong parametric distributional restrictions, we assume a specific type of triangular model in this study, in which the treatment equation has a certain form of separability and others' treatments influence one's outcome in the form of peers' weighted average treatment, possibly with some monotonic transformation. The latter assumption is similar to and more general than the mean interaction in Manski (2013), which assumes that the impacts from others can be summarized as the empirical mean of peers' treatments. Since causal inference is generally impossible if no assumptions are imposed on the interference structure (see, Imbens and Rubin 2015), assumptions similar to the above are widely adopted in the literature.

With these assumptions, we address the endogeneity by employing a control function approach (see, e.g., Blundell and Powell 2003; Florens et al. 2008; Imbens and Newey 2009), which
introduces auxiliary regressors, the **control variables**, in the estimation of the outcome equation to eliminate the endogeneity bias. Under the availability of valid instrumental variables (IVs), the rank variable in the treatment equation can serve as the control variable for own treatment variable, which is a standard result in the literature. Meanwhile, to control the endogeneity for the treatment spillovers, a straightforward choice for the control variables would be to use the peers’ rank variables by analogy. Although this approach is theoretically simple, it may not be practical unless the number of interacting partners is limited to one or two because of the curse of dimensionality. As a novel finding of this article, we demonstrate that the rank variable of a “hypothetical” individual who is on average equivalent to real peers can be used as a valid control variable by using the additive structure and interaction structure of our model. Since this rank variable is one dimensional, we can alleviate the dimensionality problem.

As is well-known, to achieve nonparametric identification of treatment parameters such as the average structural function based on a control function approach, we often require a strong support condition, such that the support of the control variable conditional on the treatment variable is equal to the marginal support of the control variable (see Imbens and Newey 2009). This is true for our model as well, but such condition may be rarely satisfied in reality. Thus, to improve the empirical tractability, we introduce additional functional form restrictions for estimation, similar to Chernozhukov et al. (2020) and Newey tractability, we introduce additional functional form restrictions be rarely satisfied in reality. Thus, to improve the empirical

To summarize, the main contributions of this study are fourfold. First, to our knowledge, this study is the first to address a continuous treatment model in which both treatment spillover and endogeneity are present. Second, we propose a novel control function approach to establish the nonparametric identification of this model under some functional form restrictions. Third, considering the empirical feasibility, we propose a semiparametric multiplicative potential outcome model and develop a three-step estimation procedure for it, which is of independent interest in semiparametric estimation theory. Fourth, by applying the proposed method to Japanese city data, we provide new empirical evidence on the relationship between the local unemployment and crime rates.

**Article organization.** In Section 2, we present our model and discuss nonparametric identification of the CATR parameter. In Section 3, we introduce an empirically tractable semiparametric model and propose our three-step estimation procedure. The asymptotic properties of the estimator are also presented in this section. The empirical analysis on the Japanese crime data is presented in Section 4, and Section 5 concludes. The proofs of technical results, Monte Carlo simulation results, and supplementary information on the empirical analysis are all summarized in the supplementary material.

**Notation.** For natural numbers and denotes an identity matrix, and denotes a matrix of zeros of dimension . For a matrix , denotes the Frobenius norm. If is a square matrix, we use and to denote its largest and smallest eigenvalues, respectively. We use and to denote the Kronecker and Hadamard (element-wise) product, respectively. For a set , denotes its cardinality. We use (possibly with a subscript) to denote a generic positive constant whose value may vary in different contexts. For random variables and means that they are independent. Lastly, for positive sequences and means that there exist such that for sufficiently large .

### 2. Nonparametric Identification

Suppose that we have a sample of agents that form social networks whose connections are represented by an adjacency matrix . These agents can be individuals, firms, or municipalities depending on the context. The networks can be directed, that is, regardless of the value of , we may
observe $A_{ij} = 1$ if $i$ affects $j$ and $A_{ij} = 0$ otherwise. The diagonal elements of $A_N$ are all zero. Throughout the article, we treat $A_N$ as nonrandom. For each $i$, we denote $i$’s “reference group” (peers, colleagues, neighbors, etc) as $P_i := \{1 \leq j \leq N: A_{ij} = 1\}$ and its size as $n_i = |P_i|$. To simplify the discussion, we assume that $n_i > 0$ for all $i$. In addition, we write $\overline{P}_i := i \cup P_i$. For a general variable $Q$, we denote $Q_{\overline{P}_i} := \{Q_j\}_{j \in \overline{P}_i}$. We define $Q_{P_i}$ similarly.

Let $T \in T$ denote the continuous treatment variable of interest. We assume that the support of interest is a real closed interval of $T$. We assume that the support of interest is $Y \in \mathbb{R}$. In this study, we explicitly allow that the peers’ treatments $T_{P_i}$ influence on $i$’s own outcome $Y_i$ via a known real-valued function $S_i$:

$$S_i(T_{P_i}) := m_i \left( \sum_{j \in \overline{P}_i} a_{ij} T_j \right),$$

where $m_i$ is a strictly increasing continuous function which may depend on $i$, and $a_{ij}$’s are weight terms satisfying $a_{ij} > 0$ if and only if $j \in \overline{P}_i$ and $\sum_{j \in \overline{P}_i} a_{ij} = 1$. For example, if $m_i$ is an identity function and $a_{ij} = n_i^{-1}$, then $S_i$ simply returns the reference-group average. This form of treatment spillover is the most commonly used in empirical studies on peer effects. If $m_i(t) = n_i t$ instead, then $S_i$ is the sum of peers’ treatments. For another example, in the spatial statistics literature, researchers often assume that $a_{ij}$ is inversely proportional to the geographical distance between $i$ and $j$. The support of $S_i$ is denoted as $S_i$, which varies across individuals in general because of the transformation $m_i$ that may be specific to $i$. As a special case, if $m_i$ is an identity function, then we have the same support for all $i$ as $S_i = T$.

Now, letting $X_i = (X_{i,j}, \ldots, X_{i,dx,j})^T$ be a $d_x \times 1$ vector of observed individual covariates, we suppose that the outcome $Y_i$ is determined in the following equation:

$$Y_i = y(T_i, S_i, X_i, \epsilon_i) \quad \text{for} \quad i = 1, \ldots, N$$

(2.1)

where $y$ is an unknown function, $S_i = S_i(T_{P_i})$, and $\epsilon_i$ is an unobservable determinant of $Y_i$. Note that $\epsilon_i$ could be a vector of unknown dimension. The non-separable setting is essential for a treatment effect model since it permits general interactions between $(T_i, S_i)$ and $\epsilon_i$, allowing for treatment heterogeneity among observationally identical individuals.2 The potential outcome when $(T_i, S_i) = (t, s) \in T S_i$ is written as $Y_{i}(t, s) = y(t, s, X_i, \epsilon_i)$. The target parameter of interest in this study is the conditional average treatment response (CATR) function:

$$\text{CATR}(t, s) := \mathbb{E}[Y_i(t, s) \mid X_i = x].$$

Next, suppose that we have a $dz \times 1$ vector of IVs, $Z_i = (Z_{i,1}, \ldots, Z_{i,dz})^T$, which determines the value of $T_i$ through the following equation:

$$T_i = \pi(X_i, Z_i) + \eta(U_i),$$

(2.2)

where $\pi$ and $\eta$ are unknown functions, and $U_i$ is a scalar unobservable variable. Here, unlike the outcome equation, we assume that there are no spillover effects from peers’ covariates $(X_{j}, Z_j)$ on $i$’s treatment level $T_i$. We allow that not only $U_i$ but also $U_{P_i}$ are potentially correlated with $\epsilon_i$, which are the sources of endogeneity for $T_i$ and $S_i$, respectively. To address the endogeneity issue, we make the following assumptions:

**Assumption 2.1.** (i) $Z_{\overline{P}_i} \perp \!\!\!\!\perp (U_{\overline{P}_i}, \epsilon_i) \mid X_{\overline{P}_i}$; (ii) the function $u \mapsto \eta(u)$ is continuous and strictly increasing, and $U_i$ is distributed as Uniform$[0, 1]$ conditionally on $X_{\overline{P}_i}$.

**Assumption 2.1(i)** is the exogeneity assumption that own and peers’ IVs are independent of the unobservables given their covariates. This assumption would be particularly plausible in an experimental setup where $Z_i$ is a randomly assigned treatment induction. For **Assumption 2.1(ii)**, we emphasize that the functional form of $\eta$ is independent of $X_{\overline{P}_i}$. Note that although a conditional mean independence suffices to identify $\pi(X_i, Z_i)$ and $\eta(U_i)$,4 imposing full independence between the instrument and the first-stage residual is important to construct a valid control variable for the spillover effect.

