Panel Instrumental Variable Regression with Varying-intensity Repeated Treatments: Theory and the China Syndrome Application

Jaerim Choi
Yonsei University
jaerimchoi@yonsei.ac.kr

Dakyung Seong
University of Sydney
dakyung.seong@sydney.edu.au

Shu Shen*
University of California, Davis
shushen@ucdavis.edu

This version: Feb, 2024

Abstract

Instrumental variable models with repeated endogenous treatments are popular in empirical research. This paper shows that if treatment effect dynamics is present and external instruments are serially correlated, the current empirical approaches adopted in the literature is invalid. Using the proposed new model and semi-parametric approach, we find strong evidence of path-dependency in the contemporaneous impact of increased Chinese import competition on U.S. manufacturing employment. Specifically, the trade shock an industry received over 1991-1999 monotonically magnifies the negative impact of the 1999-2011 trade shock. The magnifying effect is mild initially but soars after the 1991-1999 import exposure passes certain threshold.

1 Introduction

The setting of repeated endogenous treatments with varying-intensity and serially correlated external instrumental variables (IV) is popular in applied economics. For example, Autor et al. (2013) and Acemoglu et al. (2016), seminal papers on the nexus between international trade and labor economics, use a two-period panel (1991-1999 and 1999-2011) to assess the repeated endogenous treatments of exposure to rising Chinese import competition across US local labor markets and industries. Stacking data from the four

*The authors are very thankful for all the helpful comments from seminar participants at Syracuse University, the University of Toronto, the 2023 International Association of Applied Econometrics Annual Conference, the 2023 All California Econometrics Conference, and the 2023 Australia New Zealand Econometric Study Group Meeting.

Census years, Boustan (2010) investigates white departures in northern cities in response to large black migration from the rural South to northern cities between 1940 and 1970. More recently, Burchardi et al. (2020) investigate the impact of the staggered arrival of immigrants on innovation using a seven-period, county-level panel model between 1975 and 2010. Numerous other empirical studies have adopted a similar setting.

Despite its popularity in empirical research, the econometrics literature studying effects of repeated varying-intensity endogenous treatments, with external instruments, is sparse. As a result, applied researchers with such a setting have so far adopted a vanilla first-difference instrumental variable (IV) approach as the main workhorse. The approach, following the seminal work of Autor et al. (2013), eliminates fixed effects through first-difference and then solves any remaining endogeneity issues using external instruments proposed from the empirical institutional background. We argue in this paper that the existing approach can be restrictive in many empirical settings because its model overlooks the repeated feature of treatments (after first-differencing). In fact, the first contribution of our paper is to formally show that the existing empirical approach adopted in many empirical research suffers from a contradiction between the exclusion restriction and the rank condition and is, therefore, invalid, if external instruments are serially correlated *and* there exist nontrivial treatment effect dynamics in the outcome equation.

Motivated by the inconsistency result, we propose in this paper a novel model that allows 1) the effect of the current treatment to vary with treatment history and 2) the treatment in the previous period to directly affect the contemporaneous outcome. The new features we propose are relevant to many empirical applications. Take China syndrome as an example. As is documented in Autor et al. (2021), the China trade shock which commenced in the early 1990s has persisted through three distinct periods: the gradual beginning of China's export boom during the 1990s; the dramatic surge of China's export growth during the 2000s after its WTO accession; and China's export plateau after 2010. Across industries, the path of trade shock growth over time differs dramatically as there has been a natural shift in export composition in China following the growth of the Chinese economy. As a result, it is natural to expect that the impact of the China trade shock in a later period depends on trade shocks in previous periods. For instance, a certain industry hit hard by the China trade shock during an earlier period may undergo structural transformation, enabling it to have better capabilities to cope with import competition in the latter period. Our model is the first in the literature that can be used to explore such an important path-dependency feature in China shock effects.

Our new model, together with proposed semi-parametric identification and estimation strategy, uncover rich dynamics of the China trade shock not described previously in the literature (see, e.g., Autor et al., 2013, 2014; Acemoglu et al., 2016; Autor et al., 2020b,a; Feenstra et al., 2019, among many others). We find strong evidence using the industry-level dataset in Acemoglu et al. (2016) that the contemporaneous impact of increased Chinese imports on employment in 1999-2011 depends on the import exposure in 1991-1999. The path-dependency in the "China Syndrome" is monotonic, too, with the previous import exposure magnifying the negative impact of the current trade shock. More interestingly, we find that the size of the magnifying effect is mild for most industries, but becomes much larger when the import exposure increase is over 0.2 percentage points per year between 1991 and 1999. Specifically, the China shock effect in the 2000s is stable and averages to -0.25 when the import exposure change in the 1990s lies between 0 and 0.2, whereas the effects increases to an average of -1.0 when the change lies between 0.2 and 0.3. The substantially bigger contemporaneous effect estimates for industries exposed to larger earlier shocks underscore the importance of path-dependency in analyzing trade effects. Our new empirical results thus shed new light on the China shock literature by estimating a non-linear, path-dependent treatment effect, which cannot be captured in the existing empirical framework.

In the econometrics literature, our proposed model is related to the dynamic treatment effect literature, although the earlier literature has been mostly focusing on binary treatments. Heckman and Navarro (2007) and Heckman et al. (2016) establish important dynamic treatment effect concepts including the *direct effect* and *continuation values*, and propose a decomposition of the *total* longer-term effect of treatment interventions. Han (2021) offers nonparametric identification for average treatment effects as well as optimal treatment regimes. Bojinov et al. (2021) study the identification and finite population inference of dynamic treatment effects under sequential randomization or unconfoundedness. Cellini et al. (2010) study the identification of longer-term direct effects for the repeated regression discontinuity setting. Hsu and Shen (2023) formalize and extend the identification approach to allow for treatment effect heterogeneity. Gallen et al. (2023) extend Cellini et al. (2010) to study a repeated binary endogenous treatment setting where only a single time-invariant external treatment is available. In this paper, our focus is on the evaluation of repeated varying-intensity treatments, a model setting that is different from all the above-mentioned studies.

Our paper also contributes to the panel IV regression literature by introducing an outcome equation that offers more flexible modeling of effects from the time-varying continuous endogenous regressor. We rely on the presence of external instruments for model identification since, in a lot of empirical settings, the sequential exogeneity condition required for using internal instruments is not suitable. Our proposed method is, therefore, distinguished from the vast literature that uses internal instruments for panel IV identification, including Anderson and Hsiao (1982), Arellano and Bond (1991), Ahn and Schmidt (1995), and many others. For semi-parametric estimation, we propose a local polynomial conditional GMM estimator that is related to but different from the estimation approach introduced in the previous literature, including Bravo (2023), Cai and Li (2008), and Su et al. (2013), among many others. The difference is due to the formulation of our conditional moment equality, where the unknown varying coefficient of interest is a function of the endogenous regressor of the previous period. Later in Section 2.1, we detail how this setting naturally arises when studying treatment effects in a dynamic panel structure.

The paper is organized as the following. Section 2 motivates our novel model with direct carryover effect and path-dependent contemporaneous effect both empirically and theoretically. The section also presents parametric identification of the proposed model and shows its limitation. Section 3 studies semi-parametric identification of the proposed model and discusses various extensions of the benchmark model. Section 4 proposes relevant semi-parametric estimation and inference methods. Section 5 applies the proposed method to the industry-level dataset from Acemoglu et al. (2016), finding that the level of Chinese import in the previous period magnifies the contemporaneous impact of Chinese import on the US manufacturing employment. Appendices of the paper includes Monte Carlo simulation results, robustness checks for the empirical analysis, as well as all mathematical proofs of theoretical results.

2 Model and Motivation

2.1 Model Setup

Consider a series of repeated treatments $\{X_{it}\}_{t=1}^{T}$ for individual $i = 1, \dots, N$. Treatments start at $t = 1$ so $X_{it} = 0$ for any $t \leq 0$. Treatments are of variable intensity and potentially endogenous with respect to the series of outcomes $\{Y_{it}\}_{t=1}^{T}$. The identification of treatment effects relies on the presence of external instruments, denoted by $\{Z_{it}\}_{t=1}^{T}$. Both X_{it} and Y_{it} are scalar random variables, while Z_{it} is of dimension d_{zt}. To model an outcome equation, we also consider a d_{ht}-dimensional vector of control variables H_{it}. Without loss of generality, let $H_{it} = (1 \ \widetilde{H}_{it}')'$ so our model always includes an intercept.

This repeated treatment setting is popular in applied economics. For example, Autor et al. (2013) and Acemoglu et al. (2016), seminal papers in the intersection between international trade and labor economics, investigate the effect of rising Chinese import competition on US local labor markets and industries using a two-period panel model (1991-1999 and 1999-2011). The endogenous treatment of interest is the change in local labor market (or industry) exposure to Chinese imports in each time period. Boustan (2010) studies white departures to the suburbs in northern cities in response to large black influx from the rural South to northern cities using Census data between 1940 and 1970. The endogenous treatment is the change in the black population within a city over time. Burchardi et al. (2020) investigate the impact of immigration on innovation between 1975 and 2010 with five-year intervals using more than a decade of migration data from foreign countries to US counties. The endogenous treatment in this case is the number of migrants flowing into the US county in each time interval.

The econometrics literature, as far as the authors know, has not paid much attention to such an important setting. As a result, empirical studies with repeated endogenous treatments remain to predominantly use the following overly simple parametric model:

$$\textit{Existing:} \quad Y_{it} = \beta_t X_{it} + H_{it}'\gamma_t + \epsilon_{it}, \quad t = 1, \dots, T, \tag{2.1}$$

and estimate it through a standard two-stage least squared (2SLS) regression. As mentioned, identification of the model relies on the external instrumental variable Z_{it} that is correlated with X_{it} but uncorrelated with ϵ_{it}, the error term in (2.1), for each time

t. In Autor et al. (2013) and Acemoglu et al. (2016), for example, the external instrument is constructed by using imports from China to eight other high-income countries, excluding the US. The rationale behind using imports from other high-income economies is that they are similarly exposed to Chinese supply shocks, but are likely to be uncorrelated with US demand shocks.

The model (2.1) is overly simple because it only considers a *contemporaneous treatment effect* for each t. In the current study, we illustrate the importance of modeling treatment effect dynamics and show a potential pitfall of the existing approach. Specifically, we propose the following econometrics model that allows for 1) a *direct carryover effect* from the past treatment even when there is no contemporaneous treatment and 2) a *path-dependent* contemporaneous treatment effect.

$$\textit{Benchmark:} \quad Y_{it} = \beta^1_{t-1} X_{i(t-1)} + \beta^0_t \left(X_{i(t-1)}\right) X_{it} + H'_{it}\gamma_t + \varepsilon_{it}, \quad t = 1, \ldots, T. \tag{2.2}$$

As will be discussed shortly, this model naturally arises from the potential outcome framework and can encompass various forms of treatment effect dynamics that may be considered in practice. The functional parameter $\beta^0_t(.)$ is the t-th period contemporaneous treatment effect which is allowed to vary with the treatment take-up of the previous period $t-1$. As an example, in the context of Autor et al. (2013) and Acemoglu et al. (2016), $\beta^0_t(.)$ allows the impact of the "China trade shock" in the 2000s to vary with the intensity of the "China trade shock" in the 1990s. This path-dependency in the contemporaneous treatment effect could arise from several factors. For example, due to the adverse China trade shock in the previous period, innovation activities might reduce (Autor et al., 2020c), which further dampens the ability to cope with competitive pressure from China in the current period. On the contrary, the adverse trade shock in the previous period might boost innovation to escape from the competition (Bloom et al., 2016), which results in industry-level structural changes to better manage the trade shock in the current period.

In the context of Burchardi et al. (2020), the parameter function $\beta^0_t(.)$ allows for the possibility that the contemporaneous impact of the "immigration shock" on innovation varies with the intensity of the "immigration shock" in the previous five-year period, possibly through positive externalities from immigrants settled down earlier. For instance, the first arrival of immigrants helps later immigrants integrate into the local society (Battisti et al., 2022), thereby enabling new immigrants to focus on economic activities, including innovation. Previous immigrant inventors also positively influence the innovation production of their collaborators from the same ethnic origin as documented in Bernstein et al. (2022).

The parameter β^1_{t-1} in model (2.2) is the direct carryover effect of the previous period's treatment. The term "direct" is used following Heckman et al. (2016) to emphasize that β^1_{t-1} is the carryover effect of $X_{i(t-1)}$ on Y_{it} when the contemporaneous treatment X_{it} is zero. The parameter is important for counterfactual policy analysis as discussed in previous literature, including Heckman et al. (2016) in a multi-stage sequential treatment decision setting, Cellini et al. (2010) and Hsu and Shen (2023) in a dynamic regression discontinuity design, and Gallen et al. (2023) in a repeated binary endogenous treatment setting. The direct carryover effect will be extended to include nonlinearity in its functional form later in the paper.

In the context of Autor et al. (2013) and Acemoglu et al. (2016), for example, β^1_{t-1} suggests a possibility of staggered impact arising from long-term adjustments in US manufacturing industries, even after the import growth of Chinese products flattens out. In the context of Burchardi et al. (2020), the carryover effect captures innovation activities in the current period by immigrants who arrived in the previous period.

It is worth mentioning that our proposed benchmark model with dynamic treatment effects arises naturally from a potential outcome perspective. Consider a simple case with $T = 2$. Let potential outcomes given treatment intensities x_1 and x_2 be denoted by $Y_{i1}(x_1)$ and $Y_{i2}(x_1, x_2)$ such that observed outcomes satisfy $Y_{i1} = Y_{i1}(X_{i1})$ and $Y_{i2} = Y_{i2}(X_{i1}, X_{i2})$. Let W_{it} be the vector of time-varying exogenous regressors for each t and A_i be the vector of time-invariant exogenous regressors including the intercept. Suppose the potential outcome model has the following additive form and an autoregressive structure.

$$
\begin{aligned}
Y_{i1}(x_1) = Y_{i1}(0) + \beta^0_{i1} x_1, \quad & Y_{i2}(x_1, x_2) = Y_{i2}(x_1, 0) + \left(Y_{i2}(x_1, x_2) - Y_{i2}(x_1, 0)\right), \\
\text{where } & Y_{i1}(0) = W'_{i1}\gamma^w_1 + A'_i\gamma^a_1 + v_{i1}, \\
& Y_{i2}(x_1, 0) = \rho Y_{i1}(x_1) + \eta x_1 + W'_{i2}\gamma^w_2 + A'_i\gamma^a_2 + v_{i2}, \\
& Y_{i2}(x_1, x_2) - Y_{i2}(x_1, 0) = \beta^0_{i2}(x_1) x_2.
\end{aligned}
$$

Let $\beta_1^0 = \mathbb{E}\left[\beta_{i1}^0\right]$ and $\beta_2^0(x_1) = \mathbb{E}\left[\beta_{i2}^0(x_1)\right]$. The above potential outcome framework leads to the outcome equation in the benchmark model (2.2) with $\beta_1^1 = \rho\beta_1^0 + \eta$, $H_{i1} = (A_i' \; W_{i1}')'$, $\gamma_1 = ((\gamma_1^a)' \; (\gamma_1^w)')'$, $H_{i2} = (A_i' \; W_{i1}' \; W_{i2}')'$, $\gamma_2 = ((\rho\gamma_1^a + \gamma_2^a)' \; \rho(\gamma_1^w)' \; (\gamma_2^w)')'$, $\varepsilon_{i1} = v_{i1} + \left(\beta_{i1}^0 - \beta_1^0\right)x_1$, and $\varepsilon_{i2} = \rho\varepsilon_{i1} + v_{i2} + \left(\beta_{i2}^0(x_1) - \beta_2^0(x_1)\right)x_2$. The autoregressive nature of the potential outcome $Y_{i2}(x_1, 0)$ provides one mechanism of how X_{i1} could affect Y_{i2} even when $X_{i2} = 0$. All other channels of the direct carryover effect are summarized by the parameter η. In this paper, while it would be an interesting topic of future research, we do not separate out different channels of the direct carryover effect. Identification of the direct carryover effect is an important concept in the mediation literature (e.g., Flores and Flores-Lagunes, 2009, 2010).