Recall that the disturbance term $v_i$ in the outcome equation is allowed to be a vector and to enter the model in a nonadditive way. Thus, we can see that the identification result heavily depends on the functional form of the treatment equation rather than that of the outcome equation, in line with the existing literature. Below, we formally discuss the identification of the CATR parameter. Throughout this section, we use the term “identification” to indicate that the parameter of interest can be characterized through a moment of the observable random variables. Note that this does not automatically imply the estimability of the parameters because network data naturally follow a nonidentical and dependent data distribution.5 To estimate the

---

1Assuming that the functional form of $S_i$ is known a priori is a strong assumption. In the literature, there are several studies that investigate under what conditions we can make meaningful causal inference even when the “exposure” function (i.e., $S_i$) is unknown or mis-specified (e.g., Hoshino and Yanagida 2021a; Säveje, Aronow, and Hudgens 2021; Leung 2022; Vazquez-Bare 2022). All these studies consider only binary treatment situations, and whether we can establish similar results for continuous treatment models would be an important open question.

2In the literature on social interactions, researchers often assume a linear-in-means model specification (e.g., Manski 1993; Bramoullé, Djebarri, and Fortin 2020): in our context, $\gamma(T_i, S_i, X_i, \epsilon_i) = c + T_i \beta + S_i \gamma + X_i \alpha + \epsilon_i$. In this model, the marginal effects of $T_i$ and $S_i$ on the outcome are $\beta$ and $\gamma$ respectively homogeneously across individuals, which should be a strong requirement. Note that in the above literature, it is often the case that endogenous effect, the effect of others’ outcomes on one’s own outcome, is also of interest. Incorporating the endogenous effect in our non-separable framework would extremely complicate the analysis (see, e.g., Lazatti 2015) and thus is left for future research.

3Following the econometrics literature (e.g., Blundell and Powell 2003), this parameter may be called the average structural function (conditional on $x_i = x$). In the causal inference literature, it is also known as the average dose response function.

4When considering a more general treatment equation $T_i = \gamma(X_i, Z_i) + \eta(U_i, X_{\overline{P}_i})$, the second part of **Assumption 2.1(ii)** is always satisfied using $\tilde{U}_i$ and $\tilde{u}_i$ where $\tilde{U}_i = \gamma_{U_i}(X_{\overline{P}_i})$, $\tilde{u}_i(\tilde{U}_i, X_{\overline{P}_i}) = \eta_{\tilde{U}_i}(X_{\overline{P}_i})$. Under **Assumption 2.1(i)**, we have $\text{Pr}(T_i \leq u \mid X_i, Z_i) = \text{Pr}(U_i \leq u \mid X_{\overline{P}_i})$, implying that $\pi(\gamma(X_i, Z_i) + \eta(U_i, X_{\overline{P}_i}) \mid Z_i, X_{\overline{P}_i})$ can be estimated in a quantile regression framework. Furthermore, by checking whether the estimated conditional quantile function varies with the peers’ covariates, we may be able to test the validity of assuming $\eta(U_i, X_{\overline{P}_i}) \equiv \eta(U_i)$; however, directly implementing this would suffer from the curse of dimensionality.

5In this sense, the CATR parameter should be indexed by “$i$,” however, we suppress it for notational simplicity.
CATR parameter, as discussed later, we require some additional stationarity assumptions.

As is well-known, we can deal with the endogeneity of $T_i$ by including $U_i$ as a control variable in the estimation of outcome equation (2.1) (e.g., Blundell and Powell 2003; Floreens et al. 2008; Imbens and Newey 2009). For the identification of $U_i$ in Lemma 1, supplementary materials, we prove that $\pi_i(X_i, Z_i)$ and $\eta(u)$ for any $u \in (0, 1)$ are identifiable up to a location shift (see Remark 2 as well). When these parameters are treated as known, $U_i$ can be identified by $U_i = \eta^{-1}(T_i - \pi(X_i, Z_i))$ by the monotonicity of $\eta$. Equivalently, we can find such $U_i$ by solving $\min_{u \in [0, 1]} |T_i - \pi(X_i, Z_i) - \eta(u)|$.

Similarly, to account for the endogeneity of $S_i$, we need to find a control variable(s) for it. One obviously valid candidate is to use $U_{P_i}$. Once the behaviors of $U_{P_i}$ are controlled, the variation of $S_i$ comes only from $(X_{P_i}, Z_{P_i})$, which are conditionally independent of the disturbances by Assumption 2.1(i). However, this approach is not practical unless $n_i$ is limited to one or two, because of the curse of dimensionality. In the next proposition, we show that under the additive separability in (2.2) we can construct a one-dimensional control function for $S_i$.

**Proposition 1.** Under Assumption 2.1, $V_i = \{v \in [0, 1] \mid \sum_{i \in P_i} a_{ij} \eta(U_j) = \eta(v)\}$ is unique and we have $(T_i, S_i) \perp \epsilon_i \mid U_i, V_i, X_{\overline{P}_i}$.

The identification of $V_i$ can be achieved by $V_i = \eta^{-1}(\sum_{i \in P_i} a_{ij}(T_i - \pi(X_i, Z_i)))$. Proposition 1 implies that

$$\mathbb{E}[Y_i \mid T_i = t, S_i = s, U_i = u, V_i = v, X_{\overline{P}_i} = x] = \mathbb{E}[Y_i(t, s) \mid U_i = u, V_i = v, X_{\overline{P}_i} = x].$$

(2.3)

Since the left-hand side is a moment computed by the observables, the right-hand side is identifiable. Without the additive separability between $(X_i, Z_i)$ and $U_i$ as in (2.2), $V_i$ generally depends on the instruments $Z_{P_i}$, and conditioning on $V_i$ also restricts their behavior (see Kasy 2011), leading to a failure of establishing (2.3).

**Remark 1 (Interpretation of $V_i$).** We can interpret the control variable $V_i$ as the rank variable for a hypothetical friend of $i$ who is observationally equal to the weighted average of $i$’s friends. Note that this interpretation essentially comes from the assumption that $S_i$ is a scalar-valued function. That is, in our model, having multiple friends whose treatments are on average equal to $S_i$ is indistinguishable from having only one friend whose treatment level is exactly $S_i$. Note that simply using $\sum_{i \in P_i} a_{ij} \eta(U_j)$ as a control variable theoretically works, but in general, its support is unbounded. Unbounded control variables are less practical because, as presented below, our estimation procedure involves computing several integrals with respect to the control variables. In addition, a more delicate discussion is necessary to establish desirable convergence results for the functions of the control variables.

Now, we define the marginal treatment response (MTR) function:

$$MTR(t, s, u, v, x) := \mathbb{E}[Y_i(t, s) \mid U_i = u, V_i = v, X_{\overline{P}_i} = x].$$

As shown in (2.3), we can identify $MTR(t, s, u, v, x)$ as the left-hand side of (2.3). Moreover, if $MTR(t, s, u, v, x)$ is identifiable uniformly in $(u, v, x, i) \in \text{supp}(U_i, V_i, X_{\overline{P}_i} \mid X_i = x)$, we can identify $\text{CATR}(t, s, x)$ by

$$\text{CATR}(t, s, x) = \mathbb{E} \left[ \int_0^1 \int_0^1 MTR(t, s, u, v, X_{\overline{P}_i}) | U_i \mid X_{\overline{P}_i}, u, v \mid X_{\overline{P}_i}, dudv \mid X_i = x \right],$$

(2.4)

where the outer expectation is with respect to $X_{\overline{P}_i}$ conditional on $X_i = x$, and $f_{U_i|X_{\overline{P}_i}}$ is the conditional density function of $(U_i, V_i)$ given $X_{\overline{P}_i}$. For $MTR(t, s, u, v, x)$ to be well-defined on the entire $(u, v, x, i) \in \text{supp}(U_i, V_i, X_{\overline{P}_i} \mid X_i = x)$, we need the following additional condition:

**Assumption 2.2.** $\text{supp}(U_i, V_i, X_{\overline{P}_i} \mid T_i = t, S_i = s, X_i = x) = \text{supp}(U_i, V_i, X_{\overline{P}_i} \mid X_i = x)$.

From (2.3), we can see that without any support restriction, $MTR(t, s, u, v, x)$ can be obtained only for $(u, v, x, i) \in \text{supp}(U_i, V_i, X_{\overline{P}_i} \mid T_i = t, S_i = s, X_i = x)$. Assumption 2.2 requires that conditioning on $(T_i = t, S_i = s)$ should not affect the support of $(U_i, V_i, X_{\overline{P}_i})$ conditional on $X_i = x$. An important implication of this condition is that the IV must have very rich variation so that the values of $T_i$ and $S_i$ do not restrict the range of values that $U_i$ and $V_i$ can take. We summarize the result obtained so far in the next theorem.

**Theorem 1.** Suppose that Assumptions 2.1 and 2.2 hold. Then, $\text{CATR}(t, s, x)$ is identified through (2.3)–(2.4).

In order to construct an estimator for $\text{CATR}(t, s, x)$ based on Theorem 1 in practice, we need to cope with several obstacles. The first issue is the full-support condition in Assumption 2.2. Finding IVs ensuring such a support condition would be quite difficult in reality. Even without this support condition, if there are informative upper and lower bounds of $Y_i$, it would be possible to partially identify $\text{CATR}(t, s, x)$, as in Theorem 4 of Imbens and Newey (2009). For another more convenient approach to point-identify the CATR in the absence of Assumption 2.2, in the next section and thereafter, we consider introducing additional functional form restrictions on the outcome equation.

Another issue is that, since the size of the reference group and the weights $a_{ij}$ are not necessarily identical among individuals, the support $S_i$ of $S_i$ and the distribution of $V_i$ should vary with individuals in general. The dimension and distribution of $X_{\overline{P}_i}$ are also obviously heterogeneous among individuals. These heterogeneities are problematic in the estimation stage. To circumvent this problem, we simply restrict our attention to a specific subsample $N'$ such that those in this subsample have the same support $S$ and are homogeneous (in the sense of Assumption 3.0(iii)). The simplest but empirically the most typical example would be $N' = \{1 \leq i \leq N : m_i = m, n_i = c, a_{ij} = c^{-1}1[j \in P_i]\}$ for some monotonically increasing $m$ and an integer $c \geq 1$. The size of this subsample is denoted as $n = |N'|$. In the following, without loss of generality, we re-label the $N$ agents so that the first $n$ individuals belong to this subsample, that is, $N' = \{1, \ldots, n\}$.
3. Semiparametric Estimation and Asymptotic Properties

In this section, we propose our estimation procedure for CATR\((t, s, x)\) in a specific semiparametric model, and investigate its asymptotic properties as \(n\) increases to infinity. To increase \(n\), we consider a sequence of networks \(\{A_N\}\), where \(A_N\) is an \(N \times N\) adjacency matrix. Since each \(A_N\) is assumed to be non-stochastic, our analysis can be interpreted as being conditional on the realization of \(A_N\). In this sense, the model and parameters presented in this study should be viewed as triangular arrays defined along the network sequence.

3.1. A Semiparametric Model and Estimation Procedure

In Newey and Stouli (2021), they proposed the following type of potential outcome model as a baseline (in our notation): \(Y_i(t, s) = g(t, s)^\top \epsilon_i\), where \(g(t, s)\) is a vector of transformations of \((t, s)\); for example, \(g(t, s) = (1, t, s)\) (note that they do not consider models with two treatment variables). A similar parametric assumption was adopted in Chernozhukov et al. (2020) as well. In this study, we generalize their models such that \(g(t, s)\) is a general nonparametric function. In addition, we further extend their work by allowing the treatment effect to vary with the individual characteristics \(X_i\). At the same time, to preserve empirical tractability, we assume that there are no interaction effects among \((X_i, \epsilon_i)\), and that \(\epsilon_i\) is one-dimensional. Consequently, we focus on the following outcome model:

\[
Y_i = X_i^\top \beta(T_i, S_i) + g(T_i, S_i)\epsilon_i \quad \text{for } i = 1, \ldots, n \quad (3.1)
\]

where \(\beta(t, s) = (\beta_1(t, s), \ldots, \beta_d(t, s))^\top\) and \(g(t, s)\) are unknown functions to be estimated.\(^6\) For normalization, we assume that \(X_i\) includes a constant term so that \(E[\epsilon_i] = 0\) holds. We do not additionally restrict the treatment equation (2.2), but we assume that the control variables \(U_i\) and \(V_i\) can be consistently estimated at a certain convergence rate (see Remark 5). For further simplification, we strengthen Assumption 2.1(ii) as follows:

**Assumption 3.0.** (i) \( (Z_{T_i}, X_{T_i}) \parallel (U_{T_i}, \epsilon_i)\); and (ii) \((U_i, V_i)\) are identically distributed and \(E[\epsilon_i \mid U_i, V_i] = \lambda(U_i, V_i)\) for all \(i \in N'\).

Recalling the definition of \(V_i\), the first part of Assumption 3.0(ii) essentially requires that the joint distribution of \(U_{T_i}\) is stationary across all \(i \in N'\) and they adopt the same weighting scheme. For the second part, we do not require the treatment heterogeneity terms to be identically distributed. Since Assumption 3.0 is fundamental to derive our estimation procedure, it will be assumed implicitly throughout the whole subsequent discussion.

Now, the CATR for this model at \((t, s, x)\) is simply given by

\[
\text{CATR}(t, s, x) = x^\top \beta(t, s),
\]

which in fact coincides with the “unconditional” expectation \(E[y(t, s, x, \epsilon_i)]\). Thus, the task of estimating CATR\((t, s, x)\) is greatly simplified to the estimation of \(\beta\), and the support condition in Assumption 2.2 is not necessary to recover the CATR. Further, an analogous argument to (2.3) gives

\[
E[Y_i \mid T_i = t, S_i = u, U_i = u, V_i = v] = x^\top \beta(t, s) + g(t, s)E[\epsilon_i \mid U_i = u, V_i = v] = x^\top \beta(t, s) + g(t, s)\lambda(u, v).
\]

This yields the following semiparametric multiplicative regression model:

\[
Y_i = X_i^\top \beta(T_i, S_i) + g(T_i, S_i)\lambda(U_i, V_i) + \epsilon_i \quad \text{for } i = 1, \ldots, n. \quad (3.2)
\]

By construction, we have \(E[\epsilon_i \mid T_i, S_i, U_i, V_i, X_i] = 0\). It is clear that because of the multiplicative structure in (3.2), in order to identify \(g\) and \(\lambda\) separately, we need some functional form normalizations. First, recall that the condition \(E[\epsilon_i] = 0\) must be maintained. This implies the following location normalization:

\[
E[\epsilon_i] = E[\lambda(U_i, V_i)] = 0. \quad (3.3)
\]

We further need some scale normalization for identification. To this end, we set

\[
\int_{T \times S} g(t, s)\,dt\,ds = 1. \quad (3.4)
\]

Regression models with a multiplicative structure as in (3.2) can be found in the literature in different contexts (e.g., Linton and Nielsen 1995; Zhang and Wang 2015; Hu, Huang, and You 2019; Chen, Smetanina, and Wu 2020). A typical approach to estimating this type of model is to apply some marginal integration (to the estimated nonparametric function or to the data itself in advance of the estimation). However, note that none of these prior studies considered a functional-coefficient specification as in ours. Thus, we need to develop a new estimation procedure for our model, which should be of independent interest in the semiparametric estimation literature. Specifically, we deal with the multiplicative structure by nonparametrically estimating the model ignoring the functional form restriction in the first step, and then splitting the estimated function into two multiplicative components by marginal integration in the second step in a similar manner to Chen, Smetanina, and Wu (2020).\(^7\) The first stage estimation involves a four-dimensional nonparametric regression on \((T_i, S_i, U_i, V_i)\), which often produces unstable estimates under a moderate sample size. Thus, we consider using a penalized regression in this step.

The whole estimation procedure is as follows. Before estimating the regression model in (3.2), we need to consistently estimate the function \(\pi\) and \(\eta\) to obtain consistent estimates of \((U_i, V_i)\). Excluding this preliminary step, our procedure for estimating the CATR consists of three steps. In the first step, we estimate the model in (3.2) ignoring the multiplicative structure.

\(^6\) If we add one more continuous treatment, then \(\beta\)'s and \(g\) will become three-dimensional functions, which are practically difficult to estimate because of the curse of dimensionality. To estimate such a model, we would need to introduce additional functional form restrictions on (3.1).

\(^7\) Although it is possible to estimate the model in a single step by estimating \(g\) and \(\lambda\) separately from the beginning, the resulting method requires solving a high-dimensional non-convex optimization problem, which is computationally challenging, as pointed out in Zhang, Zhong, and Wang (2020). To circumvent this issue, Zhang, Zhong, and Wang (2020) proposed an iterative computational algorithm.
using a penalized series regression approach. In the next step, we obtain estimates of $g$ and $\lambda$ by marginally integrating the non-parametric function obtained in the previous step. Finally, we re-estimate the coefficient functions $\beta$ using a local linear kernel regression. With these two additional steps, we can achieve an oracle property for the estimation of the CATR parameter (see Theorem 4). The proposed estimation procedure is new to the literature but may be viewed as a minor extension of a commonly used two-step series-kernel estimation method, in which Chen, Smetanina, and Wu’s (2020) marginal integration step is inserted between the series and kernel estimation steps.