The above potential outcome model allows individual treatment effects to be heterogeneous in an arbitrary way. Apparently, randomness in treatment effects needs to be restricted by exclusion restrictions for IV-based identification strategies to operate. More details about the exclusion restrictions will be discussed in Section 3.

2.2 Caveats of Ignoring Treatment Effect Dynamics

In this section, we delve into the panel setting of interest by discussing potential caveats with the existing 2SLS approach. For illustration purposes, we temporarily assume that both the external instrument Z_{it} and the control H_{it} are scalar-valued. Moreover, for each i, the instrument Z_{it} is assumed to follow the autoregressive law of motion such that

$$Z_{it} = \alpha_0 + \alpha_1 Z_{i(t-1)} + e_{it}, \tag{2.3}$$

where $E[e_{it}] = 0$ and $E[Z_{i(t-1)}e_{it}] = 0$ for each i and t. We note that the serial correlation of the instrument Z_{it} is commonly observed in practice. For example, the sample correlation coefficient between Z_{i1} and Z_{i2} is 0.86 where i denotes a region in Autor et al. (2013); the sample correlation coefficient between Z_{i1} and Z_{i2} is 0.63 where i denotes an industry in Acemoglu et al. (2016). Such high correlations are not surprising; China's comparative advantage with the rest of the world had been in labor-intensive industries and was not likely to have changed drastically from the 1990s to the 2000s.

Let $\widehat{\beta}_t$ denote the standard IV estimator based on the existing model (2.1) using Z_{it} as the external instrument for X_{it}. The following lemma shows that the estimator is inconsistent for the contemporaneous effect of X_{it} on Y_{it} when the external instruments are serially correlated and the outcome equation model either 1) has a nontrivial carryover effect or 2) features nontrivial path-dependency in the contemporaneous effect.

Lemma 2.1 *Suppose the contemporaneous exogeneity and standard rank condition hold for the external instrument Z_{it} such that $\mathbb{E}[Z_{it}\varepsilon_{it}] = 0$ and $cov(Z_{it}, X_{it}) \neq 0$ for all $t = 1, \dots, T$. In addition, $\mathbb{E}[X_{i(t-1)}e_{it}] = 0$. If the external instrument Z_{it} is serially correlated, then, for any $t \geq 2$, the 2SLS estimator $\widehat{\beta}_t$ defined above is*

(a) inconsistent for the contemporaneous treatment effect β_t^0 if the true model has a nontrivial direct carryover effect, i.e.,

$$Y_{it} = \gamma_t + \beta_{t-1}^1 X_{i(t-1)} + \beta_t^0 X_{it} + \varepsilon_{it}, \quad \beta_{t-1}^1 \neq 0. \tag{2.4}$$

(b) inconsistent for the weighted average contemporaneous treatment effect

$$\bar{\beta}_t^0 = \int \beta_t^0(x) dF_{X_{t-1}}(x),$$

where $F_{X_{t-1}}(.)$ is the cumulative distribution function of $X_{i(t-1)}$, if the true model has a path-dependent contemporaneous effect, i.e.,

$$Y_{it} = \gamma_t + \beta_t^0 \left(X_{i(t-1)}\right) X_{it} + \varepsilon_{it}, \quad \beta_t^0(.) \neq C \text{ for any constant } C. \tag{2.5}$$

The first part of Lemma 2.1 holds because the consistency of $\widehat{\beta}_t$ under DGP (2.4) requires the exclusion restriction such that $\mathbb{E}[Z_{it}(\beta_{t-1}^1(X_{i(t-1)} - \mathbb{E}[X_{i(t-1)}]) + \varepsilon_{it})] = \beta_{t-1}^1 cov(Z_{it}, X_{i(t-1)}) = 0$. Plugging the linear projection of Z_{it} in equation (2.3) and conditions of the lemma, we find that $cov(Z_{it}, X_{i(t-1)}) = cov(\alpha_0 + \alpha_1 Z_{i(t-1)} + e_{it}, X_{i(t-1)}) = \alpha_1 cov(Z_{i(t-1)}, X_{i(t-1)})$.[1] Given the nontrivial direct carryover effect (i.e., $\beta_{t-1}^1 \neq 0$) and the serial correlation of Z_{it} (i.e., $\alpha_1 \neq 0$), a contradiction arises between the exclusion restriction and the rank condition of period $t-1$.[2] The second part of the lemma holds

[1]The condition $\mathbb{E}[X_{i(t-1)}e_{it}] = 0$ is only a sufficient condition required to ensure that $\alpha_1 cov(Z_{i(t-1)}, X_{i(t-1)})$ does not happen to cancel out with $cov(e_{it}, X_{i(t-1)})$.

[2]It is worth noting that the inconsistency result of $\widehat{\beta}_t$ cannot be generalized to the case where there is no carryover effect (i.e., $\beta_{t-1}^1 = 0$ in equation (2.4)) but the outcome shock is correlated with last period's treatment (i.e., $\mathbb{E}[X_{i(t-1)}\varepsilon_{it}] \neq 0$). Consider, for example, a DGP where $Z_{i2} = \alpha_0 + \alpha_1 Z_{i1} + e_{i2}$ following equation (2.3), $X_{i1} = \gamma_0 + \gamma_Z Z_{i1} + \nu_{i1}$, and $\varepsilon_{i2} = \nu_{i1} + v_{i1}$. If $(\nu_{i1}, v_{i1}) \perp (Z_{i1}, e_{i2})$, then it is clear that $\mathbb{E}[Z_{i2}\varepsilon_{i2}] = 0$ and $\mathbb{E}[\varepsilon_{i2}X_{i1}] \neq 0$ hold simultaneously.

because for $\widehat{\beta}_t$ to be a consistent estimator of $\bar{\beta}_t^0$ under DGP (2.5), we need the exclusion restriction $\mathbb{E}[Z_{it}\left((\beta_t^0(X_{i(t-1)}) - \bar{\beta}_t^0)X_{it} + \varepsilon_{it}\right)] = \mathbb{E}[Z_{it}(\beta_t^0\left(X_{i(t-1)}\right) - \bar{\beta}_t^0\right)X_{it}] = cov\left(Z_{it}X_{it}, \beta_t^0(X_{i(t-1)})\right) = 0$. The restriction does not hold in general because Z_{it} is correlated with $X_{i(t-1)}$.

As is discussed earlier, the commonly used IV in the China syndrome literature is highly serially correlated. In addition, there is strong empirical evidence (see Section 5) that, in the China Syndrome application, the contemporaneous impact of increased Chinese import competition on U.S. manufacturing employment depends on the import exposure of the preceding period. Given the empirical evidence, Lemma 2.1 essentially implies that the current empirical strategy adopted in the China Syndrome literature is invalid. A new estimation strategy taking into account treatment effect dynamics, therefore, needs to be developed.

Discussions:

The inconsistency results stated above are rooted in the repeated endogenous treatments design and the serial correlation of external instruments. It is hence worthwhile to differentiate our benchmark model (2.2) from a single treatment setting with effect heterogeneity, where the exclusion restriction of the existing 2SLS approach is very likely to hold. Specifically, consider the following model with cross-sectional observations:

$$Y_i = \beta_0 + \beta^W W_i + \beta^X(W_i)X_i + \varepsilon_i,$$

where Y_i, X_i, and W_i are the outcome, endogenous treatment, and exogenous variable, respectively. In the above, the treatment effect of X_i on Y_i is potentially heterogeneous with respect to covariate W_i. In such a model, the vanilla 2SLS regression of Y_i on X_i instrumented by Z_i would provide a valid estimator for the average treatment effect $\mathbb{E}[\beta^X(W_i)]$ if, in addition to $\mathbb{E}[\varepsilon_i Z_i] = 0$, one also has $\mathbb{E}[Z_i(W_i - \mathbb{E}[W_i])] = 0$ and $\mathbb{E}[Z_i(\beta^X(W_i) - \mathbb{E}[\beta^X(W_i)])X_i] = 0$. Both conditions would hold if, for example, $Z_i \perp W_i | X_i$. In the repeated endogenous treatments design, however, the treatment effect heterogeneity relies on the level of the past treatment variable. The last period's treatment cannot be uncorrelated with this period's external instrument as long as the external instrument has a nontrivial serial correlation. Lemma 2.1, therefore, implies that treatment effect dynamics, if potentially present, need to be modeled explicitly when formulating the outcome equation models.

2.3 Parametric Identification

In the appendix, we give formal assumptions for parametric identification of the proposed *benchmark* model under the sequential exogeneity condition of external instruments. We discuss two simple special cases here for the ease of understanding.

When $T = 2$ and $\beta_2^0(.) = \beta_2^0$, the benchmark model in (2.2) reduces to

$$Y_{i1} = \gamma_1 + \beta_0^1 X_{i1} + \varepsilon_{i1},$$
$$Y_{i2} = \gamma_2 + \beta_1^1 X_{i1} + \beta_2^0 X_{i2} + \varepsilon_{i2}.$$

The first period could be identified by a 2SLS regression of Y_{i1} on X_{i1} instrumented by Z_{i1}. The second period could be identified by a 2SLS regression of Y_{i2} on X_{i1} and X_{i2} instrumented by Z_{i1} and Z_{i2}, assuming sequential exogeneity of Z_{it}.[3] Similarly, if $\beta_2^0(x) = \eta_{2,1} + \eta_{2,2}x$, the benchmark model for $t = 2$ reduces to

$$Y_{i2} = \gamma_2 + \beta_1^1 X_{i1} + (\eta_{2,1} + \eta_{2,2} X_{i1}) X_{i2} + \varepsilon_{i2},$$

which could be identified by a 2SLS regression of Y_{i2} on X_{i1}, X_{i2}, and $X_{i1}X_{i2}$ instrumented by Z_{i1}, Z_{i2}, and $Z_{i1}Z_{i2}$.

The above parametric approach can face several obstacles in empirical applications. First, the parametric 2SLS approach requires functional-form knowledge of the path-dependent contemporaenous effect $\beta_2^0(.)$. Second, having a 2SLS regression with multiple endogenous regressors can be demanding in terms of the first-stage rank condition, especially when the sample size is small, such as in the China syndrome application we study in Section 5. Third, the parametric 2SLS approach requires the strict exogeneity of the external instrument, which may not be a good assumption in some empirical applications when general equilibrium effects are present.

[3] The second-period model could also be identified by a 2SLS regression of Y_{i2} on X_{i1} and X_{i2} instrumented by X_{i1} and Z_{i2}, assuming sequential exogeneity of endogenous treatment. However, if there is a feedback effect from X_{i1} to the second-period outcome shock or that X_{i1} is passively correlated to ε_{i2} through the serial correlation of outcome shocks, the additional sequential exogeneity condition $\mathbb{E}[X_{i1}\varepsilon_{i2}] = 0$ is not likely to hold.

In the next section, we explore an alternative semi-parametric identifying strategy with a novel conditional exclusion restriction. The conditional exclusion restriction only requires contemporary exogeneity of the external instrument. In addition, the new strategy allows for semi-parametric identification of the proposed model, imposing no assumptions on the functional form of the path-dependent treatment effect of interest.

3 Semi-parametric Identification

3.1 Identification With Exogenous Control Vector

We first consider identification assuming the control vector $\widetilde{H}_{it}$ is exogenous. Let $\ddot{Z}_{it} = (\omega_t(Z_{it})' \ H'_{it})'$ with some known function $\omega_t(.)$. Let $\mathcal{X}_{t-1}$ denote the support of $X_{i(t-1)}$. In some applications, $\mathcal{X}_{t-1}$ may be a particular subset of empirical interests.

Assumption 3.1 (semi-parametric identification) *For all $t = 1, \ldots, T$, assume that*

(a) (exclusion restriction I) $\mathbb{E}[\varepsilon_{it}] = 0$ and $\mathbb{E}[\varepsilon_{it}|X_{i(t-1)}, Z_{it}, \widetilde{H}_{it}] = \mathbb{E}[\varepsilon_{it}|X_{i(t-1)}]$. For $t = 1$, since $X_{i0} = 0$, the assumption reduces to $\mathbb{E}[\varepsilon_{i1}|Z_{i1}, \widetilde{H}_{i1}] = 0$.

(b) (rank condition) $\mathbb{E}[\ddot{Z}_{it}(X'_{it} \ H'_{it})|X_{i(t-1)} = x]$ is full rank for all $x \in \mathcal{X}_{t-1}$.

Assumption 3.1.(a) denotes the contemporaneous mean independence between external instruments and error terms *after* conditioning on the last treatment. For a two-period model, it reduces to $\mathbb{E}[\varepsilon_{i1}|Z_{i1}, \widetilde{H}_{i1}] = 0$ and $\mathbb{E}[\varepsilon_{i2}|X_{i1}, Z_{i2}, \widetilde{H}_{i2}] = \mathbb{E}[\varepsilon_{i2}|X_{i1}]$.

As a well-known relationship between unconditional and conditional mean independence, Assumption 3.1.(a) is not weaker (nor stronger) than its unconditional counterpart $\mathbb{E}[\varepsilon_{it}|Z_{it}, \widetilde{H}_{it}] = 0$. In applied settings, in fact, conditional mean independence tends to be regarded as more robust than unconditional independence. For instance, if $X_{i(t-1)}$ has impacts on both Z_{it} and ε_{it}, then Z_{it} is not unconditionally exogenous with respect to ε_{it}, but the conditional exclusion restriction in Assumption 3.1.(a) can still hold.

Recall that $H_{it} = (1 \ \widetilde{H}'_{it})'$ so the first element of γ_t is the intercept. Denote the intercept by $\gamma_{t,1}$ and let $\gamma_{t,-1}$ include all the remaining elements of γ_t except $\gamma_{t,1}$. Let $g_t(x) = \gamma_{t,1} + \beta^1_{t-1}x + \mathbb{E}[\varepsilon_{it}|X_{i(t-1)} = x]$. Assumption 3.1.(a) implies that for all $x \in \mathcal{X}_{t-1}$

and $t = 1, \ldots, T$,

$$
\begin{aligned}
&\mathbb{E}\left[\ddot{Z}_{it}\left(Y_{it} - \left(g_t(x) + \beta_t^0(x)X_{it} + \widetilde{H}_{it}'\gamma_{t,-1}\right)\right)|X_{i(t-1)} = x\right] \\
=&\mathbb{E}\left[\ddot{Z}_{it}\left(\mathbb{E}\left[\varepsilon_{it}|Z_{it}, \widetilde{H}_{it}, X_{i(t-1)} = x\right] - \mathbb{E}\left[\varepsilon_{it}|X_{i(t-1)} = x\right]\right)|X_{i(t-1)} = x\right] = 0.
\end{aligned}
$$

For $t = 1$, the above conditional moment equality reduces to a classic unconditional moment equality such that $\mathbb{E}\left[\ddot{Z}_{it}\left(Y_{it} - \left(\beta_1^0 X_{i1} + H_{i1}'\gamma_1\right)\right)\right] = \mathbb{E}\left[\ddot{Z}_{i1}\varepsilon_{i1}\right] = 0$ since $X_{i0} = 0$ and $g_1(.) = \gamma_{1,1}$. For all $t \geq 2$, the above conditional moment equality implies the identification of $(g_t(x)\ \beta_t^0(x)\ \gamma_{t,-1}')'$, for all $x \in \mathcal{X}_{t-1}$. Whether the conditional moment equality is just-identified or over-identified depends on the dimension of the function $\omega(.)$.

The above identification strategy is not impacted if the proposed model in (2.2) is generalized to include a nonlinear direct carryover effect and/or path-dependent slopes of exogenous controls:

$$
Y_{it} = \beta_t^0\left(X_{i(t-1)}\right)X_{it} + H_{it}'\gamma_t(X_{i(t-1)}) + \varepsilon_{it}, \tag{3.1}
$$

where $\gamma_t(.) = (\beta_{t-1}^1(.)\ \gamma_{t,-1}'(.))'$ with the intercept term soaked into the nonlinear carryover effect function $\beta_{t-1}^1(.)$, for all $t = 1, \ldots, T$. Under model (3.1), Assumption 3.1.(a) implies that for all $x \in \mathcal{X}_{t-1}$ and $t = 1, \ldots, T$,

$$
\mathbb{E}\left[\ddot{Z}_{it}\left(Y_{it} - \left(g_t(x) + \beta_t^0(x)X_{it} + \widetilde{H}_{it}'\gamma_{t,-1}(x)\right)\right)|X_{i(t-1)} = x\right] = 0, \tag{3.2}
$$

where $g_t(x) = \beta_{t-1}^1(x) + \mathbb{E}[\varepsilon_{it}|X_{i(t-1)} = x]$. Again, the conditional moment equality identifies the path-dependent contemporaneous treatment effect $\beta_t^0(.)$ treating the carryover effect function as a nuisance parameter. However, if one is willing to assume sequential exogeneity of the endogenous treatment, or that $\mathbb{E}\left[\varepsilon_{it}|X_{i(t-1)}\right] = 0$, the function $g_t(.)$ in equation (3.2) reduces to the direct carryover effect function.

3.2 Identification with Potentially Endogenous Additional Controls

Note that Assumption 3.1.(a) requires additional controls in $\widetilde{H}_{it}$ to be exogenous. The restriction can be relaxed to the following assumption if we do not pursue separate identification of effects from these additional controls.

Assumption 3.2 (semi-parametric identification: exclusion restriction II) *Assume that* $\mathbb{E}[\varepsilon_{it}] = 0$, $\mathbb{E}[\varepsilon_{it}|X_{i(t-1)}, Z_{it}] = \mathbb{E}[\varepsilon_{it}|X_{i(t-1)}]$, *and* $\mathbb{E}[\widetilde{H}_{it}|X_{i(t-1)}, Z_{it}] = \mathbb{E}[\widetilde{H}_{it}|X_{i(t-1)}]$, *for all $t = 1, \ldots, T$. For $t = 1$, since $X_{i0} = 0$, the assumption reduces to $\mathbb{E}[\varepsilon_{i1}|Z_{i1}] = 0$ and $\mathbb{E}[\widetilde{H}_{i1}|Z_{i1}] = \mathbb{E}[\widetilde{H}_{it}]$.*

Assumption 3.2 allows the control vector $\widetilde{H}_{it}$ to be endogenous, as long as the external instrument is mean independent of both $\widetilde{H}_{it}$ and the error term ε_{it} after conditioning on the past treatment $X_{i(t-1)}$.

Assumption 3.2 implies that for all $x \in \mathcal{X}_{t-1}$,

$$\begin{aligned} & \mathbb{E}\left[\omega_t(Z_{it})\left(Y_{it} - \left(\beta_t^0(x)X_{it} + H_{it}'\widetilde{\gamma}_t(x)\right)\right)|X_{i(t-1)} = x\right] \\ =& \mathbb{E}\left[\omega_t(Z_{it})\varepsilon_{it}|X_{i(t-1)} = x\right] - \mathbb{E}\left[\omega(Z_{it})\mathbb{E}[\varepsilon_{it}|X_{i(t-1)} = x]|X_{i(t-1)} = x\right] = 0, \end{aligned} \tag{3.3}$$

where $\widetilde{\gamma}_t(x) = \gamma_t(x) + \mathbb{E}[H_{it}H_{it}'|X_{i(t-1)} = x]^{-1}\mathbb{E}[H_{it}\varepsilon_i|X_{i(t-1)} = x]$ for all $x \in \mathcal{X}_{t-1}$.[4] The first equality is explained in the appendix. The second equality holds because $\mathbb{E}[\varepsilon_{it}|X_{i(t-1)}, Z_{it}] = \mathbb{E}[\varepsilon_{it}|X_{i(t-1)}]$ by Assumption 3.2.

In addition, Assumption 3.2 implies that for all $x \in \mathcal{X}_{t-1}$

$$\begin{aligned} & \mathbb{E}\left[H_{it}\left(Y_{it} - \left(\beta_t^0(x)X_{it} + H_{it}'\widetilde{\gamma}(x)\right)\right)|X_{i(t-1)} = x\right] \\ =& \mathbb{E}\left[H_{it}\varepsilon_{it}|X_{i(t-1)} = x\right] - \mathbb{E}\left[H_{it}H_{it}'|X_{i(t-1)} = x\right]\left(\widetilde{\gamma}(x) - \gamma(x)\right) = 0. \end{aligned}$$

Summing up, Assumption 3.2 implies that

$$\mathbb{E}\left[\ddot{Z}_{it}\left(Y_{it} - \left(\beta_t^0(x)X_{it} + H_{it}'\widetilde{\gamma}_t(x)\right)\right)|X_{i(t-1)} = x\right] = 0. \tag{3.4}$$

Compared to the identification in the last section in (3.1), allowing $\widetilde{H}_{it}$ to be endogenous implies that we cannot separately identify its effect on the outcome. However, if the external instrument used to identify $\beta_t^0(.)$ does not move with $\widetilde{H}_{it}$ conditional on $X_{i(t-1)}$, the endogeneity of $\widetilde{H}_{it}$ does not influence the identification of $\beta_t^0(.)$. For identification of $\beta_t^0(.)$ only, there is no need to distinguish whether Assumption 3.1.(a) or Assumption 3.2 is used as the exclusion restriction. In practice, researchers can choose either of them depending on whether the additional control vector in the empirical application is potentially endogenous.

Discussions:

[4] The definition implies that $H_{it}'\left(\widetilde{\gamma}_t(x) - \gamma_t(x)\right) = H_{it}'\mathbb{E}[H_{it}H_{it}'|X_{i(t-1)} = x]^{-1}\mathbb{E}[H_{it}\varepsilon_i|X_{i(t-1)} = x] \equiv \mathbb{L}[\varepsilon_i|X_{i(t-1)} = x, \widetilde{H}_{it}]$, the population level linear projection of ε_i on $\widetilde{H}_{it}$ conditional on $X_{i(t-1)} = x$.

The identification strategy discussed above can be adapted to the general cross-sectional setting discussed at the end of Section 2.2. As is discussed there, unlike in the repeated treatment setting, it is more likely to have W_i and the external instrument Z_i to satisfy certain exclusion conditions such that omitting the direct effect of W_i on Y_i and the heterogeneity in the effect of X_i on Y_i is fine. However, if the exclusion conditions detailed at the end of Section 2.2 cannot be argued in some empirical applications, or researchers are genuinely interested in effect heterogeneity, the identification strategy discussed in this section could be easily adapted to this cross-sectional setting.

It is worth mentioning that the cross-section extension in discussion also incorporates the popular fuzzy regression discontinuity (RD) set-up. If we label the binary endogenous treatment in the RD setting using T_i and let the external instrument be $Z_i = 1(W_i \geq c)$ where W_i is the endogenous running bvariable and c the RD threshold, we can write down the fuzzy RD model with covariates (see e.g., Calonico et al., 2019) as

$$\begin{aligned} &Y_i(0) = \beta_W(W_i) + \widetilde{H}_i'\gamma_{-1}(W_i) + \varepsilon_i, \quad Y_i(1) = Y_i(0) + \beta_T(W_i), \\ \Rightarrow\ &Y_i = Y_i(0)(1-T_i) + Y_i(1)T_i = \beta_W(W_i) + \beta_T(W_i)T_i + \widetilde{H}_i'\gamma_{-1}(W_i) + \varepsilon_i, \end{aligned}$$

just as the outcome equation stated at the end of Section 2.2.

The identification assumption of fuzzy RD is essentially that, for any $w \in (c-\epsilon, c+\epsilon)$ with small $\epsilon > 0$, $\mathbb{E}[\varepsilon_i|W_i = w, Z_i] = \mathbb{E}[\varepsilon_i|W_i = w]$ and $\mathbb{E}[\widetilde{H}_i|W_i = w, Z_i] = \mathbb{E}[\widetilde{H}_i|W_i = w]$, similar to Assumption 3.2. The RD treatment effect $\beta_T(c)$ for compliers at the RD cutoff c is identified by a conditional moment equality similar to (3.4):

$$\mathbb{E}\left[(Z_i\ H_i')'\left(Y_i - \left(\beta_T(w)T_i + H_i'\widetilde{\gamma}(w)\right)\right)|W_i = w\right] = 0,$$

for $w \in (c-\epsilon, c+\epsilon)$ and any small $\epsilon > 0$, where the nuisance parameter function $\widetilde{\gamma}(w) = (\beta_w(w)\ \gamma_{-1}(w)')' + \mathbb{E}[H_iH_i'|W_i = w]^{-1}\mathbb{E}[H_i\varepsilon_i|W_i = w]$.

4 Semi-parametric Estimation and Inference

This section studies semi-parametric estimation of the path-dependent contemporaneous effect function $\beta_t^0(.)$ and the functional intercept $g_t(.)$. Sections 4.1 and 4.2 discuss estimation and pointwise inference procedures following the identification result in conditional moment equality (3.2)[5], where all slopes of the outcome equation are assumed

[5] As is discussed earlier, the difference between identification results summarized by equation (3.2) and equation (3.4) lies only in the interpretation of slope coefficients of additional controls $\widetilde{H}_{it}$. Therefore,

to be functional. Section 4.3 proposes a semi-parametric estimator for the average contemporaneous effect $\bar{\beta}_t^0(.)$ defined in Lemma 2.1. The proposed average effect estimator enjoys a parametric rate of convergence. Without loss of generality, we set $t = 2$ in this section for the purpose of illustration.

4.1 Estimation of Functional Coefficients

Let $\theta_2(x) = (\beta_2^0(x)\; g_2(x)\; \gamma'_{2,-1}(x))'$ collect all parameters of interest. Let $\ddot{X}_{i2} = (X'_{i2}\; H'_{i2})'$. The identification result in the previous section, for the case of $t = 2$, can be summarized by the following conditional moment equality:

$$\mathbf{\Lambda}_{\ddot{Z}Y}(x) - \mathbf{\Lambda}_{\ddot{Z}\ddot{X}}(x)\theta_2(x) = 0, \tag{4.1}$$

where $\mathbf{\Lambda}_{\ddot{Z}Y}(x) = \mathbb{E}[\ddot{Z}_{i2}Y_{i2}|X_{i1} = x]$ is a $(K+d_{h2})\times 1$ matrix and $\mathbf{\Lambda}_{\ddot{Z}\ddot{X}}(x) = \mathbb{E}[\ddot{Z}_{i2}\ddot{X}'_{i2}|X_{i1} = x]$ is a $(K + d_{h2}) \times (1 + d_{h2})$ matrix.

Let $\mathbf{W}(x)$ be a pre-determined $(K+d_{h2})\times(K+d_{h2})$ weighting matrix. For all $x \in \mathcal{X}_1$, define $\widehat{\theta}_2(x)$ as the solution to

$$\min_{\theta_2(x)} \left(\widehat{\mathbf{\Lambda}}_{\ddot{Z}Y}(x) - \widehat{\mathbf{\Lambda}}_{\ddot{Z}\ddot{X}}(x)\theta_2(x)\right)' \mathbf{W}(x) \left(\widehat{\mathbf{\Lambda}}_{\ddot{Z}Y}(x) - \widehat{\mathbf{\Lambda}}_{\ddot{Z}\ddot{X}}(x)\theta_2(x)\right), \tag{4.2}$$

where $\widehat{\mathbf{\Lambda}}_{\ddot{Z}Y}(x)$ and $\widehat{\mathbf{\Lambda}}_{\ddot{Z}\ddot{X}}(x)$ are local polynomial estimators of $\mathbf{\Lambda}_{\ddot{Z}Y}(x)$ and $\mathbf{\Lambda}_{\ddot{Z}\ddot{X}}(x)$, respectively. It is clear that

$$\widehat{\theta}_2(x) = \left(\widehat{\mathbf{\Lambda}}'_{\ddot{Z}\ddot{X}}(x)\mathbf{W}(x)\widehat{\mathbf{\Lambda}}_{\ddot{Z}\ddot{X}}(x)\right)^{-1} \widehat{\mathbf{\Lambda}}'_{\ddot{Z}\ddot{X}}(x)\mathbf{W}(x)\widehat{\mathbf{\Lambda}}_{\ddot{Z}Y}(x). \tag{4.3}$$

If $K = 1$, the conditional moment equality in (4.1) is just identified, and

$$\widehat{\theta}_2(x) = \left(\widehat{\mathbf{\Lambda}}_{\ddot{Z}\ddot{X}}(x)\right)^{-1} \widehat{\mathbf{\Lambda}}_{\ddot{Z}Y}(x).$$

Let $[.]_{[j]}$ denote the j-th element of the original vector. We define the estimator for the contemporaneous treatment effect function $\beta_2^0(.)$ and the estimator of the intercept function $g_2(.)$ as, respectively,

$$\widehat{\beta}_2^0(.) = \left[\widehat{\theta}_2(.)\right]_{[1]} \text{ and } \widehat{g}_2(.) = \left[\widehat{\theta}_2(.)\right]_{[2]}.$$

Our proposed estimation approach is related to the augmented moment equality approach discussed in Bravo (2023). In particular, if we follow Bravo (2023) and carry out

we illustrate semi-parametric estimation and inference following the identification result in (3.2).

a p-th order local polynomial approximation of the varying-coefficient function $\theta_2(.)$, we obtain the augmented estimation function

$$\mathbb{E}\left[\ddot{Z}_{i2}^{h,x}\left(Y_{i2}-\sum_{l=0}^{\ell}\frac{1}{p!}\ddot{X}_{i2}'\theta_2^{(p)}(x)\left(\frac{X_{i1}-x}{h}\right)^p\right)|X_{i1}=x\right]\approx 0, \tag{4.4}$$

with $\ddot{Z}_{i2}^{h,x}=\left(1\ \frac{X_{i1}-x}{h}\ \ldots\ (\frac{X_{i1}-x}{h})^p\right)'\otimes\ddot{Z}_{i2}$ being the augmented instrument vector including additional local instruments with bandwidth h.[6] The parameter vector of interest in (4.4) is $\ddot{\theta}_2^p(x)=\left(\theta_2'(x)\ \left(\theta_2^{(1)}(x)\right)'\ \cdots\ \left(\theta_2^{(p)}(x)\right)'\right)'$, which then could be estimated using the same semi-parametric estimation strategy as discussed above for $\theta_2(.)$.

In this paper, we particularly focus on the case $p=0$ in (4.4), which is just the original conditional moment equality (4.1) stated in the beginning of the section. This is because a higher order expansion of $\beta_2^0(x)$, the first element of $\theta_2(x)$, results in more demanding first-stage rank conditions that are not compatible with our empirical application. However, the proposed estimation approach can be readily adapted with allowing a higher order expansion of the rest of $\theta_2(x)$ by adding interaction terms of $(X_{i1}-x)^p$, for $p=1,2,\ldots$, with the original control vector H_{i2}. The identification results in (3.2) and (3.4) remain valid with the augmented control vector because the moment equalities are conditional on fixed values of X_{i1}.

4.2 Asymptotic Properties of the Proposed Functional Estimator

Asymptotic properties of the proposed estimator $\widehat{\theta}_2(x)$ are summarized in the following. Let ℓ be the order of local polynomials used to estimate the population counterparts of $\widehat{\mathbf{\Lambda}}_{\ddot{Z}\ddot{X}}(x)$ and $\widehat{\mathbf{\Lambda}}_{\ddot{Z}Y}(x)$. We specifically consider the case with $\ell=0$ and $\ell=1$ for local constant and local linear conditional mean estimation. Let $\kappa(.)$ be the kernel function and h be the bandwidth. Let $\kappa_h(X_{i1}-x)=\kappa((X_{i1}-x)/h)/h$. Let $\mu_k=\int u^k\kappa(u)du$ and $\nu_k=\int u^k\kappa^2(u)du$. Further, define $\mathbf{\Omega}(x)=\mathbf{W}(x)\mathbf{\Lambda}_{\ddot{Z}\ddot{X}}(x)\left(\mathbf{\Lambda}_{\ddot{Z}\ddot{X}}'(x)\mathbf{W}(x)\mathbf{\Lambda}_{\ddot{Z}\ddot{X}}(x)\right)^{-1}$ and $\mathbf{\Sigma}(x)=\frac{\nu_0}{f_{X_1}(x)}\mathbb{E}\left[\widetilde{\varepsilon}_{i2}^2\ddot{Z}_{i2}\ddot{Z}_{i2}'\middle|X_{i1}=x\right]$, where $\widetilde{\varepsilon}_{i2}=\varepsilon_{i2}-\mathbb{E}[\varepsilon_{i2}|X_{i1}]$.