**Preliminary step: control variables estimation.** Under Assumptions 2.1(ii) and 3.0(i), it holds that $\Pr(T_i \leq \pi(X_i, Z_i) + \eta(u)) | X_i, Z_i) = u$ for any $u \in (0, 1)$. This implies that we can estimate $\pi(X_i, Z_i)$ and $\eta(u)$ using the CQR method: for pre-specified $0 < u_1 < u_2 < \cdots < u_L < 1$, 

$$\widehat{\eta}(u_1), \ldots, \widehat{\eta}(u_L), \widehat{\pi} \\
:= \arg\min_{\eta_1, \ldots, \eta_L, \pi} \sum_{i=1}^n \left( T_i - \pi(X_i, Z_i) - \eta_i \right)$$

subject to some location normalization (e.g., $\eta(0.5) = 0$), where $q_n(x) = x(x - 1)$. Then, we compute the residual, $r_{si} := T_i - \widehat{\pi}(X_i, Z_i)$, for each $i$. Note that $r_{si}$ is an estimator of $\eta(U_i)$, whose convergence rate is governed by that for $\widehat{\pi}(u)$. The estimator for $U_i$ can be obtained by $\widehat{U}_i := \arg\min_{u \in [0, 1]} |r_{si} - \widehat{\eta}(u)|$. Similarly, we can estimate $V_i$ by $\widehat{V}_i := \arg\min_{v \in [0, 1]} \sum_{j \in P_i} a_{ij} r_{sj} - \widehat{\eta}(v)$. In practice, these minimization problems are solved by grid search with a sufficiently large $L$.

**Remark 2 (Identification and estimation of the treatment equation).** Here, we briefly comment on the identification and estimation of the treatment equation. If one considers estimating $\pi$ without explicit functional form assumptions, such a model has been investigated in Kai, Li, and Zou (2010), where they applied a local polynomial smoothing to the CQR problem. As shown in Kai, Li, and Zou (2010), we can estimate $\pi$ with the standard nonparametric convergence rate under mild conditions. However, as we demonstrate later (see Remark 5 as well), the full nonparametric estimation of $\pi$ is generally unacceptable for achieving the desirable asymptotic behavior for our CATR estimator. Thus, in the numerical studies in this article, we use a linear model specification: $\pi(X_i, Z_i) = \gamma X_i + Z_i$, which corresponds to the model originally considered in Zou and Yuan (2008). They showed that the coefficients ($\gamma$, $\gamma_Z$) can be estimated at the $\sqrt{n}$ rate under the standard linear independence condition on $(X_i, Z_i)$. One may consider a semiparametric CQR model as an intermediate case of these two (e.g., Kai, Li, and Zou 2011).

**First step: penalized series estimation.** Let $(p_{T_1}(t), p_{S_1}(s), p_{U_j}(u), p_{V_j}(v)) : j = 1, 2, \ldots$ be the basis functions on $T$, $S$, $[0, 1]$, and $[0, 1]$, respectively. We define $p_T(t) := (p_{T_1}(t), \ldots, p_{T_K}(t))^\top$, $p_S(s) := (p_{S_1}(s), \ldots, p_{S_K}(s))^\top$, and $p_T(s) := p_T(t) \otimes p_S(s)$. Then, we consider series approximating $\beta(t, s)$ by $\beta_T(t, s) \approx p_T(t, s) \otimes \theta_{\beta_T}$ for some $K_T \times 1$ coefficient vector $\theta_{\beta_T}$, for each $l = 1, \ldots, dx$, where $K_T := K_T \times 1$. Similarly, define $p_U(u) := (p_{U_1}(u), \ldots, p_{U_K}(u))^\top$, $p_V(v) := (p_{V_1}(v), \ldots, p_{V_K}(v))^\top$, $p_{U,V}(u, v) := p_U(u) \otimes p_V(v)$, and $K_U := K_U \times 1$. We assume that $K_T$ and $K_U$ increase as $n$ increases at the same speed so that there exists an increasing sequence $\{k_n\}$ satisfying $(K_T, K_U) = (n, k_n)$.

For the estimation of $g(t, s, \lambda(u, v), \lambda^*)$, to impose the location condition (3.3) in the estimation, we would like to normalize the basis function such that $\gamma_U(u, v) \approx p_{U,V}(u, v) - \mathbb{E}[p_{U,V}(U, V)]$ (recall that we have been focusing on a subsample in which $\{(U_i, V_i)\}$ are identically distributed). However, since both $\mathbb{E}[p_{U,V}(U, V)]$ and $(U_i, V_i)$ are unknown, we instead use 

$$\widehat{p}_{U,V}(u, v) := p_{U,V}(u, v) - \frac{1}{n} \sum_{i=1}^n p_{U,V}(\widehat{U}_i, \widehat{V}_i).$$

Then, letting $p(t, s, u, v) := p_{TS}(t, s) \otimes p_{U,V}(u, v)$, we consider approximating $g(t, s, \lambda(u, v), \lambda^*) \approx \lambda_U(t, s, u, v)^\top \theta_{\lambda_U}$ with a $K_{TS} \times 1$ coefficient vector $\theta_{\lambda_U}$, where $K_{TS,U} := K_{TS} \times K_U$. Using these approximations, we have 

$$Y_i \approx p_X(X_i, T_i, S_i)^\top \theta_\beta + \hat{p}(T_i, S_i, \widehat{U}_i, \widehat{V}_i)^\top \theta_{\lambda_U} + \epsilon_i, \quad i = 1, \ldots, n$$

(3.5) where $p_X(X_i, T_i, S_i) = X_i \otimes p_{TS}(T_i, S_i)$, and $\theta_\beta = (\theta_{\beta_T}, \ldots, \theta_{\beta_T})^\top$. Based on this model approximation, we estimate $\theta = (\theta_\beta, \theta_{\lambda_U})^\top$ by the penalized least squares method.

Let $D$ be a $(dxK_T + K_U)$-dimensional positive semidefinite symmetric matrix such that $p_{max}(D) = O(1)$. This matrix serves as a penalization matrix, in which typical elements of $D$ are, for example, the integrated derivatives of the basis functions. Write $\Pi_n := (p_X(X_i, T_i, S_i), \hat{p}(T_i, S_i, \widehat{U}_i, \widehat{V}_i)^\top, \Pi_n = ((\Pi_1, \ldots, \Pi_n), \Pi_n, \ldots, \Pi_n)^\top$ and $\lambda_n = (Y_1, \ldots, Y_n)$. Then, for a sequence of tuning parameters $\{\tau_n\}$ tending to zero at a certain rate as $n \to \infty$, the penalized least squares estimator of $\theta$ is given by 

$$\hat{\nu}_n := \left( \nu_{1, \beta}, \nu_{1, \lambda_U} \right)^\top = \left[ \Pi_n^\top \Pi_n + \tau_n D_n \right]^{-1} \Pi_n^\top \lambda_n$$

(3.6)

The first-stage estimator of $\hat{\beta}_{1, \nu}(t, s) := p_{TS}(t, s) \otimes \hat{\beta}_{1, \nu}(t, s)$ for $l = 1, \ldots, dx$. Similarly, $g(t, s, \lambda(u, v))$ can be estimated as $p(t, s, u, v)^\top \hat{\beta}_{1, \nu}$.

**Second step: marginal integrations.** In the second step, we first recover the $\lambda$ function. Recalling the scale normalization in (3.4), the estimator of $\lambda(u, v)$ can be naturally defined by 

$$\hat{\lambda}_n(u, v) := \int_{T \cap S} \left( \hat{p}(t, s, u, v)^\top \hat{\beta}_{1, \nu} \right) dtds$$

$$= [p_{TS} \otimes \hat{p}_{UV}(u, v)]^\top \hat{\beta}_{1, \nu}$$

(3.7)

where $p_{TS} := \int_{T \cap S} p_{TS}(t, s) dtds$.

For the estimation of $g(t, s)$, note that simply dividing $p(t, s, u, v)^\top \hat{\beta}_{1, \nu}$ by $\hat{\lambda}_n(u, v)$ results in an inefficient estimator and is not well-defined if $\hat{\lambda}_n(u, v)$ is close to zero. To obtain a more efficient estimator using the information at all $(u, v)$ values, we apply a least-squares principle to $p(t, s, u, v)^\top \hat{\beta}_{1, \nu}$ $\approx \hat{\lambda}_n(u, v) v(t, s)$; that is, we define our estimator $\hat{g}_n(t, s)$ as the solution of $\min_{\hat{g}_n(t, s)} \int_0^1 \int_0^1 \left( \hat{p}(t, s, u, v)^\top \hat{\beta}_{1, \nu} - \hat{\lambda}_n(u, v) v(t, s) \right)^2 du dv$. By simple calculation, we can find that the solution has a closed form expression:

$$\hat{g}_n(t, s) := \int_0^1 \int_0^1 \left( \hat{p}(t, s, u, v)^\top \hat{\beta}_{1, \nu} - \hat{\lambda}_n(u, v) v(t, s) \right) du dv$$

$$= \left[ p_{TS}(t, s) \otimes \hat{p}_{UV}(u, v) \right]^\top \hat{\beta}_{1, \nu}$$

(3.8)
where \( \hat{P}_{UV} = \int_0^1 \int_0^1 \tilde{P}_{UV}(u, v)\omega_n(u, v)du dv \) and \( \tilde{\omega}_n(u, v) := \tilde{\lambda}_n(u, v)/\int_0^1 \int_0^1 \tilde{\lambda}_n(u', v')^2 du' dv' \).

**Third step: kernel smoothing.** In the final step, we re-estimate \( \beta \) using a local linear kernel regression. Let \( W \) be a symmetric univariate kernel density function and \( \{(h_T, h_s)\} \) be the sequence of bandwidths tending to zero as \( n \) increases. The bandwidths may depend on the evaluation point \( (t, s, x) \) of the CATR parameter in general (see Remark 4). Further, we define \( X_i(t, s) := (X_i^T, X_i^T(T-T)/h_T, X_i^T(S-s)/h_s)^T \), and \( W_i(t, s) := (h_Th_s)^{-1}W((T_i-T)/h_T)W((S_i-s)/h_s) \) for a given \( (t, s) \in TS \). Then, our final estimator \( \tilde{\beta}_n(t, s) \) of \( \beta(t, s) \) is obtained by

\[
\tilde{\beta}_n(t, s) := S_d \left[ \sum_{i=1}^n X_i(t, s)X_i(t, s)^TW_i(t, s) \right]^{-1} \times \sum_{i=1}^n X_i(t, s)[Y_i - \tilde{\omega}_n(T_i, S_i)\hat{\lambda}_n(\tilde{U}_i, \tilde{V}_i)]W_i(t, s),
\]

(3.9)

where \( S_d := [I_{dx}, 0_{dx \times 2dx}] \), and the estimator of CATR \( r(t, s, x) \) is given by \( \hat{\text{CATR}}_n(t, s, x) = x^T\tilde{\beta}_n(t, s) \).

### 3.2. Asymptotic Properties

To investigate the asymptotic properties of the proposed estimators, we first introduce assumptions on the dependence structure underlying the data. Typical applications of treatment spillover models would be data from school classes, households, working places, neighborhoods, municipalities, etc., which are usually subject to some local dependence. To account for such dependency, we assume that all \( N \) agents are located in a (latent) \( d \)-dimensional space \( D \subseteq \mathbb{R}^d \) for some \( d < \infty \). For spatial data applications, the space \( D \) is typically defined by a geographical space with \( d = 2 \). For nonspatial data, it is possible that \( D \) is a space of general social and demographic characteristics (or a mixture of geographic space and such spaces), and in this case we should view it as an embedding of individuals rather than their actual locations. The set of observation locations, which we denote by \( D_N \subset D \), may differ across different \( N \). With a little abuse of notation, for each \( i \), we use the same \( i \) to denote his/her location. Let \( \Delta(i, j) \) denote the Euclidean distance between \( i \) and \( j \).

**Assumption 3.1.** For all \( i, j \in D_N \) such that \( i \neq j \), (i) \( \Delta(i, j) \geq 1 \) (without loss of generality); and (ii) there exists a threshold distance \( \Delta \in \mathbb{Z} \) satisfying \( A_{ij} = 0 \) if \( \Delta(i, j) > \Delta \).

Assumption 3.1(i) rules out the infill asymptotics (Cressie 1993). Assumption 3.1(ii) is a “homophily” assumption in the space \( D \). This may not be too restrictive since the definition of \( D \) can be freely modified depending on the context. This framework would accommodate many empirically relevant situations. For example, clustered sample data with finite cluster size can be seen as a special case of it. An important implication from these two assumptions is that the size of reference group is uniformly bounded above by some constant of order \( \Delta^2 \). Thus, we do not allow for the existence of “dominant units,” which may have increasingly many interacting partners as the network grows. Assuming \( \Delta \) to be an integer is only to simplify the proof.

**Assumption 3.2.** (i) \( \{(X_i, Z_i, \epsilon_i, U_i) : i \in D_N, N \geq 1\} \) is an \( \alpha \)-mixing random field with mixing coefficient \( c(k, l, r) \leq (k + l)^\alpha \sigma(r) \) for some constant \( 0 \leq \alpha < \infty \) and function \( \sigma(r) \) that satisfies \( \sum_{r=1}^\infty (r + 2\Delta)^{d-1} \sigma(r) < \infty \) and (ii) \( \sup_{i \in D_N} \|X_i\| < \infty \).

For the precise definition of the \( \alpha \)-mixing random field and the mixing coefficient in this context, see Definition 1 in Appendix A (see also Jenish and Prucha 2009). Combined with Assumption 3.1(iii), Assumption 3.2(i) implies that \( \{S_i\}, \{V_i\} \), and \( \{\tilde{\epsilon}_i\} \) are also \( \alpha \)-mixing processes. For Assumption 3.3, the second assumption restricts the dependence structure in the treatment heterogeneity, and the third requires that all potential correlations among \( \epsilon \)'s are through the correlations of the other individual characteristics. Note that \( \tilde{\epsilon}_i \) can be written as \( \tilde{\epsilon}_i = g(T_i, S_i)[\epsilon_i - \lambda(U_i, V_i)] \). Thus, with this assumption we impose that the \( \tilde{\epsilon}_i \)'s have the fourth order moments and that they are independent of each other conditional on the individual characteristics.

**Assumption 3.3.** (i) For all \( j, (P_{TJ}(t), P_{SJ}(s), P_{UV}(u), P_{VV}(v)) \) are uniformly bounded on \( T, S, [0, 1], \) and \( [0, 1] \), respectively; (ii) \( P_{UV}(u) \) and \( P_{VV}(v) \) are differentiable such that \( \sup_{u \in [0, 1]} |\partial P_{UV}(u)/\partial u| \leq \psi_u \) and \( \sup_{v \in [0, 1]} |\partial P_{VV}(v)/\partial v| \leq \psi_v \); (iii) \( \|P_{TS}^\delta\| = O(1) \) and \( \|P_{VV}^\delta\| = O(1) \), where

\[
P_{UV} = \int_0^1 \int_0^1 P_{UV}(u, v)\omega(u, v)du dv \quad \text{with} \quad \omega(u, v) := \frac{\lambda(u, v)}{\int_0^1 \int_0^1 \lambda(u', v')^2 du' dv'};
\]

(iv) there exist positive constants \( (c, \xi) \) such that \( \|P_{UV}(u, v) - P_{UV}(u', v')\| \leq c\xi \|u, v - u', v'\| \) for any \( (u, v), (u', v') \in [0, 1]^2 \) and (v) there exist positive constants \( (c, \xi) \) such that \( \|P_{TS}(t, s) - P_{TS}(t', s')\| \leq c\xi \|t, s - t', s'\| \) for any \( (t, s), (t', s') \in TS \).

**Assumption 3.5.** Define \( \Pi_n \) analogously to \( \hat{\Pi}_n \) by replacing the estimates of \( (U_i, V_i, E[P_{UV}(U, V)]) \) with their true values. There exist \( c_1, c_2 \) such that \( \epsilon_i < \rho_{min}(E(\Pi_n^T \Pi_n))/\rho_{max}(E(\Pi_n^T \Pi_n)) \leq c_2 \) uniformly in \( (K_{TS}, K_{VV}) \) for sufficiently large \( n \).

Assumption 3.4(ii) is introduced for analytical simplicity, which imposes some restrictions on the choice of basis functions. For example, B-spline basis and Fourier series satisfy this assumption. The uniform boundedness implies that \( \sup_{T \in T} \|P_{T}(t)\| = O(\sqrt{K_T}) \), and the similar result applies to \( P_S \).
Assumption 3.6. For all $l = 1, \ldots, dx$, $\beta_l$ is twice continuously differentiable, and there exists a vector $\theta_{\beta_l}^*$ and a positive constant $\mu_l$ such that $\sup_{(t,s) \in TS} |\beta_l(t,s) - P_{TS}(t,s)^\top \theta_{\beta_l}^*| = O(K_{TS}^{-\mu_l})$. Similarly, $(g, \lambda)$ are continuously differentiable, and there exist vectors $\theta_{g,\lambda}^*$ and positive constants $(\mu_g, \mu_\lambda)$ such that $\sup_{(t,s) \in TS} |g(t,s) - P_{TS}(t,s)^\top \theta_{g,\lambda}^*| = O(K_{TS}^{-\mu_g})$ and $\sup_{(u,v) \in [0,1]^2} |k(u,v) - \bar{P}_{UV}(u,v)^\top \theta_{g,\lambda}^*| = O(K_{UV}^{-\mu_\lambda})$.