Let $\mathcal{X}_1^*$ be an interior subset of $\mathcal{X}_1$, the support of X_{i1}. Researchers are interested in estimating the varying coefficient function $\theta_2(.)$ on $\mathcal{X}_1^*$. The following theorem summarizes the asymptotic properties of the proposed kernel-based conditional GMM estimator $\widehat{\theta}_2(.)$.

[6] When $p=1$, (4.4) coincides with the conditional moment equality considered in Cai and Li (2008) and Su et al. (2013).

Theorem 4.1 *Suppose that the data* $\{Y_{i2}, X_{i1}, X_{i2}, Z_{i2}, \widetilde{H}_{i2}\}_{i=1}^N$ *follow the data generating process in equation* (3.1) *and that Assumption 3.1 (or having Assumption 3.2 replacing 3.1.(a)) and Assumption D.1 in the appendix hold. Let* $\mathsf{c}_f(x) = f^{(1)}(x)/f(x)$*. Then for each* $x \in \mathcal{X}_1^*$,

$$\sqrt{Nh}\left(\widehat{\theta}_2(x) - \theta_2(x) - h^2\mu_2\mathbf{\Omega}'(x)\mathbf{B}_\ell(x)\right) \to_d N(0, \mathbf{\Omega}'(x)\mathbf{\Sigma}(x)\mathbf{\Omega}(x)),\ \ \ell = 0, 1,$$

where the bias term $\mathbf{B}_0(x) = \mathbf{\Lambda}^{(1)}_{\ddot{Z}\ddot{X}}(x)\theta_2^{(1)}(x) + \mathbf{\Lambda}_{\ddot{Z}\ddot{X}}(x)\left(\mathsf{c}_f(x)\theta_2^{(1)}(x) + \theta_2^{(2)}(x)/2\right)$ *for local constant estimation and* $\mathbf{B}_1(x) = \mathbf{\Lambda}^{(1)}_{\ddot{Z}\ddot{X}}(x)\theta_2^{(1)}(x) + \mathbf{\Lambda}_{\ddot{Z}\ddot{X}}(x)\theta_2^{(2)}(x)/2$ *for local linear estimation. In the just-identified case, we have*

$$\sqrt{Nh}\left(\widehat{\theta}_2(x) - \theta_2(x) - h^2\mu_2\mathbf{\Omega}'(x)\mathbf{B}_\ell(x)\right) \to_d N\left(0, (\mathbf{\Lambda}^{-1}_{\ddot{Z}\ddot{X}}(x))'\mathbf{\Sigma}(x)\mathbf{\Lambda}^{-1}_{\ddot{Z}\ddot{X}}(x)\right).$$

Theorem 4.1 states the pointwise asymptotic normality of the functional coefficients. Similar results can be found in earlier articles on the varying coefficient models, including Su et al. (2013), Cai et al. (2019) and Bravo (2023) for their own estimators. If a subset of the functional parameter $\theta_2(.)$ reduces to constant, as in the *benchmark* model (2.2), a faster convergence rate could be obtained for the degenerating subset of parameters. More will be discussed in Section 4.3.

Let $\widehat{\varepsilon}_{i2} = Y_{i2} - \ddot{X}'_{i2}\widehat{\theta}_2(X_{i1})$ and $\widehat{f}_{X_1}(x)$ be a consistent estimator of $f_{X_1}(x)$ for each $x \in \mathcal{X}_1$. The asymptotic variance stated in Theorem 4.1 could be estimated by $\widehat{\mathbf{\Omega}}(x) = \mathbf{W}(x)\widehat{\mathbf{\Lambda}}_{\ddot{Z}\ddot{X}}(x)(\widehat{\mathbf{\Lambda}}_{\ddot{Z}\ddot{X}}(x)'\mathbf{W}(x)\widehat{\mathbf{\Lambda}}_{\ddot{Z}\ddot{X}}(x))^{-1}$ and

$$\widehat{\mathbf{\Sigma}}(x) = \frac{h}{N\widehat{f}^2_{X_1}(x)}\sum_{i=1}^N \widehat{\varepsilon}^2_{i2}\ddot{Z}_{i2}\ddot{Z}'_{i2}\kappa_h^2(X_{i1} - x),$$

which are consistent estimators for $\mathbf{\Omega}(x)$ and $\mathbf{\Sigma}(x)$, respectively. The following proposition formalizes.

Proposition 4.1 *Suppose that the conditions in Theorem 4.1 hold. Then for each* $x \in \mathcal{X}_1^*$, $\widehat{\mathbf{\Omega}}(x) \to_p \mathbf{\Omega}(x)$ *and* $\widehat{\mathbf{\Sigma}}(x) \to_p \mathbf{\Sigma}(x)$*, and thus* $\widehat{\mathbf{\Omega}}'(x)\widehat{\mathbf{\Sigma}}(x)\widehat{\mathbf{\Omega}}(x) \to_p \mathbf{\Omega}'(x)\mathbf{\Sigma}(x)\mathbf{\Omega}(x)$.

4.3 Average Effects

The kernel estimator defined above for the path-dependent contemporaneous treatment effect also implies a semi-parametric estimator for the average contemporaneous treatment effect parameter defined in Lemma 2.1. In empirical studies, it might be of interest to estimate the average effect of treatment even if the treatment is potentially heterogeneous.

Let $1_{\{A\}}$ be an indicator function that takes 1 if A is true and 0 otherwise. Let $\vartheta_2 = \mathbb{E}[\theta_2(X_{i1})|X_{i1} \in \mathcal{X}_1^*] = p^{-1} \int_{\mathcal{X}_1^*} \theta_2(x) dF_{X_1}(x)$, where $p = \mathbb{E}[1_{\{X_{i1} \in \mathcal{X}_1^*\}}]$. Define

$$\widehat{\vartheta}_2 = \frac{1}{\sum_{i=1}^N 1_{\{X_{i1} \in \mathcal{X}_1^*\}}} \sum_{i=1}^N \widehat{\theta}_2(X_{i1}) 1_{\{X_{i1} \in \mathcal{X}_1^*\}}$$

as the estimator for ϑ_2. Recall that the set $\mathcal{X}_1^*$ is not necessarily the whole support of X_{i1}. Practitioners may restrict their interest to a particular subset of empirical interests or where data is rich. In the empirical example of "China Syndrome", for example, we set $\mathcal{X}_1$ to $[0, 0.5]$ given the extreme right-skewedness of the X_{i1} distribution. Given $\widehat{\vartheta}_2$, we can define $\widehat{\bar{\beta}_2^0} = \left[\widehat{\vartheta}_2\right]_{[1]}$ as an estimator for the average contemporaneous treatment effect $\bar{\beta}_2^0 = [\vartheta_2]_{[1]}$.

Theorem 4.2 *Suppose that conditions in Theorem 4.1 hold and the bandwidth additionally satisfies that $Nh^4 \to 0$ at a polynomial rate of N. Let $N_s = \sum_i 1_{\{X_{i1} \in \mathcal{X}_1^*\}}$ with $N_s/N \to p \in (0, 1]$ as $N \to \infty$. Then, $\widehat{\vartheta}_2$ satisfies that*

$$\sqrt{N_s}(\widehat{\vartheta}_2 - \vartheta_2) \to_d N(0, \mathbf{\Sigma}_1^* + \mathbf{\Sigma}_2^*),$$

where $\mathbf{\Sigma}_1^* = \mathsf{c}_\kappa \mathbb{E}[\mathbf{\Omega}'(X_{i1}) \mathbb{E}[\tilde{\varepsilon}_{i2}^2 \ddot{Z}_{i2} \ddot{Z}_{i2}' | X_{i1}] \mathbf{\Omega}(X_{i1}) | X_{i1} \in \mathcal{X}_1^*]$, $\mathbf{\Sigma}_2^* = \mathbb{V}[\theta_2(X_{i1})|X_{i1} \in \mathcal{X}_1^*]$ and $\mathsf{c}_\kappa = \int\int \kappa(u)\kappa(u-s) du ds$. If we further assume that $\theta_2(x) = \theta_2$ for all $x \in \mathcal{X}_1^*$, then

$$\sqrt{N_s}(\widehat{\vartheta}_2 - \vartheta_2) \to_d N(0, \mathbf{\Sigma}_1^*).$$

The convergence rate of the average treatment effect is faster than the pointwise convergence rate of the functional coefficient estimator for using all data with $X_{i1} \in \mathcal{X}_1$ rather than only those in a shrinking window defined by the bandwidth h. To achieve this parametric convergence rate, a stronger bandwidth condition is required in the theorem. Similar conditions are also required in the literature. See, for example, Su et al. (2013).

The asymptotic variance of $\widehat{\vartheta}_2$ consists of two terms where one is associated with the estimation error of $\widehat{\theta}_2(.)$ and the other is associated with the heterogeneity of $\theta_2(.)$ on $\mathcal{X}_1^*$. If the function $\theta_2(.)$ is not path-dependent, the second term $\mathbf{\Sigma}_2^*$ degenerates to zero.

The variances $\mathbf{\Sigma}_1^*$ and $\mathbf{\Sigma}_2^*$ are estimable respectively by using the following matrices:

$$\widehat{\mathbf{\Sigma}}_1^* = \widehat{p} \cdot N^{-1} \sum_j \widehat{\varepsilon}_{j2}^2 \widehat{\zeta}(X_{j1}) \ddot{Z}_{j2} \ddot{Z}_{j2}' \widehat{\zeta}'(X_{j1}),$$

$$\widehat{\mathbf{\Sigma}}_2^* = N_s^{-1} \sum_{i: X_{i1} \in \mathcal{X}_1^*} (\widehat{\theta}_2(X_{i1}) - \widehat{\vartheta}_2)(\widehat{\theta}_2(X_{i1}) - \widehat{\vartheta}_2)',$$

where $\widehat{\zeta}(x) = N_s^{-1} \sum_{i: X_{i1} \in \mathcal{X}_1^*} \kappa_h(X_{i1} - x) \widehat{f}_{X_1}^{-1}(X_{i1}) \widehat{\mathbf{\Omega}}'(X_{i1})$ and $\widehat{\varepsilon}_{j2}$ is defined in Section 4.2.

The following proposition states consistency properties of the variance estimators.

Proposition 4.2 *Suppose that the conditions in Theorem 4.2 hold. Then,* $\widehat{\mathbf{\Sigma}}_1^* \to_p \mathbf{\Sigma}_1^*$ *and* $\widehat{\mathbf{\Sigma}}_2^* \to_p \mathbf{\Sigma}_2^*$.

The parametric convergence rate of the proposed average estimator suggests a two-step estimation procedure for the partially linear benchmark model. The first step involves estimating the average effect $\widehat{\vartheta}_2$ and obtaining a modified outcome $Y_{i2}^* = Y_{i2} - H_{i2}' \left[\widehat{\vartheta}_2\right]_{[-1]}$ with the effect of H_{i2} partialed out. The second-step semi-parametric conditional GMM estimation then uses Y_{i2}^* as the outcome, $\ddot{X}_{i2} = (X_{i2}\ 1)'$ as the regressor set, and $\ddot{Z}_{i2} = (\omega_2(Z_{i2})'\ 1)'$ as the instrument set. The parametric convergence rate of the first-step average effect estimator implies that first-step estimation error can be ignored asymptotically in the asymptotic distribution of the second step semi-parametric estimator. As a result, the proposed two-step estimator would enjoy the same asymptotic properties as described in Theorem 4.1 with the second-step definitions of outcome, regressor set, and instrument set.[7]

5 Empirical Application: Path-dependent China Shock Effects

Using China's spectacular, supply-driven export growth as a trade shock, combined with a first-difference IV strategy, Autor et al. (2013) investigated the impacts of import

[7]This estimation procedure is differentiated from the two-step estimation methods proposed in Fan and Huang (2005), Cai et al. (2019), Fan et al. (1998), and Fan and Li (2003). Specifically, when estimating constant coefficients, the estimators suggested by Fan and Huang (2005), Fan and Li (2003), and Cai et al. (2019) use the functional coefficient estimates computed on the entire support of X_{i1} in our notation, while our approach utilizes information from areas where data richness is expected. Given that real-world data, including our empirical application, are often sparsely observed, our two-step estimator, which only employs densely observed data points, is expected to produce more robust estimation results for constant coefficients, thereby improving the pointwise estimation result for the functional coefficient.

penetration on local labor market outcomes. Subsequent studies have adopted or slightly modified its empirical approach and examined the effects of the trade shock in a variety of other contexts, including worker-level outcomes (Autor et al., 2014), industry-level outcomes (Acemoglu et al., 2016), innovation (Autor et al., 2020b) and political outcomes (Autor et al., 2020a). In addition to import competition, Feenstra et al. (2019) revisited the analysis by adding the expansion of US exports to the framework. However, none of the aforementioned studies have yet considered treatment effect dynamics.

In this section, we revisit the China Syndrome application using the proposed model with treatment effect dynamics. We use the industry-level data from Acemoglu et al. (2016).[8] The outcome of interest (Y_{it}) is the annual log employment change over 1991-1999 ($t = 1$) and over 1999-2011 ($t = 2$) in industry i. The endogenous treatment of interest (X_{it}) is the annual change in US exposure to Chinese import over the same two periods in industry i. The external instrument (Z_{it}) is the annual change in exposure to Chinese import in eight other high-income countries defined in Acemoglu et al. (2016) over the same two periods.

Investigating treatment effect dynamics is especially interesting in the China syndrome application because the empirical literature has suggested different channels through which the contemporaneous effect of import competition could be path-dependent. For instance, if an industry was hit hard by Chinese import competition in the first decade, the industry may undergo a structural change and transforms its production process from low-quality to high-quality (Bloom et al., 2016). Consequently, in the second period, the industry exhibiting structural transformation may have better capabilities than other industries with no such transformation in responding to the contemporaneous China import competition. On the other hand, the industry hit hard by Chinese import competition may reduce innovation activities (Autor et al., 2020c). In such a case, the industry experiencing slowdown in innovation may have worse competence in coping with

[8] We use the industry-level data in Acemoglu et al. (2016) rather than the location-level data, which is more popular in the China syndrome literature. This is because external instruments in location-level regressions take a shift-share form, which can cause complications in inference as is explained in Borusyak et al. (2022) and Adão et al. (2019). Meanwhile, Borusyak et al. (2022) also showed that the identification power of shift-share instruments comes exclusively from industry-level shocks used to in their construction.

Chinese import competition than other industries with no such innovation deceleration. That is, either scenario can result in path-dependent contemporaneous treatment effects in the second period. However, to our best knowledge, no such an empirical framework has been proposed to estimate path-dependent contemporaneous treatment effects.

We use three different variations of the control vector ($\widetilde{H}_{it}$): 1) intercept only, 2) a full set of one-digit manufacturing sector fixed effects, and 3) the sector fixed effects as well as all production controls and pre-trend controls considered in Acemoglu et al. (2016), including production workers as a share of total employment, the log average wage, the ratio of capital to value added (in 1991), computer investment as a share of total investment, high-tech equipment as a share of total investment (in 1990), and changes in the log average wage and in the industry's share of total employment over 1976–1991. The data set is a balanced panel of 392 four-digit manufacturing industries over two time periods.