The constants $(\mu_l, \mu_g, \mu_\lambda)$ generally depend on the choice of the basis function and the dimension and the smoothness of the function to be approximated. For example, if the target function belongs to a $k$-dimensional Holder class with smoothness $\rho$, it typically holds that $\mu = \rho/k$ for splines, wavelets, etc (Chen 2007). Here, define $\theta_{\beta_l, \beta_l}^* = \theta_{\beta_l}^* \otimes \theta_{\beta_l}^*$ so that $\bar{P}(t,s,u,v)^\top \theta_{\beta_l, \beta_l}^* = (P_{TS}(t,s)^\top \theta_{\beta_l}^*) \cdot (\bar{P}_{UV}(u,v)^\top \theta_{g,\lambda}^*)$ holds, where $\bar{P}(t,s,u,v) = P_{TS}(t,s) \otimes \bar{P}_{UV}(u,v)$. Then, we can easily see that

$$\sup_{(t,s,u,v) \in TS \times [0,1]^2} |g(t,s)\lambda(u,v) - \bar{P}(t,s,u,v)^\top \theta_{\beta_l, \beta_l}^*|$$

$$= \sup_{(t,s,u,v) \in TS \times [0,1]^2} |g(t,s)\lambda(u,v) - (P_{TS}(t,s)^\top \theta_{\beta_l}^*) \cdot (\bar{P}_{UV}(u,v)^\top \theta_{g,\lambda}^*)| = O(K_{TS}^{-\mu_g} + K_{UV}^{-\mu_\lambda}).$$

Assumption 3.7. There exists a constant $\nu \in (0,1/2]$ such that $(\sup_{u \in D_u} \bar{U}_1 - U_1, \sup_{v \in D_v} \bar{V}_1 - V_1) = O_p(n^{-\nu}).$

Assumption 3.8. As $n \to \infty$, (i) $\kappa_u^4/n \to 0$ and $\zeta_4 \sqrt{\kappa_u n^{-\nu}} \to 0$, where $\zeta_4 = \zeta U\sqrt{K_{UV}} + \zeta V\sqrt{K_{U1}}$; and (ii) $\kappa_4^2 \ln \kappa_u/n \to 0$.

For Assumption 3.7, if we adopt a parametric model specification for the treatment equation in (2.2), then the assumption holds with $\nu = 1/2$. As discussed in Remark 5, we require $\nu$ to be at least larger than $1/3$ under optimal bandwidths. Assumption 3.8(i) is used to establish a matrix law of large numbers. Recalling that $(K_{TS}, K_{UV}) \approx \kappa_u n$, the first part of the assumption requires that $K_{TS}$ and $K_{UV}$ must grow slower than $n^{3/4}$. It is easy to see that $\zeta_4$ gives the order of $||\bar{P}_{UV}(u,v)/\partial u||$ and $||\bar{P}_{UV}(u,v)/\partial v||$. For example, when one uses a tensor product B-splines, it can be shown that $\zeta_4 = O(\kappa_u)$ (see, e.g., Hoshino and Yanagi 2021b). Condition (ii) implies the first part of (i). This assumption can be relaxed if we can strengthen Assumption 3.3 so that the error terms $[\tilde{e}_i]$ have the moments of order higher than four.

Now, in the following theorem, we derive the convergence rate for the first-stage series estimator.

**Theorem 2.** Suppose that Assumptions 3.1–3.3, 3.4(i), (ii), 3.5–3.7, and 3.8(i) hold. Then, we have

(i) $||\hat{\beta}_{n,\beta} - \theta_{\beta}^*|| = O_p \left( \frac{\kappa_u}{\sqrt{n}} + b_\mu + \tau_n^* + n^{-\nu} \right)$

(ii) $||\hat{\beta}_{n,g,\lambda} - \theta_{g,\lambda}^*|| = O_p \left( \frac{\kappa_u}{\sqrt{n}} + b_\mu + \tau_n^* + n^{-\nu} \right)$

where $b_\mu := K_{TS}^{-\mu_\mu} + K_{UV}^{-\mu_\lambda}$ and $\tau_n^* := \tau_n \sqrt{\theta^* \cdot D \theta^*}$.

The above results should be standard in the literature. In particular, result (i) indicates that we can estimate CATR $(t,s,x)$ consistently by $x^\top \hat{\theta}_{n,\beta}$, although less efficiently when compared with our final estimator. The magnitude of $\tau_n^*$ depends not only on the penalty matrix $D$ but also on the choice of the basis function. Under $\rho_{\max}(D) = O(1)$, it is clear that the upper bound for $\tau_n^*$ is $O(\kappa_0 \kappa_u)$. If the coefficients are decaying in the order of series, $\tau_n^* = O(\kappa_0)$ would hold, as in Han (2020).

For later use, we introduce the following miscellaneous assumptions.

**Assumption 3.9.** (i) $\inf_{(u,v) \in [0,1]^2} \lambda_{UV}(u,v) > 0$, where $\lambda_{UV}$ is the joint density of $(U, V)$; (ii) there exists $c$ such that $\rho_{\max}(E[\bar{P}_{UV}(U, V) \bar{P}_{UV}(U, V)^\top]) < c$ uniformly in $(K_{U}, K_{V})$; and (iii) $\kappa_0 \kappa_u n^{-\nu} = O(1)$ and $\kappa_4^2 (b_\mu + \tau_n^* + n^{-\nu}) = O(1)$.

In the next theorem, we establish the convergence rates for the second-stage estimators.

**Theorem 3.** Suppose that Assumptions 3.1–3.8 hold. Then, we have

(i) $\sup_{(u,v) \in [0,1]^2} |\hat{\lambda}_n(u,v) - \lambda(u,v)| = O_p \left( \frac{\sqrt{\kappa_u \ln \kappa_u}}{n} + \sqrt{\kappa_u} (b_\mu + \tau_n^* + n^{-\nu}) \right)$

(ii) If Assumption 3.9 additionally holds, we have

$$\sup_{(t,s) \in TS} |\hat{g}_{n}(t,s) - g(t,s)| = O_p \left( \frac{\sqrt{\kappa_u \ln \kappa_u}}{n} + \sqrt{\kappa_u} (b_\mu + \tau_n^* + n^{-\nu}) \right).$$

The above theorem clearly shows that the marginal integration procedure successfully resolves the slower convergence of the first-stage estimator. Note however that both estimators $\hat{\lambda}_n$ and $\hat{g}_n$ do not attain the optimal uniform convergence rate of Stone (1982).\footnote{Whether our marginal integration-based estimation method can achieve the optimal rate is left as an open question. For example, for $\hat{\lambda}_n$, in order to achieve the optimal uniform convergence rate, the bias term should be of order $O(K_{UV}^{-\mu_\lambda}/n)$, where $p$ is a smoothness parameter (see the discussion given after Assumption 3.6). For the attainability of the optimal uniform rate for standard linear series regression estimators, see Huang (2003), Belloni et al. (2015), and Chen and Christensen (2015), among others.}

We next derive the limiting distribution of our final estimator for CATR $(t,s,x)$, where $(t,s)$ is a given interior point of $TS$. Let $f_i(t,s)$ be the joint density of $(T_i, S_i)$, and define $\Omega_{1,(t,s)} := \mathbb{E} [X_i | X_i^\top \tau_i = t, S_i = s]$, $\Omega_{2,(t,s)} := \frac{1}{n} \sum_{i=1}^n \Omega_{1,(t,s)} f_i(t,s)$, and $\Omega_{2,(t,s)} := \mathbb{E} [X_i | X_i^\top \tau_i = t, S_i = s]$. Then, Assumption 3.10. (i) For all $i \in D_N$, $f_i(t,s)$ and $\Omega_{1,(t,s)}$ are continuously differentiable and $\Omega_{2,(t,s)}$ is continuous on
\( \mathcal{TS} \); and (ii) \( \Omega_1(t, s) := \lim_{n \to \infty} \Omega_{1,n}(t, s) \) and \( \Omega_2(t, s) := \lim_{n \to \infty} \Omega_{2,n}(t, s) \) exist and are positive definite.

**Assumption 3.11.** The kernel \( W \) is a probability density function that is symmetric and continuous on the support \([-c_W, c_W]\).

**Assumption 3.12.** (i) \( (h_T, h_S) \sim n^{-1/6} \); and (ii) \( \xi_n^{(1/6)} - v \to 0 \) and \( n^{1/3}(\tau_n^* + b_H + n^{-v}) \to 0 \) as \( n \to \infty \).

**Assumption 3.11** is fairly standard. The compact support assumption is used just for simplicity, and it can be dropped at the cost of lengthier proof. For **Assumption 3.12**, it will be later shown that condition (i) is the optimal rate for the bandwidths. We use condition (ii) to ensure that the final CATR estimator becomes oracle efficient.

Let \( I_1(t, s) := 1 \left[ \frac{|t - s|}{h_T^2} \leq c_W, \frac{|s - t|}{h_S^2} \leq c_W \right], I_2(t, s) := \text{diag}\left( I_{1(t, s)}, \ldots, I_{1(t, s)} \right), P_{n,TS} := (P_{TS}(T_1, S_1), \ldots, P_{TS}(T_n, S_n))^\top \), and \( P_{n,UV} := (P_{UV}(U_1, V_1), \ldots, P_{UV}(U_n, V_n))^\top \).

**Assumption 3.13.** There exist constants \((c_{TS}, c_{UV})\) such that \( \max(P_{n,TS}I_1(t, s)P_{n,TS}/n) < c_{TS} \) and \( \max(P_{n,UV}I_1(t, s)P_{n,UV}/n) < c_{UV} \) uniformly in \((K_{TS}, K_{UV}, h_T, h_S)\) for sufficiently large \( n \).

**Assumption 3.14.** (i) \( \sup_{(a,b) \in D_N} \sup_{(t_0,s_0) \in (TS)^2} f_{j_1}(t_0, s_0) \) is the joint density of \((T_0, S_0, T_j, S_j)\); and (ii) \( \sup_{(a,b) \in D_N} \mathbb{E}[\hat{\xi}_j | T_0, S_0, T_j, S_j] < \infty \).

**Assumption 3.15.** (i) \( \sum_{t=1}^{\infty} e^{r - d - 1} \tilde{a}(r)^{1/2} < \infty \) for some \( e > d/2 \); (ii) \( \omega(\bar{C}^N, \infty, r - 2\bar{C}) = O(r^{-d}) \) for some \( d' > d \) and a positive constant \( C \) (see (A.1)); and (iii) \( \sum_{t=1}^{\infty} (t + 2\bar{C})^{d/(d-2)} - 1 \tilde{a}(r) < \infty \) for some \( 3 \leq \ell < 4 \).

Assumptions 3.13, 3.14, and 3.15 are technical requirements to derive the asymptotic normality of our CATR estimator. In particular, Assumption 3.15(ii) and (iii) are introduced to use the central limit theorem for mixing random fields developed by Jenish and Prucha (2009) in our context. Now we are ready to state our main theorem:

**Theorem 4.** Suppose that Assumptions 3.1–3.15 hold. Then, for a given interior point \((t, s) \in TS \) and a finite \( x \) in the support of \( X \), we have

\[
\sqrt{n h_T h_S} \left( \text{CATR}(t, s, x) - \text{CATR}(t, s, x) \right) = \cdots
\]

where \( \phi_i = \int \phi(W(\theta))^i d\theta, \beta_{TT}(t, s) = \partial^2 \beta(t, s)/(\partial t)^2 \), and \( \beta_{SS}(t, s) := \partial^2 \beta(t, s)/(\partial s)^2 \).

The proof of Theorem 4 is straightforward from Lemma 9, supplementary materials, and thus is omitted. In Lemma 8, supplementary materials, we show that the estimation error for our CATR estimator caused by the estimations of \( g \) and \( \lambda \) are of order \( o_p(n h_T h_S)^{-1/2} \). Thus, the asymptotic distribution presented in the theorem is in fact equivalent to that obtained when the estimators \( \hat{g}_n \) and \( \hat{\lambda}_n \) are replaced by their true counterparts; that is, our CATR estimator has an oracle property.

**Remark 3 (Covariance matrix estimation).** For statistical inference, we need to consistently estimate the asymptotic covariance matrix. The matrix \( \Omega_1(t, s) \) can be easily estimated by the kernel method—see Lemma 5, supplementary materials. Similarly, we can estimate \( (\phi_i^2)\tilde{\Omega}_2(t, s) \) using the kernel method with the error terms \( \tilde{e}_i \) being replaced by the residuals. That is, letting \( \tilde{e}_i(t, s) := Y_i - X_i^\top \beta_0(t, s) - \tilde{g}_n(T_i, S_i) \tilde{\lambda}_n(\tilde{U}_i, \tilde{V}_i) \) for \( i = 1, \ldots, n \), we can show that \( \tilde{\Omega}_{2,n}(t, s) := [h_T h_S/n] \sum_{i=1}^n X_i^\top \tilde{e}_i(t, s) X_i W_i(t, s)^{1/2} \) is consistent for \( (\phi_i^2)\tilde{\Omega}_2(t, s) \) with an additional mild assumption—see Lemma 10, supplementary materials.

**Remark 4 (Bandwidth selection).** As a result of Theorem 4, the asymptotic mean squared error (AMSE) of CATR\(n(t, s, x)\) is given by

\[
\text{AMSE}(t, s, x) = \frac{\langle \phi_i^2 \rangle^2}{4} \left[ x^\top \beta_{TT}(t, s) h_T^2 + x^\top \beta_{SS}(t, s) h_S^2 \right]^2 + \frac{\langle \phi_i^2 \rangle^2}{4} + x^\top \beta_{TT}(t, s) \sigma_n(T) + x^\top \beta_{SS}(t, s) \sigma_n(S) \right]^2.
\]

From this, we can derive the optimal bandwidth parameters that minimize the AMSE. In particular, suppose that the bandwidths are given by \( h_T = C(t, s, x) \sigma_n(T) n^{-1/6} \) and \( h_S = C(t, s, x) \sigma_n(S) n^{-1/6} \) for some constant \( C(t, s, x) > 0 \), where \( \sigma_n(T) \) and \( \sigma_n(S) \) are the sample standard deviations of \( T \) and \( S \), respectively. Then, we can obtain \( C(t, s, x) \) that minimizes the AMSE as follows:

\[
C(t, s, x) = \left( \frac{2\langle \phi_i^2 \rangle^2}{4} x^\top \beta_{TT}(t, s) \sigma_n(T) + x^\top \beta_{SS}(t, s) \sigma_n(S)^2 \right) \times \cdots
\]

Although the optimal bandwidth involves several unknown quantities, their approximate values are obtainable using the first-stage series estimates (with or without regularization).

**Remark 5.** The functional form of \( \pi(X_s, Z_s) \) determines the possible range for \( v \). In view of **Assumption 3.12**(ii), \( \nu \) must satisfy \( 1/3 < \nu \). In addition, assuming \( \xi_1 = O(\kappa_n) \), the second part of **Assumption 3.8**(i) is reduced to \( \kappa_n^{-3/2} n^{-v} \to 0 \). Thus, for example when \( \kappa_n = O(n^{1/3}) \) (in view of the first part of **Assumption 3.8**(i)), the condition is further reduced to \( 3/10 < \nu \), being consistent with **Assumption 3.12**(ii). These imply that the treatment model does not have to be fully parametrically specified in general (i.e., \( \nu = 1/2 \)), but at the same time a full nonparametric specification may not be acceptable.

---

4. **Causal Impacts of Unemployment on Crime**

It is often considered that unemployment and crime are endogenously related because of their simultaneity (e.g., Levitt 2001). In economic theory, criminal activity is typically characterized by the balance between the cost and benefit of committing illegal actions.
Table 1. Variables used.

| Variables*        | (Shorthand) |
|-------------------|-------------|
| Y                 | Crime rate: 100 × # crimes / population |
| T                 | Unemployment rate: 100 × unemployment / labor force population |
| X                 | Population density: ln([daytime population / area]) |
|                   | Per-capita retail sales: ln([retail sales / population]) |
|                   | Elderly ratio: population over 65 / whole population |
| Z                 | Ratio of single households: # single households / # total households |
|                   | Availability of childcare facilities: 1000 × # childcare facilities / # total households |

* The variables are all at city level.

activities (Becker 1968). Thus, poor local labor market conditions result in a relative increase in the benefits of crime, increasing the number of crime incidents. On the other hand, crime drives away business owners and customers, exacerbating the working condition. In addition, from the viewpoint of criminals, if their communities are economically deteriorating with high unemployment, they might commit crimes in more “beneficial” neighborhoods. This would suggest the need to account for the spatial spillover effects of unemployment.

There are several prior studies that investigate the relationship between unemployment and crime using some IV-based methods (e.g., Raphael and Winter-Ebmer 2001; Lin 2008; Altindag 2012). For example, using U.S. state-level data, Raphael and Winter-Ebmer (2001) conducted two-stage least squares regression analysis with military spending and oil costs as the IVs for the unemployment. Their estimation results suggest that unemployment is indeed an important determinant of property crime rates. In this article, we present an empirical analysis on the causal effect of local unemployment rate on crime based on Japanese city-level data. Our empirical study aims at extending the earlier works in two ways. Our model is based on a more flexible potential outcome framework and allows for the existence of spillover effects from neighboring regions.