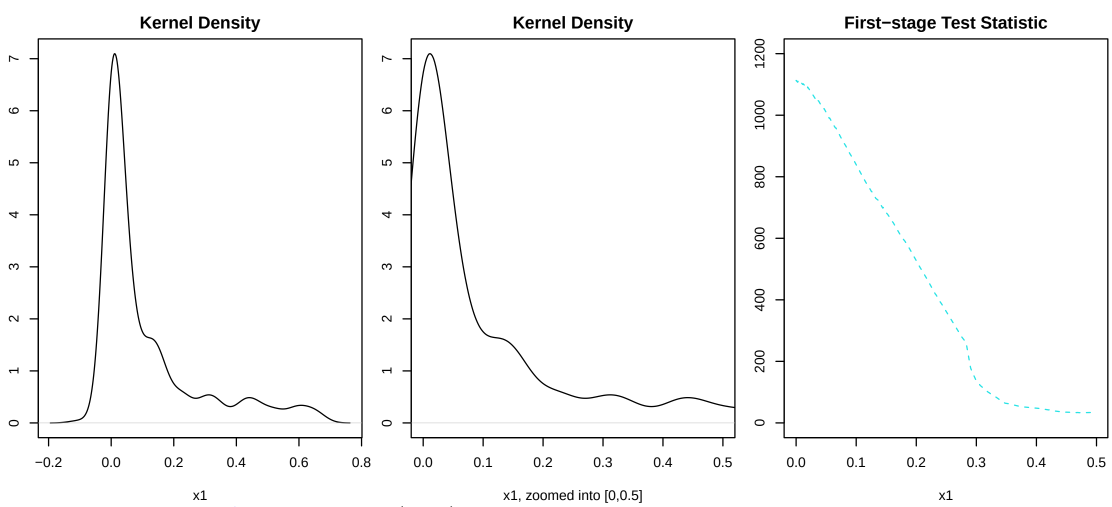
An important first step in studying the potentially path-dependent treatment effect is to determine the initial treatment timing upon which our identification condition is based. In this application, US imports from China had been almost negligible in the 1980s and started to increase in the early 1990s, see Autor et al. (2014, Fig. 1) and Acemoglu et al. (2016, Fig. 2). Therefore, it is reasonable to consider the period from 1991 to 1999 as the first period where the treatment was given to US industries. Furthermore, around the turn of the century, China's accession to the WTO accelerated US imports from China. Given these circumstances, we focus on estimating the potentially path-dependent treatment effect in the second period (1999-2011)—that is, the impact of the China trade shock in the later period (1999-2011) depending on the magnitude of the treatment in the previous period (1991-1999).

Table 1 presents estimation results of several parametric estimators. The estimator *para 1*, as is reported in columns (1), (4), and (7), follows the model in equation (2.1). It is also the existing parametric approach adopted by the empirical literature. The estimator *para 2*, as is reported in columns (2), (5), and (8), is based on the 2SLS regression of Y_2 on (X_2, X_1, X_1X_2) instrumented by (Z_2, X_1, X_1Z_2). The estimator *para 3*, as is reported in columns (3), (6), and (9), is based on the 2SLS regression of Y_2 on (X_2, X_1, X_1X_2) instrumented by (Z_2, Z_1, Z_1Z_2). As is discussed in Section 2.3 as well as through the Monti Carlo simulations in Appendix B, the three parametric estimators rely on different exclusion restrictions.

Besides the difference in exclusion restrictions, columns (1)-(3), (4)-(6) and (7)-(9) in Table 1 differ in the additional control vector in use. Columns (1)-(3) are intercept only. Columns (4)-(6) include sector fixed effects. Columns (7)-(9) include both sector fixed effects and industry production and pretrend controls. The coefficient estimate reported in Column (1) of Table 1, -1.08, is comparable with the coefficient estimate of -1.16 reported in column (7) of Table 2 in Acemoglu et al. (2016). The estimates are not exactly identical to each other because Acemoglu et al. (2016) weigh observations by 1991 employment while we do not use a weighted regression.

Across the columns in Table 1, the contemporaneous effect estimators for $t = 2$ tend to lose statistical significance as more controls are added to the parametric 2SLS regression. In addition, parametric approaches to allowing for direct carryover effect and path-dependency in contemporaneous effect do not seem to be successful. The direct carryover coefficients and the parametric path-dependency coefficients reported in columns (5), (6), (8) and (9) do not have statistical precision.

Figure 1: Distribution of X_1 and First-stage F Statistics



Note: Data are from Acemoglu et al. (2016). Kernel densities reported in the left and middle graphs use the *density* command in *R* and the default bandwidths. The first-stage F test statistics reported in the right graph are from local 2SLS regressions using the specification reported in Column (7) of Table 1 but with observations weighted by kernel weights, calculated using the same kernel function and bandwidth as in the semi-parametric estimations reported in Figure 2.

Table 1: Parametric 2SLS Regression Results

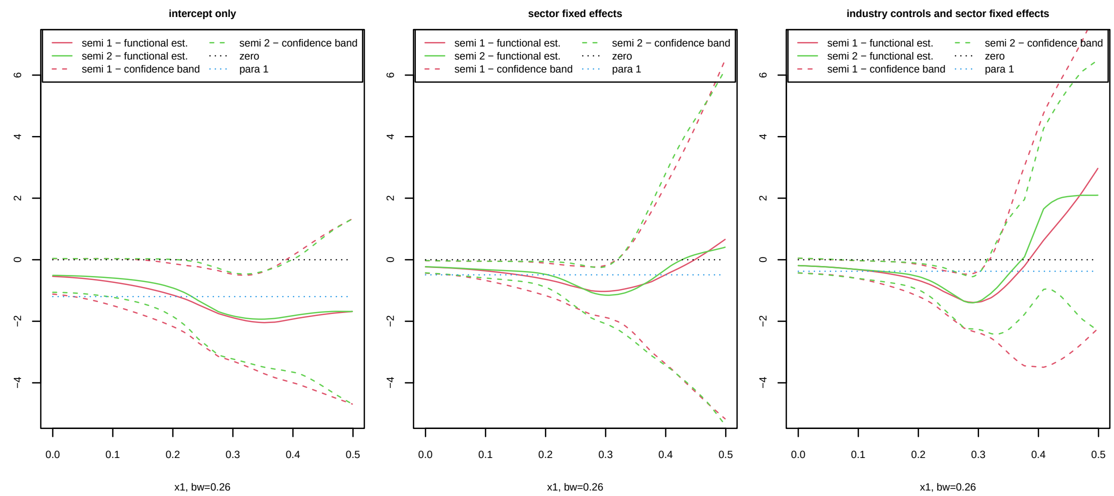
	Intercept Only			Sector FEs Only			Controls and Sector FEs		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	para 1	*para 2*	*para 3*	*para 1*	*para 2*	*para 3*	*para 1*	*para 2*	*para 3*
X_2	-1.08***	-1.01***	-0.37	-0.47**	-0.44*	-0.11	-0.31	-0.31	-0.08
	(0.21)	(0.25)	(0.34)	(0.19)	(0.23)	(0.28)	(0.20)	(0.24)	(0.28)
X_1		-0.99***	-3.26***		-0.22	-1.46		-0.03	-0.92
		(0.38)	(1.13)		(0.33)	(1.04)		(0.33)	(1.20)
X_1X_2		0.13	0.21		0.02	0.05		0.03	0.01
		(0.09)	(0.15)		(0.08)	(0.13)		(0.07)	(0.13)
Sector FEs	N	N	N	Y	Y	Y	Y	Y	Y
Controls	N	N	N	N	N	N	Y	Y	Y
F-S Test	742	203	12	621	201	10	574	198	7
N	392	392	392	392	392	392	392	392	392

Note: Data are from Acemoglu et al. (2016). Parametric 2SLS regressions are carried out with *Stata*. F-S Test is the minimum eigenvalue statistic reported following the *ivregress 2sls* command in *Stata*. *, ** and *** indicate significance at 10%, 5%, and 1% level, respectively.

Next, we use the proposed semi-parametric method to estimate the potentially path-dependent contemporaneous effect of Chinese import competition. Figure 2 reports functional estimates of $\beta_2^0(x_1)$ evaluated at $x_1 \in [0, 0.5]$, for industries that experienced a Chinese import exposure rise of 0-0.5 percentage points per year over the first decade, i.e., 1991-1999. Given the peculiar heavy right-skewness (see Figure 1), the X_1 range we report in Figure 2 includes over 80% of the industries. The two semi-parametric estimators reported in Figure 2, *semi 1* and *semi 2*, are both just-identified. They differ in whether to allow for a local linear expansion of the control vector's coefficient functions. Details are discussed in the simulation section in Appendix B. The two semi-parametric estimators gives very similar results as is seen in Figure 2 and lead to the same empirical findings.

We see from Figure 2 that all semi-parametric estimators report a negative second-period contemporaneous effect for $X_1 \in [0, 0.3]$. The size of the negative effect tends to decrease as more control variables are introduced to the model. Nonetheless, statistical significance is preserved for $X_1 \in [0, 0.3]$, as additional controls soak up variations in the

Figure 2: Second-period Contemporaneous Treatment Effect Estimates



Note: Data are from Acemoglu et al. (2016). Semi-parametric kernel function and bandwidth are chosen following the rule-of-thumb discussed in Appendix B with $\varrho = 3.5$. Robustness checks with $\varrho = 3.25$ and $\varrho = 3.75$ are reported in the appendix and see similar empirical results. The blue dotted line of *para 1* in the left/middle/right graph corresponds to the estimates reported in columns (1)/(4)/(7) of Table 1, respectively.

error term of the outcome equation. The lack of statistical precision after X_1 exceeds around 0.3 is due to sparse data, and is expected from both the density graph and the the first-stage F test statistics graph in Figure 1.

The most interesting empirical finding of Figure 2 is the downward-sloping functional form of the path-dependent contemporaneous China shock effect, shared by all three graphs of Figure 2. The results indicate that previous decade's Chinese import exposure magnifies the negative impact of the current decade's Chinese import exposure on employment. The magnifying effect is seen to be fairly small if the previous decade's Chinese import exposure is relatively mild but becomes much larger (i.e., steeper slope) when X_1 is larger than around 0.2.

To conclude the empirical section, we report the average contemporaneous treatment effect estimates in Table 2. The first row of the table reports semi-parametric average contemporaneous effect estimates integrated over the X_1 range of $[0, 0.3]$, where the functional estimates reported in Figure 2 are quite precisely estimated. Compared to parametric estimates reported in columns (1), (4), and (7) of Table 1, the semi-parametric average effect estimates are around 40% smaller when no control variables are considered

Table 2: Average Contemporaneous China Shock Effect in 1999-2011

	Intercept Only		Sector FEs Only		Controls and Sector FEs	
	(1)	(2)	(3)	(4)	(5)	(6)
	semi 1	*semi 2*	*semi 1*	*semi 2*	*semi 1*	*semi 2*
$\mathcal{X}_1^* \in [0, 0.3]$	-0.68**	-0.60**	-0.33***	-0.30***	-0.31**	-0.29**
	(0.29)	(0.27)	(0.11)	(0.11)	(0.13)	(0.14)
$\mathcal{X}_1^* \in [0, 0.2]$	-0.62**	-0.55*	-0.29***	-0.27**	-0.25**	-0.24*
	(0.30)	(0.28)	(0.11)	(0.11)	(0.13)	(0.13)
$\mathcal{X}_1^* \in [0.2, 0.3]$	-1.49***	-1.35***	-0.85***	-0.79***	-1.04***	-0.96***
	(0.53)	(0.52)	(0.25)	(0.29)	(0.32)	(0.35)

Note: Data are from Acemoglu et al. (2016). Semi-parametric kernel function and bandwidth are chosen following the rule-of-thumb discussed in Appendix B with $\varrho = 3.5$. Robustness checks with $\varrho = 3.25$ and $\varrho = 3.75$ are reported in the appendix and see similar empirical results. *, ** and *** indicate significance at 10%, 5%, and 1% level, respectively.

and 30% smaller when sector fixed effects are considered. In terms of statistical significance, all semi-parametric results reported in Table 2 are statistically significant. The parametric regression results (see Table 1), on the other hand, lose statistical significance as controls and fixed effects are added into the model.

The second and third rows of Table 2 report contemporaneous effect estimates averaged over the X_1 range of $[0, 0.2]$ and $[0.2, 0.3]$, respectively. Across all three model specifications, the contemporaneous China shock effect in 1999-2011 is much larger in the third row, i.e., when an industry's China shock exposure in the last decade is over 0.2 percentage points per year. In particular, when all production controls, pre-trend controls, and sector fixed effects are controlled, estimates in columns (5) and (6) suggest that a 1 percentage point increase in industry import penetration reduces domestic industry employment by about 0.25 percentage point when averaged over $X_1 \in [0, 0.2]$, and about 1 percentage point when averaged over $X_1 \in [0.2, 0.3]$. – The contemporaneous China shock effect in 1999-2011 is four folds as large for industries exposed to substantial China shock in the past decade, compared to industries exposed to small or moderate shock in the 90's.

The bigger contemporaneous effect estimates of industries exposed to larger earlier shocks underscore the importance of path-dependency in the contemporaneous treatment effect in the 2000s. While our goal is not to uncover an underlying mechanism behind the magnifying effect, the result could be cautiously interpreted as evidence of decreased innovation activities caused by an earlier exposure to Chinese import penetration during the 1990s. For example, industries significantly affected by Chinese import penetration might have decreased their innovation activities in the first period (Autor et al., 2020c), which further weakens their ability to cope with Chinese import competition in the following period.

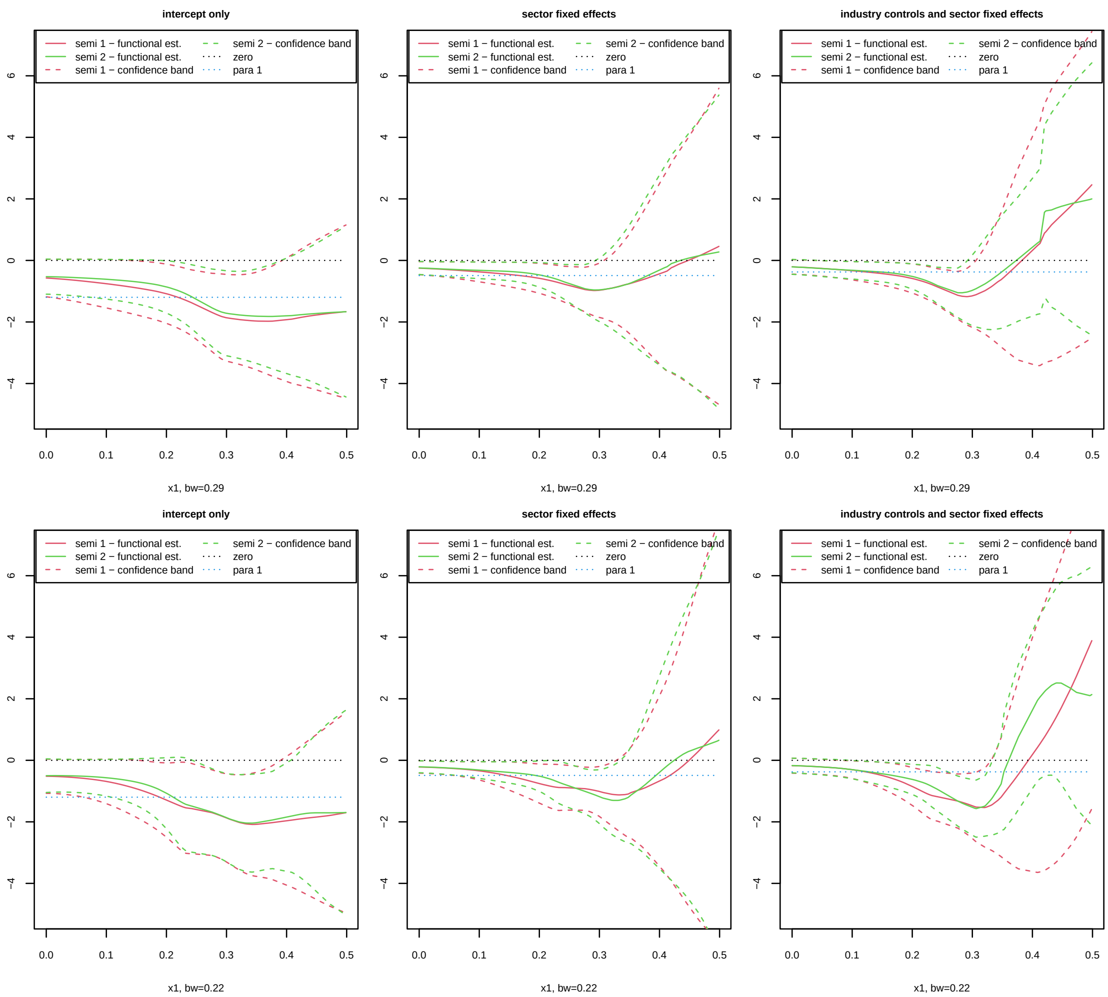
6 Conclusion

In the paper, we propose a new panel IV model featuring treatment effect dynamics. Specifically, our new model allows for a direct carryover effect of the preceding treatment and path-dependency in the contemporaneous treatment effect. We show that in the presence of treatment effect dynamics, existing textbook 2SLS estimators become inconsistent if external instruments are serially correlated. To address this issue, we propose a novel semi-parametric identification and estimation procedure and study asymptotic properties of the suggested estimators. We show that the proposed estimators have satisfactory small sample performance. When applied to revisit the seminal study by Acemoglu et al. (2016) on the China syndrome, our proposed method reveals important empirical findings that have not been discovered previously. In particular, we find that the contemporaneous impact of increased Chinese import competition on US manufacturing employment is magnified by the import exposure in the preceding decade. The size of the magnifying effect is mild if the last decade's import exposure was small or moderate. But the interaction between the past and current trade shocks becomes much more significant when the import exposure over the last decade exceeds 0.2 percentage points per year.

Appendix A: Robustness Checks for Empirical Analysis

In this section, we report robustness checks of the empirical results using alternative bandwidths for semi-parametric estimation. Figure 3 and Table 3 show that empirical findings drawn in Section 5 are not sensitive to alternative bandwidths choices.

Figure 3: Second-period Contemporaneous Treatment Effect Estimates



Note: Data are from Acemoglu et al. (2016). Semi-parametric kernel function and bandwidth are chosen following the rule-of-thumb discussed in Appendix B with $\varrho = 3.25$ in the top panel and $\varrho = 3.75$ in the bottom panel.

Table 3: Robustness Checks with Alternative Bandwidths

	Intercept Only		Sector FEs Only		Controls and Sector FEs	
	(1)	(2)	(3)	(4)	(5)	(6)
	semi 1	*semi 2*	*semi 1*	*semi 2*	*semi 1*	*semi 2*
	Panel B: $\kappa = 3.25$					
$\mathcal{X}_1^* \in [0, 0.3]$	-0.70**	-0.61**	-0.34***	-0.31***	-0.31**	-0.29**
	(0.31)	(0.29)	(0.12)	(0.12)	(0.13)	(0.14)
$\mathcal{X}_1^* \in [0, 0.2]$	-0.65**	-0.57*	-0.30***	-0.28**	-0.26**	-0.26*
	(0.32)	(0.29)	(0.12)	(0.12)	(0.13)	(0.14)
$\mathcal{X}_1^* \in [0.2, 0.3]$	-1.43***	-1.24***	-0.78***	-0.71***	-0.88***	-0.80***
	(0.46)	(0.45)	(0.23)	(0.26)	(0.27)	(0.31)
	Panel B: $\kappa = 3.75$					
$\mathcal{X}_1^* \in [0, 0.3]$	-0.67**	-0.60**	-0.32***	-0.29***	-0.31**	-0.28**
	(0.27)	(0.27)	(0.11)	(0.11)	(0.14)	(0.14)
$\mathcal{X}_1^* \in [0, 0.2]$	-0.60**	-0.54**	-0.28**	-0.25**	-0.24*	-0.23*
	(0.28)	(0.27)	(0.11)	(0.11)	(0.13)	(0.14)
$\mathcal{X}_1^* \in [0.2, 0.3]$	-1.58**	-1.49**	-0.89***	-0.81***	-1.18***	-1.01***
	(0.62)	(0.61)	(0.27)	(0.30)	(0.36)	(0.38)

Note: Data are from Acemoglu et al. (2016). Semi-parametric kernel function and bandwidth are chosen following the rule-of-thumb discussed in Appendix B with $\varrho = 3.25$ in panel A and $\varrho = 3.75$ in panel B. *, ** and *** indicate significance at 10%, 5%, and 1% level, respectively.

Appendix B Monte Carlo Simulations

In this section, we study the small sample performance of the proposed semi-parametric varying coefficient estimator and the average effect estimator whose asymptotic properties are studied respectively in Sections 4.2 and 4.3. We focus on a two-period model. Throughout simulations in this section, first-period random variables are generated as:

$$(e, v, \eta, \psi, \epsilon) \sim_{i.i.d.} N(0, 0.2^2), \quad X_1 = \psi + \nu, \quad Z_1 \sim 0.5X_1 + \epsilon,$$

where $\nu \sim_{i.i.d.}$ *exponential*(1) and independent of (e, v, η, ψ). This simulation setup is designed for a right-skewed distribution of X_1, as is observed in our empirical application in Section 5. The distribution of X_1 is plotted in the right graph of Figure 4. In both simulation and empirical sections of the paper, we set $\mathcal{X}_1^*$ to $[0, 0.5]$, where data are relatively densely observed.

For the second period, we consider various DGPs, starting from a baseline DGP obtained by fitting parametric OLS regressions to the empirical dataset. Across all DGPs, we consider two sample sizes: 250 and 1,000, and report the estimated contemporaneous effect functions averaged across 1,000 simulations.

We conduct two variations of the semi-parametric estimation approach discussed in Section 4.1 and report their performance. Recall that the instrument set $\ddot{Z}_t = (\omega_t(Z_t)'\ H_t')'$. For the purpose of estimating $\beta_2(x_1)$, for all $x_1 \in [0, 0.5]$, the first semi-parametric estimator (*semi 1*) defines $\omega_t(Z_2) = Z_2$ and $H_2 = 1$. The second semi-parametric estimator (*semi 2*) keeps $\omega_t(Z_2) = Z_2$ but defines $H_2 = (1\ X_1 - x)'$ to allow for a better approximation of the intercept function.

We compare these two semi-parametric estimators to parametric estimators employed in the literature. Specifically, we consider three parametric estimators of $\beta_2^0(.)$. The first parametric estimator (*para 1*) is based on model *existing* or equation (2.1). The second (*para 2*) is based on the 2SLS regression of Y_2 on (X_2, X_1, X_1X_2) instrumented by (Z_2, X_1, X_1Z_2), while the last (*para 3*) is based on the 2SLS regression of Y_2 on (X_2, X_1, X_1X_2) instrumented by (Z_2, Z_1, Z_1Z_2). As is discussed in Section 2.3, the three parametric estimators rely on different exclusion restrictions.

For semi-parametric estimation, under-smoothing is required for the average estimator to ensure satisfactory asymptotic properties. Following Chernozhukov et al. (2013), we use the rule-of-thumb bandwidth:

$$h = \widehat{h}_{ROT} \times \widehat{s} \times N^{1/5-1/\varrho},$$

where $\widehat{s}$ is the standard deviation of X_1 and ϱ is an under-smoothing tuning parameter. $\widehat{h}_{ROT}$ minimizes the weighted Mean Integrated Square Error of the local linear estimation of Y_2 on studentized X_1. We also follow Chernozhukov et al. (2013) to use the quartic kernel function (i.e., $\kappa(s) = 15/16(1-s^2)^2 \cdot 1_{\{|s|\leq 1\}}$) and set the value of ϱ to 3.5. In both simulation and empirical sections, we also run simulations with $\varrho = 3.25$ and $\varrho = 3.75$ as a robustness check.

As is discussed earlier, the baseline DGP is estimated from the empirical dataset of

Acemoglu et al. (2016) using OLS regressions. Key features of the DGP is reported in Figure 5. It is worth noting that since the OLS ignores potential endogeneity in the model, relationships depicted in Figure 5 shall not be compared with the causal analysis conducted in Acemoglu et al. (2016) or in our revisit reported in Section 5.

DGP A: the baseline model

$$
\begin{aligned}
&Z_2 = 0.420 + 1.182 Z_1 + e,\\
&X_2 = 0.083 + 0.156 X_1 + (0.82 + 0.002 X_1) Z_2 + v,\\
&Y_2 = -3.415 - 2.461 X_1 + 0.259 X_1^2 + (-0.916 + 0.636 X_1 - 0.095 X_1^2) X_2 + u,\\
&\quad \text{with } u = 0.6\eta + 0.6v.
\end{aligned}
$$

Under the baseline DGP, X_2 is an endogenous regressor for the outcome equation of Y_2, while X_1 is exogenous. In addition, both Z_1 and Z_2 are valid instruments under this DGP.

Figure 4: The Baseline Data Generating Process (DGP A)

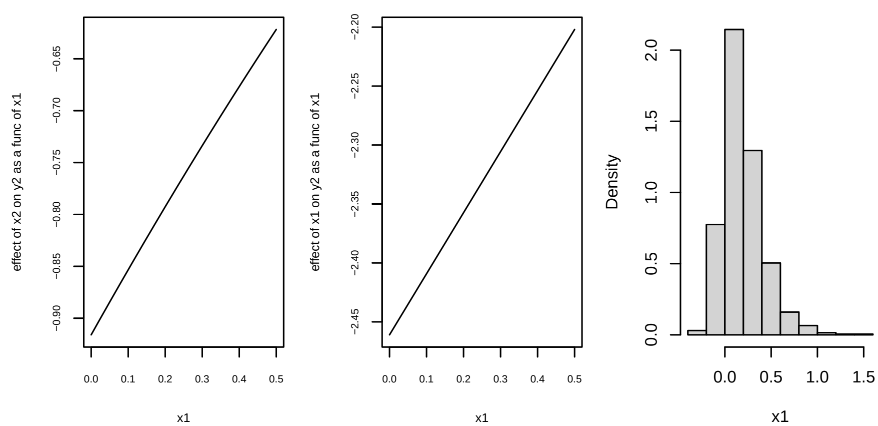
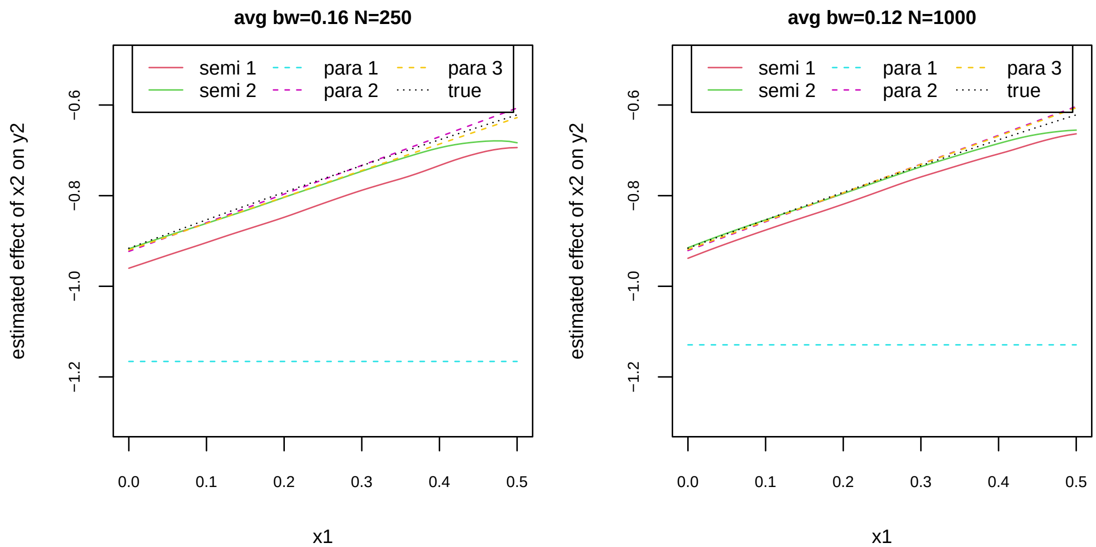


Figure 4 shows that under the baseline DGP, both the carryover and contemporaneous treatment effects at $t = 2$ are negative, while their magnitudes (in absolute values) tend to decrease as X_1 gets larger. Moreover, within the considered range of X_1, curvatures of the effect functions are hardly noticeable despite their quadratic functional forms, implying that both *para 2* and *para 3* are close to be correctly specified.

Figure 5 compares the performance of various contemporaneous effect estimators. The left graph is for $N = 250$ and the right is for $N = 1,000$. We see that the estimator *para 1*, which is popular in empirical studies, deviates dramatically from the true con-

Figure 5: Simulation Results, DGP A



Note: Simulations are carried out 1,000 times. Semi-parametric kernel function and bandwidth are chosen following the rule-of-thumb with $\varrho = 3.5$.

temporaneous effect function. The other two parametric estimators, *para 2* and *para 3*, perform reasonably well, which is expected by their almost correct model specifications under the baseline DGP.

The semi-parametric estimator *semi 2* performs comparably to *para 2* and *para 3* in regions with higher density of X_1 (i.e., $X_1 \in [0, 0.3]$). Meanwhile, under the baseline DGP, estimation errors of the other semi-parametric estimator *semi 1* are non-negligible but improve dramatically with the increase of the sample size.

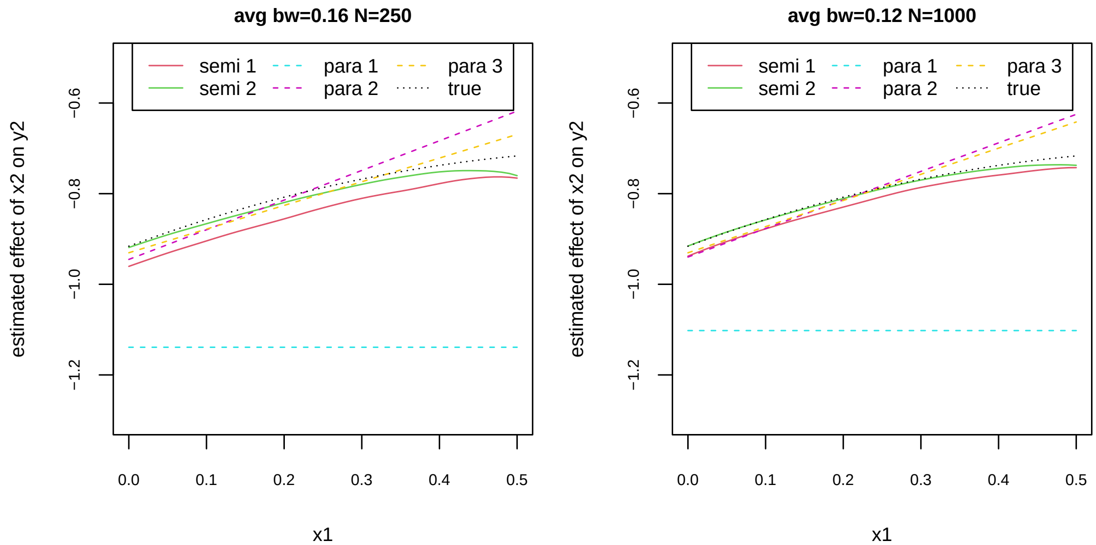
Next, we modify the second-period contemporaneous effect function to allow it to have an exaggerated curvature within the considered range. Under the new DGP, all parametric estimators are expected to perform poorly due to substantial model misspecifications.

DGP B: the model with an exaggerated curvature

$$Y_2 = -3.415 - 2.461X_1 + 1.295X_1^2 + (-0.916 + 0.636X_1 - 0.475X_1^2)X_2 + u.$$

Figure 6 summarizes estimation results for DGP B. First we notice that none of the parametric estimators can reproduce the quadratic true effect function, regardless of the

Figure 6: Simulation Results, DGP B



Note: Simulations are carried out 1,000 times. Semi-parametric kernel function and bandwidth are chosen following the rule-of-thumb with $\varrho = 3.5$.

sample size. The estimator *semi 2* continues to report simulation averages very close to the true contemporaneous effect function. The other semi-parametric estimator *semi 1* also captures the curvature of the second-period contemporaneous effect function well, although its simulation average continues to show a noticeable deviation from the true function when the sample size is small. When $N = 1,000$, *semi 1* performs substantially better and outperforms all parametric estimators.