The variables used and their definitions are summarized in Table 1. The outcome variable of interest is the city-level crime rate, which is based on the number of crimes recorded in each city in 2006. The crime data cover all kinds of criminal offenses (the breakdown is unfortunately unavailable). The treatment variable is the regional unemployment rate as of 2005. As an IV for the unemployment rate, we employ the availability of child daycare facilities in 2006. It would be legitimate to assume that the availability of childcare facilities does not directly affect the crime rate. Meanwhile, in the labor economics literature, there is empirical evidence that expanding childcare services is effective in increasing female labor participation (e.g., Baunachscher and Schlotter 2015). Thus, the availability of childcare facilities can be viewed as an indicator of the local working environment, particularly for young females, and would contribute to reducing the total unemployment rate. For other control variables, we include population density, annual retail sales per capita, the ratio of elderly people, and the ratio of single households. The sales data are as of 2006, and the others are those in 2005.10

The network $A_N$ is defined by whether the cities share a common boundary. Specifically, assuming that there is a limit on the number of cities each city can interact with, we set $A_{ij} = 1$ if city $j$ is adjacent to $i$ and in the $k$-nearest neighbors of $i$. Below, we report the results when $k = 2$ for illustration. We also have tried several different specifications for $A_N$, and confirmed that the results are overall similar (for more details, see Appendix D.2). For all cases, the treatment spillover variable $S_i$ is defined by $i$’s reference-group mean, say $\overline{unempl}$. Then, the model estimated is as follows:

$$
\begin{align*}
\text{crime} &= (1, \text{density}, \text{sales}, \text{elderly}, \text{single})^\top \beta(\overline{unempl}, \overline{unempl}) + g(\text{unempl}, \text{unempl}) \epsilon \\
\text{unempl} &= (\text{density}, \text{sales}, \text{elderly}, \text{single, childcare})^\top \gamma + \eta(U).
\end{align*}
$$

Note that since the treatment equation does not involve any network interactions explicitly, the CQR estimation can be implemented using all data regardless of the specification of $A_N$. For the estimation of the CATR parameter, we need to select a subsample to maintain distributional homogeneity of $(U, V)$. To this end we focus on the cities that have exactly two interacting partners (recalling that $k = 2$). After excluding observations with missing data, the estimation of CATR was performed on a sample of size 1773. The estimation procedure is the same as that in the Monte Carlo experiments in Appendix C. That is, we set the penalty parameter to $\tau = 5/n$ and use the B-spline basis with two internal knots.11 The descriptive statistics of the data are summarized in Table 5 in Appendix D.1.

The estimation results for the treatment equation are provided in Table 6 in Appendix D.1, where we can find that childcare is significantly negatively related to the unempl variable, as expected. Figure 1 presents the estimated CATR($t, s, x$) for different values of $(t, s, x)$. In each panel, $t$ ranges over 0.1 to 0.9 empirical quantiles of $[T_i]$ and $s$ is either at 0.2 or 0.6 quantile of $[S_i]$. Because of the correlation between $T$ and $S$, the CATR estimates at more extreme $t$ cannot be estimate reliably and thus they are not reported (see Figure 2 in Appendix D.1 for the joint density of $(T, S)$). For the value of $x$, we evaluate at the empirical median in the left panel. We consider two more cases for $x$: the median value for the cities in the bottom 20% level of population density among all cities (middle panel), and for the cities in the top 20% level (right panel). The former would be considered as a typical city in a rural area, while the latter as a typical city in an urban area. In the figure, we also report the results obtained when $T$ and $S$ are treated as exogenous (this corresponds to the CQR estimator given in Appendix C).

Now, we report our main empirical findings. First, we can observe that the CATR weakly increases in general as the unemployment rate increases. That is, as the number of unemployed people increases, the crime rate tends to increase, which is consistent with the findings in the prior studies. Second, in the left and right panels, the CATR with large $S$ tends to be significantly greater than the CATR with small $S$, indicating that the average unemployment rate of surrounding cities does affect

10 All this information is freely available from e-Stat (a portal site for Japanese Government Statistics). https://www.e-stat.go.jp/en/regional-statistics/ssdsvview/municipality

11 We have confirmed that the results reported here have a certain robustness to other choices of penalty parameters and basis orders. However, we have also observed that when the number of internal knots is three or higher, it is better to employ some regularized estimator when computing the second derivatives appearing in (3.10) to stabilize the estimates.
the city's own crime rate in such cases. However, interestingly, the spillovers are no longer prominent when the city is in a rural area. One interpretation is that the cities in such areas are relatively independent from other cities and villages, and thus the spillover effects may be less impactful. A more comprehensive figure that summarizes the CATR estimates when $S$ is at 0.2, 0.3, …, 0.8 quantiles is presented in Figure 3 in Appendix D.2, where we can more clearly observe these tendencies. Lastly, we find that ignoring the potential endogeneity of $(\text{unempl}_i, \text{unempl}_j)$ leads to certain differences in the estimates, indicating the importance of accounting for the endogeneity.

5. Conclusion

In this article, we considered a continuous treatment effect model that admits potential treatment spillovers through social networks and the endogeneity for both one’s own treatment and that of others. We proved that the CATR, the conditional expectation of the potential outcome, can be nonparametrically identified under some functional form restrictions and the availability of appropriate IVs. We also considered a more empirically tractable semiparametric treatment model and proposed a three-step procedure for estimating the CATR. The consistency and asymptotic normality of the proposed estimator were established under certain regularity conditions. As an empirical illustration, using Japanese city-level data, we investigated the causal effects of unemployment rate on the number of crime incidents. As a result, we found that the unemployment rate indeed tends to increase the number of crimes and that, interestingly, the unemployment rates of neighboring cities matter more for non-rural cities than rural cities. These results illustrate the usefulness of the proposed method.

Supplementary Materials

The online supplementary file contains the proofs of all technical results, Monte Carlo simulations, and additional results for our empirical analysis.

Acknowledgments

I am grateful to the editor Ivan Canay, the associate editor, three anonymous referees, Takuya Ishihara, and the participants of SEA 2022 for their insightful comments and valuable suggestions, which greatly improved the article.

Disclosure Statement

The author reports there are no competing interests to declare.

Funding

This work was supported financially by JSPS Grant-in-Aid for Scientific Research 20K01597.

References

Altindag, D. T. (2012), “Crime and Unemployment: Evidence from Europe,” International Review of Law and Economics, 32, 145–157. [600]
Aronow, P. M., Eckles, D., Samii, C., and Zonseizin, S. (2021), “Spillover Effects in Experimental Data,” in Advances in Experimental Political Science, eds. J. Druckman and D. P. Green, pp. 289–319, Cambridge: Cambridge University Press. [591]
Bauernschuster, S., and Schlotter, M. (2015), “Public Child Care and Mothers’ Labor Supply—Evidence from Two Quasi-Experiments,” Journal of Public Economics, 123, 1–16. [600]
Becker, G. S. (1968), “Crime and Punishment: An Economic Approach,” Journal of Political Economy, 76, 169–217. [600]
Belloni, A., Chernozhukov, V., Chetverikov, D., and Kato, K. (2015), “Some New Asymptotic Theory for Least Squares Series: Pointwise and Uniform Results,” Journal of Econometrics, 186, 345–366. [598]
Blundell, R., and Powell, J. L. (2003), “Endogeneity in Nonparametric and Semiparametric Regression Models,” in Advances in Economics and Econometrics: Theory and Applications, Eighth World Congress, Econometric Society Monographs, eds. M. Dewatripont, L. P. Hansen, and S. J. Turnovsky, 312–357, Cambridge: Cambridge University Press. [591, 593, 594]
Bramoulle, Y., Djebarri, H., and Fortin, B. (2020), “Peer Effects in Networks: A Survey,” Annual Review of Economics, 12, 603–629. [593]
Chen, L., Smetanina, E., and Wu, W. B. (2020), “Estimation of Nonstationary Nonparametric Regression Model with Multiplicative Structure,” The Econometrics Journal, 25, 176–214. [395, 596]
Chen, X. (2007), “Large Sample Sieve Estimation of Semi-nonparametric Models,” in Handbook of Econometrics (Vol. 6), eds. J. J. Heckman and E. E. Leamer, pp. 5549–5632, Amsterdam: Elsevier. [598]
Chen, X., and Christensen, T. M. (2015), “Optimal Uniform Convergence Rates and Asymptotic Normality for Series Estimators Under Weak Dependence and Weak Conditions,” Journal of Econometrics, 188, 447–465. [598]
Chernozhukov, V., Fernández-Val, I., Newey, W., Stouli, S., and Vella, F. (2020), “Semiparametric Estimation of Structural Functions in Nonseparable Triangular Models,” Quantitative Economics, 11, 503–533. [592, 595]
Cressie, N. A. C. (1993), Statistics for Spatial Data (Revised ed.), Wiley Series in Probability and Statistics, New York: Wiley. [597]
Del Prete, D., Forastiere, L., and ScabolaZZ a, V. L. (2020), “Causal Inference on Networks under Continuous Treatment Interference,” arXiv, 2004.13459. [591]