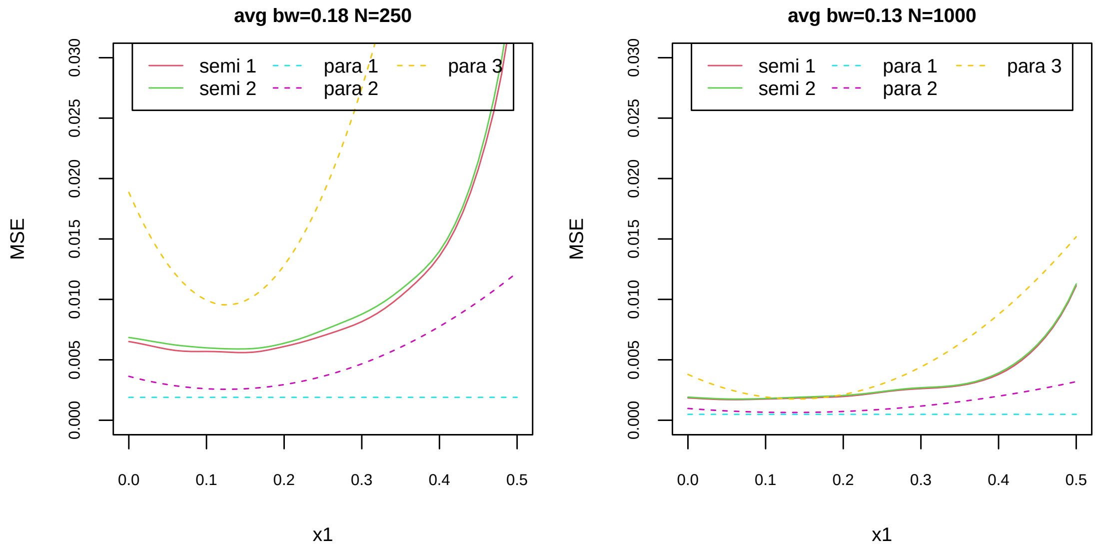
The third DGP modifies DGP A in the direction opposite to that considered in DGP B. Specifically, we explore the performance of the estimators when there is no path-dependency in the contemporaneous treatment effect. Under this scenario, all parametric and semi-parametric estimators are consistent. Hence, instead of comparing estimated curves, we compute the empirical mean squared error (MSE) of each estimator to make a more meaningful comparison of their finite sample performances.

DGP C: no carryover effect or path-dependency in contemporaneous effect

$$Y_2 = -3.415 - 0.916X_2 + u.$$

Figure 7 reports the empirical MSEs of each estimator as a function of X_1. Not sur-

Figure 7: Simulation Results, DGP C (Empirical MSEs)



Note: Simulations are carried out 1,000 times. Semi-parametric kernel function and bandwidth are chosen following the rule-of-thumb with $\varrho = 3.5$. Robustness results with alternative bandwidths are reported in the appendix.

prisingly, the simplest parametric estimator, *para 1*, gives the smallest MSE, since its model is correctly specified under this DGP. The two kernel-based semi-parametric estimators have efficiency loss, especially in low density regions, for only using data within local estimation windows. The most interesting finding of Figure 7 is that the parametric estimator *para 3* performs worse, in terms of empirical MSE, than the semiparametric estimators for both simulation sample sizes. This unsatisfactory small sample performance of *para 3* speaks about its demanding rank condition arose from the use of three endogenous regressors.

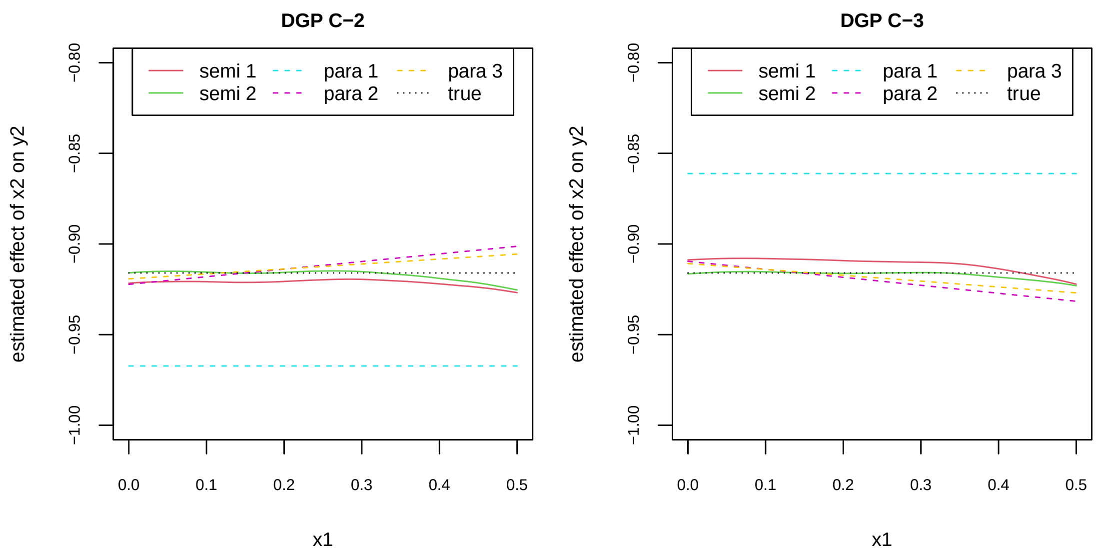
Last but not least, we consider a scenario where exclusion restrictions of external instruments are valid *ONLY* after conditioning on the endogenous treatment from the last period. Specifically, we consider the following two DGPs with the outcome equation being the same as the one in DGP C.

DGPs C-2, C-3: no sequential exogeneity

$$\text{DGP C-2:} \quad u = 0.6\eta + 0.6v + 0.5\psi,$$

$$\text{DGP C-3:} \quad u = 0.6\eta + 0.6v - 0.5\psi.$$

Figure 8: Simulation Results DGPs C-2 and C-3



Note: Simulations are carried out 1,000 times. $N = 1,000$. Semi-parametric kernel function and bandwidth are chosen following the rule-of-thumb with $\varrho = 3.5$.

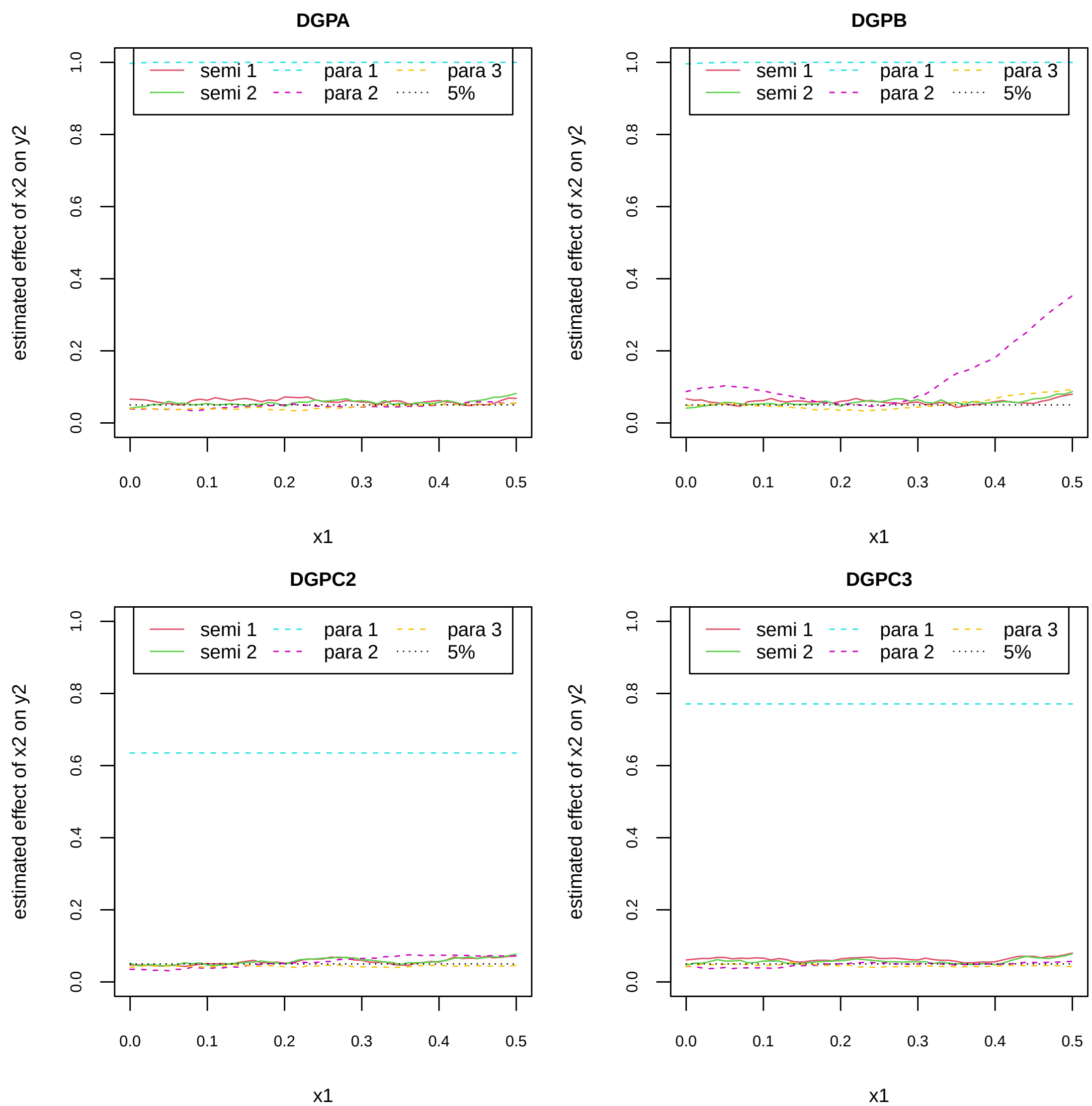
DGPs C-2 and C-3 keep the underlying model of DGP C except for modifying the construction rule of the outcome error u. Under the new DGPs, the unconditional exclusion restriction $\mathbb{E}[Z_2 u_2] = 0$ is invalid due to the nontrivial correlation between u and ψ.[9] As a result, all three parametric estimators are inconsistent. On the other hand, the conditional exclusion restriction required for our semi-parametric estimation method is satisfied.[10] The simulation results are summarized in Figure 8. Similar to the case of DGP B, only the semi-parametric estimators, in particular *semi 2* and in regions with higher density of X_1 (i.e., $X_1 \in [0, 0.3]$), shows reasonable small sample performance.

To investigate the small sample performance of the proposed inference procedure described in Proposition 4.1, we carry out pointwise t-tests following each parametric and semi-parametric estimation method. Figure 9 reports the rejection proportion of each test under various DGPs.

[9] Specifically, under DGP C-2, $\mathbb{E}[Z_2 u_2] = \mathbb{E}[(0.420 + 1.182 \cdot (0.5(\psi + \nu) + \epsilon) + e)(0.6\eta + 0.6v + 0.5\psi)] = 1.182 \cdot 0.25\mathbb{E}[\psi^2] \neq 0$, and similarly under DGP C-3, $\mathbb{E}[Z_2 u_2] = -1.182 \cdot 0.025\mathbb{E}[\psi^2] \neq 0$.

[10] Specifically, given the independence between the ϵ,e,ψ, and ν, we have $\mathbb{E}[u_2|Z_2, X_1] = \mathbb{E}[0.6\eta + 0.6v + 0.5\psi|Z_2, X_1] = \mathbb{E}[0.5\psi|Z_2, X_1] = E[0.5\psi|1.182\epsilon + e, X_1] = E[0.5\psi|X_1]$.

Figure 9: Size Control, DGPs A, B, C-2, and C-3



Note: Simulations are carried out 1,000 times. $N = 1,000$. Semi-parametric kernel function and bandwidth are chosen following the rule-of-thumb with $\varrho = 3.5$.

Testing results reported in Figure 9 are in line with simulation results reported in Figures 5, 6, and 8 for estimator properties. The *existing* parametric method, or *para 1*, popularly adopted in the empirical literature, performs the worst across all DGPs A, B, C-2, and C-3. The estimator *para 2* performs much better than *para 1*, although it can still lose size control under unfavorable DGPs. The most sophisticated parametric estimator, *para 3*, shows much more robust size performance in Figure 9 except for some small distortions under DGP B. However, as is already shown in Figure 7, this more flexible parametric estimator takes on huge efficiency losses in small samples. In contrast, both semi-parametric estimators, *semi 1* and *semi 2*, have good pointwise size control for the reported range of X_{i1} values.

Table 4: Small Sample Performance of the Average Effect Estimator

		N=250			N=1,000		
	True	*Para 1*	*Semi 1*	*Semi 2*	*Para 1*	*Semi 1*	*Semi 2*
				DGP A			
Avg. Est.	-0.782	-1.166	-0.834	-0.794	-1.129	-0.808	-0.786
RMSE	-	0.395	0.081	0.061	0.350	0.040	0.030
T-Test Size	0.050	0.785	0.090	0.033	0.997	0.095	0.037
				DGP B			
Avg. Est.	-0.807	-1.139	-0.851	-0.819	-1.102	-0.828	-0.810
RMSE	-	0.343	0.075	0.060	0.298	0.036	0.030
T-Test Size	0.050	0.765	0.070	0.033	0.996	0.068	0.037
				DGP C			
Avg. Est.	-0.916	-0.916	-0.922	-0.923	-0.916	-0.917	-0.917
RMSE	-	0.044	0.057	0.058	0.022	0.029	0.029
T-Test Size	0.050	0.043	0.040	0.036	0.043	0.039	0.037
				DGP C-2			
Avg. Est.	-0.916	-0.972	-0.928	-0.920	-0.967	-0.921	-0.916
RMSE	-	0.071	0.054	0.054	0.056	0.027	0.027
Test Size	0.050	0.209	0.040	0.037	0.635	0.035	0.034
				DGP C-3			
Avg. Est.	-0.916	-0.857	-0.912	-0.922	-0.861	-0.910	-0.917
RMSE	-	0.070	0.050	0.052	0.058	0.026	0.026
Test Size	0.050	0.338	0.045	0.039	0.771	0.043	0.039

Note: Simulations are carried out 1,000 times. Semi-parametric kernel function and bandwidth are chosen following the rule-of-thumb with $\varrho = 3.5$.

Before concluding this section, we study the small sample performance of the semi-parametric average estimator proposed in Section 4.3. Small sample performances are measured by the average estimator across simulations, root MSE (RMSE), and the size of associated t-test with simulation results reported in Table 4. We see that proposed semi-parametric average estimators based on both *semi 1* and *semi 2* perform significantly better than the *existing* parametric estimator *para 1* except under DGP C. Under DGP C, however, where the true model is free of any treatment effect dynamic features, the existing parametric estimator *para 1* has better small sample performance. Among the two semi-parametric average estimators, the one based on *semi 2* performs better in small samples, which is in line with results shown in the rest of the simulation section.

Appendix C: Parametric Identification

Without loss of generality, we discuss the case $T = 2$ for parametrically identifying the *benchmark* model (2.2).

Let $\boldsymbol{\nu}_{i,s:t} = (\nu'_{is} \ \ldots \ \nu'_{it})'$ denote the random vector that stacks $\nu_{i\ell}$ from period s to period t. Let $\vee$ denote the larger and $\wedge$ the smaller of two numbers. Let $f_{it} = \omega_t(\mathbf{Z}_{i,[(t-s)\vee 1]:t})$ be the d_{ft}-dimensional vector generated by external instruments from period $[(t-s)\vee 1]$ to period t with known function $\omega_t(.)$. Then unknown parameters in model (2.2) are identified through classic parametric 2SLS or GMM estimation strategies under the following assumptions. For $t = 1$, the enodgeneous regressor is X_{i1} and the external instrument set is f_{i1}. For $t = 2$, the enodgeneous regressor is $\psi_{i,d_\beta} X_{i2}$ and the external instrument set is f_{i2}.

Assumption C.1 (parametric identification) *Assume that*

1. *(known functional form)* $\beta_2^0(X_{i1}) = \psi'_{i,d_\beta}\eta_{d_\beta}$, *where* $\psi_{i,d_\beta} = (1 \ \psi_2(X_{i1}) \ \ldots \ \psi_{d_\beta}(X_{i1}))'$ *and* η_{d_β} *is a* d_β*-dimensional parameter vector;*

2. *(exclusion restriction)* $\mathbb{E}[\varepsilon_{it}|\mathbf{Z}_{i,[(t-s)\vee 1]:t}, \widetilde{H}_{it}] = 0$*, for* $s = 0, 1$ *and* $t = 1, 2$*;*

3. *(rank condition)* $\mathbb{E}[(f'_{i1} \ H'_{i1})'(X'_{i1} \ H'_{i1})]$ *and* $\mathbb{E}[(f'_{i2} \ H'_{i2})'(X'_{i1} \ (\psi_{i,d_\beta}X_{i2})' \ H'_{i2})]$ *are both of full rank.*

Assumption C.1.1 is a standard parametric functional form assumption. Assumptions C.1.2 and C.1.3 are standard exclusion restriction and rank condition for parametric IV regressions. If $s = 1$, the assumption reduces to $E[\varepsilon_{i1}|Z_{i1}, \widetilde{H}_{i1}] = 0$ and

$E[\varepsilon_{i2}|Z_{i1}, Z_{i2}, \widetilde{H}_{i2}] = 0$. If pre-intervention periods of the external instrument (e.g., Z_{i0}, $Z_{i(-1)}$) are observed, Assumption C.1.2 could also be modified to utilize such information.

Appendix D: Additional Assumptions and Some Useful Lemmas

The following assumption is required for the asymptotic results stated in Section 4.

Assumption D.1 *(a) The observations $\{Y_{i2}, X_{i1}, X_{i2}, Z_{i2}, \widetilde{H}_{i2}\}_{i=1}^N$ are i.i.d.*

(b) The density function $f_{X_1}(.)$ of X_{i1} is twice continuously differentiable with bounded derivatives and bounded away from zero on $\mathcal{X}_1$, the compact support of X_{i1}.

(c) The function $\theta_2(.)$ is three times continuously differentiable on $\mathcal{X}_1$.

(d) The kernel function $\kappa(.)$ is a symmetric density function with compact support.

(e) The matrix $\Lambda_{\ddot{Z}\ddot{X}}(.)$ is twice continuously differentiable on $\mathcal{X}_1$, and $\mathbb{E}[\widetilde{\varepsilon}_{i2}^2 \ddot{Z}_{i2}\ddot{Z}_{i2}' | X_{i1} = .]$ is Lipschitz continuous on $\mathcal{X}_1$.

(f) There exists an $s > 2$ such that $\sup_{x\in\mathcal{X}_1} \mathbb{E}[\|\ddot{Z}_{i2}\|^{2s}|X_{i1} = x] < \infty$, $\sup_{x\in\mathcal{X}_1} \mathbb{E}[|X_{i2}|^{2s}|X_{i1} = x] < \infty$ and $\sup_{x\in\mathcal{X}_1} \mathbb{E}[|Y_{i2}|^{2s}|X_{i1} = x] < \infty$ and $N^{2\delta-1}h \to \infty$ for some $\delta < 1-s^{-1}$.

(g) $h \to \infty$, $Nh^3 \to \infty$, $Nh^7 \to 0$ as $N \to \infty$.

Next, we define some useful notations. Let ι_j be a $(\ell+1)$-dimensional vector whose j-th element is one and all the rest elements are non-zero. Let $\widetilde{X}_{i1}'(x) = (1\ (X_{i1} - x)\ \ldots (X_{i1}-x)^\ell)$. For all values of x, the ℓ-th order local polynomial estimator of $\widehat{\mathbf{\Lambda}}_{\ddot{Z}Y}(x)$ is defined as

$$\widehat{\mathbf{\Lambda}}_{\ddot{Z}Y}(x) = N^{-1}\sum_i \kappa_h(X_{i1} - x)\ddot{Z}_{i2}Y_{i2}\widetilde{X}_{i1}'(x)\widehat{\mathbf{M}}(x)^{-1}\iota_1,$$

with $\widehat{\mathbf{M}}(x) = N^{-1}\sum_i \kappa_h(X_{i1} - x)\widetilde{X}_{i1}(x)\widetilde{X}_{i1}'(x)$. Similarly, the l-th column of the local polynomial estimator $\widehat{\mathbf{\Lambda}}_{\ddot{Z}\ddot{X}}(x)$ is defined as

$$\left[\widehat{\mathbf{\Lambda}}_{\ddot{Z}\ddot{X}}(x)\right]_{.,l} = N^{-1}\sum_i \kappa_h(X_{i1} - x)\ddot{Z}_{i2}\ddot{X}_{i2,l}\widetilde{X}_{i1}'(X)\widehat{\mathbf{M}}(x)^{-1}\iota_1,$$

where $\ddot{X}_{i2,l}$ is the l-th element of $\ddot{X}_{i2}$.

To derive asymptotic properties of the above defined estimators, we first introduce several matrix notations. Let $\widetilde{\mathbf{X}}_1(x)$ be a $N \times (\ell+1)$ matrix whose i-th row is given by $\widetilde{X}'_{i1}(x)$ and $\mathbf{K}(x)$ be a $N \times N$ diagonal matrix whose diagonal entries are given by $\{\kappa_h(X_{i1}-x)\}_{i=1}^N$. For notational convenience, we hereafter let $d_{\ddot{Z}}$ and $d_{\ddot{X}}$ denote the dimensions of $\ddot{Z}_{i2}$ and $\ddot{X}_{i2}$. Moreover, let $\widetilde{Y}_{i2} = \ddot{Z}_{i2}Y_{i2}$ and $\widetilde{X}_{i2} = \mathrm{vec}(\ddot{Z}_{i2}\ddot{X}'_{i2})$ and $\widetilde{\mathbf{Y}}_2$ and $\widetilde{\mathbf{X}}_2$ be $N \times d_{\ddot{Z}}$ and $N \times d_{\ddot{Z}}d_{\ddot{X}}$ matrices collecting the vectors $\widetilde{Y}_{i2}$ and $\widetilde{X}_{i2}$. Then,

$$
\begin{aligned}
\widehat{\mathbf{M}}(x) &= N^{-1}\widetilde{\mathbf{X}}'_1(x)\mathbf{K}(x)\widetilde{\mathbf{X}}_1(x), \\
\widehat{\mathbf{\Lambda}}_{\ddot{Z}Y}(x) &= N^{-1}\widetilde{\mathbf{Y}}'_2\mathbf{K}(x)\widetilde{\mathbf{X}}_1(x)\widehat{\mathbf{M}}(x)^{-1}\iota_1, \\
\mathrm{vec}(\widehat{\mathbf{\Lambda}}_{\ddot{Z}\ddot{X}})(x) &= N^{-1}\widetilde{\mathbf{X}}'_2\mathbf{K}(x)\widetilde{\mathbf{X}}_1(x)\widehat{\mathbf{M}}(x)^{-1}\iota_1.
\end{aligned}
$$

For each integer j, we let $\mathbf{M}_j = (\mu_{i+k+j-2})_{1\le i,k\le \ell+1}$ with $\mu_k = \int u^k\kappa(u)du$ for an integer k. Let $\mathbf{M} \equiv \mathbf{M}_0$. It is easy to see that and $\mathbf{M}_j\iota_s = \mathbf{M}_{j-1}\iota_{s+1}$ for $s = 1,\ldots,\ell$. The matrix $\mathbf{D}$ is defined by a $(\ell+1)\times(\ell+1)$ diagonal matrix whose diagonal entries are given by $\{1\ h\ \cdots h^\ell\}$. Moreover, we let $\mathbf{L}_{\ddot{Z}\ddot{X}}(x)$ be a $d_{\ddot{X}}d_{\ddot{Z}} \times (\ell+1)$ matrix whose ith column is given by $\mathrm{vec}(\mathbf{\Lambda}^{(i-1)}_{\ddot{Z}\ddot{X}})/(i-1)!$. Following classic kernel derivations in, for example, Fan and Gijbels (1996), we know that, for all $x \in \mathcal{X}_1$, under the conditions of Theorem 4.1,

$$
\mathbf{D}^{-1}\widehat{\mathbf{M}}(x)\mathbf{D}^{-1} = \mathbf{M}f_{X_1}(x) + h\mathbf{M}_1 f^{(1)}_{X_1}(x) + O_p(a_h), \tag{D.1}
$$

$$
\mathrm{vec}(\widehat{\mathbf{\Lambda}}_{\ddot{Z}\ddot{X}}(x)) = \mathrm{vec}(\mathbf{\Lambda}_{\ddot{Z}\ddot{X}}(x)) + O_p((Nh)^{-1/2}) + O_p(h_\ell), \tag{D.2}
$$

where $a_h = (Nh)^{-1/2} + h^2 = o(h)$ since $Nh^3 \to \infty$, and h_ℓ is h^2 for both local constant local linear estimation.

The following two lemmas state useful results for proving the Theorems and Propositions of the main paper.

Lemma D.1 *Let* $\widehat{\mathbf{B}}(x) = N^{-1}\sum_i \ddot{Z}_{i2}\ddot{X}'_{i2}\theta_2(X_{i1})\widetilde{X}'_{i1}(x)\kappa_h(X_{i1}-x)\widehat{\mathbf{M}}^{-1}(x)\iota_1 - \widehat{\mathbf{\Lambda}}_{\ddot{Z}\ddot{X}}(x)\theta_2(x)$. *Suppose that the conditions in Theorem 4.1 hold. Then,*

$$
\widehat{\mathbf{B}}(x) = h^2\mu_2\mathbf{B}_\ell + O_p(h^3),
$$

for $\ell = 0, 1$, *with* $\mathbf{B}_\ell$ *defined in Theorem 4.1.*

Lemma D.2 *Suppose that the conditions in Theorem 4.1 hold. Then, the following holds uniformly for $x \in \mathcal{X}_1$.*

$$\sup_{x\in\mathcal{X}_1} \left\|\mathbf{D}^{-1}\widehat{\mathbf{M}}(x)\mathbf{D}^{-1} - \left(\mathbf{M}f_{X_1}(x) + h\mathbf{M}_1 f_{X_1}^{(1)}(x)\right)\right\| = O_p(c_h + h^2) \tag{D.3}$$

$$\sup_{x\in\mathcal{X}_1} \|\mathbf{\Psi}(x)\| = O_p(c_h), \tag{D.4}$$

$$\sup_{x\in\mathcal{X}_1} \|\widehat{\mathbf{\Psi}}(x) - \mathbf{\Psi}(x)\| = O_p(c_h(c_h + h)), \tag{D.5}$$

$$\sup_{x\in\mathcal{X}_1} \left\|vec(\widehat{\mathbf{\Lambda}}_{\ddot{Z}\ddot{X}}(x) - \mathbf{\Lambda}_{\ddot{Z}\ddot{X}}(x))\right\| = O_p(c_h + h^2), \tag{D.6}$$

$$\sup_{x\in\mathcal{X}_1} \|(\widehat{\mathbf{\Omega}}'(x) - \mathbf{\Omega}'(x))\mathbf{\Psi}(x)\| = O_p(c_h(c_h + h^2)), \tag{D.7}$$

$$\sup_{x\in\mathcal{X}_1} \left\|\widehat{\mathbf{B}}(x)\right\| = O_p(hc_h + h^2) \tag{D.8}$$

where $c_h = (\log(1/h)/(Nh))^{1/2}$ *and* $\mathbf{\Psi}(x) \equiv \frac{1}{N}\sum_i \kappa_h(X_{i1}-x)\widetilde{\varepsilon}_{i2}\ddot{Z}_{i2}\widetilde{X}_{i1}'(x)\mathbf{D}^{-1}\mathbf{M}^{-1}\iota_1 f_{X_1}^{-1}(x)$.

Proof of Lemma D.1:

Given that $\widehat{\mathbf{\Lambda}}_{\ddot{Z}\ddot{X}}(x)\theta_2(x) = (\theta_2(x) \otimes \mathbf{I}_{d_{\ddot{Z}}})'\text{vec}(\widehat{\mathbf{\Lambda}}_{\ddot{Z}\ddot{X}}(x))$, we find that

$$\begin{aligned}\widehat{\mathbf{B}}(x) &= N^{-1}\sum_i \left((\theta_2(X_{i1}) - \theta_2(x)) \otimes \mathbf{I}_{d_{\ddot{Z}}}\right)' \widetilde{X}_{i2}\iota_1'\widehat{\mathbf{M}}^{-1}(x)\widetilde{X}_{i1}(x)\kappa_h(X_{i1} - x) \\ &= \widetilde{\mathbf{B}}_1(x) + \widetilde{\mathbf{B}}_2(x) + \mathbf{R}_{\mathbf{B}}(x),\end{aligned} \tag{D.9}$$

where

$$\begin{aligned}\widetilde{\mathbf{B}}_1(x) &= (\theta_2^{(1)}(x) \otimes \mathbf{I}_{d_{\ddot{Z}}})'N^{-1}\sum_i h\left(\frac{X_{i1}-x}{h}\right)\widetilde{X}_{i2}\iota_1'\widehat{\mathbf{M}}^{-1}(x)\widetilde{X}_{i1}(x)\kappa_h(X_{i1} - x) \\ &\equiv (\theta_2^{(1)}(x) \otimes \mathbf{I}_{d_{\ddot{Z}}})'N^{-1}\sum_i \widehat{b}_{i1}(x), \\ \widetilde{\mathbf{B}}_2(x) &= (\theta_2^{(2)}(x)/2 \otimes \mathbf{I}_{d_{\ddot{Z}}})'N^{-1}\sum_i h^2\left(\frac{X_{i1}-x}{h}\right)^2\widetilde{X}_{i2}\iota_1'\widehat{\mathbf{M}}^{-1}(x)\widetilde{X}_{i1}(x)\kappa_h(X_{i1} - x) \\ &\equiv (\theta_2^{(2)}(x)/2 \otimes \mathbf{I}_{d_{\ddot{Z}}})'N^{-1}\sum_i \widehat{b}_{i2}(x), \\ \mathbf{R}_{\mathbf{B}}(x) &= N^{-1}\sum_i(\theta_2^{(3)}(\xi_i)/6 \otimes \mathbf{I}_{d_{\ddot{Z}}})'h^3\left(\frac{X_{i1}-x}{h}\right)^3\widetilde{X}_{i2}\iota_1'\widehat{\mathbf{M}}^{-1}(x)\widetilde{X}_{i1}(x)\kappa_h(X_{i1} - x).\end{aligned}$$

where $\mathbf{R}_{\mathbf{B}}(x)$ is the Taylor expansion remainder term with ξ_i lying between x and X_{i1} for all i.

For the term $\widetilde{\mathbf{B}}_1(x)$, we know that

$$
\begin{aligned}
&\mathbb{E}[N^{-1}\sum_i \widehat{b}_{i1}(x)|X_{i1}] = N^{-1}\sum_i \mathrm{vec}(\mathbf{\Lambda}_{\ddot{Z}\ddot{X}}(X_{i1}))\,(X_{i1}-x)\,\widetilde{X}_{i1}'(x)\widehat{\mathbf{M}}^{-1}(x)\iota_1\kappa_h(X_{i1}-x)\\
&= N^{-1}\sum_i \left(\mathbf{L}_{\ddot{Z}\ddot{X}}(x)\widetilde{X}_{i1}(x)\,(X_{i1}-x) + h^{\ell+2}\frac{1}{(\ell+1)!}\mathrm{vec}(\mathbf{\Lambda}^{(\ell+1)}(x))\left(\frac{X_{i1}-x}{h}\right)^{\ell+2}\right)\\
&\qquad\qquad \times \widetilde{X}_{i1}'(x)\widehat{\mathbf{M}}^{-1}(x)\iota_1\kappa_h(X_{i1}-x) + O_p(h^{\ell+3}).
\end{aligned}
$$

Let $\mathbf{H}(x)$ be a diagonal matrix whose ith entry is given by $X_{i1}-x$. First, note that

$$
\begin{aligned}
&N^{-1}\sum_i \mathbf{L}_{\ddot{Z}\ddot{X}}(x)\widetilde{X}_{i1}(x)\,(X_{i1}-x)\,\widetilde{X}_{i1}'(x)\widehat{\mathbf{M}}^{-1}(x)\iota_1\kappa_h(X_{i1}-x) \qquad \text{(D.10)}\\
&= N^{-1}\mathbf{L}_{\ddot{Z}\ddot{X}}(x)\widetilde{\mathbf{X}}_1'(x)\mathbf{H}(x)\mathbf{K}(x)\widetilde{\mathbf{X}}_1(x)\widehat{\mathbf{M}}^{-1}(x)\iota_1\\
&= N^{-1}\sum_{s=1}^{\ell+1}[\mathbf{L}_{\ddot{Z}\ddot{X}}(x)]_s\iota_s'\widetilde{\mathbf{X}}_1'(x)\mathbf{H}(x)\mathbf{K}(x)\widetilde{\mathbf{X}}_1(x)\widehat{\mathbf{M}}^{-1}(x)\iota_1\\
&= N^{-1}\sum_{s=1}^{\ell}[\mathbf{L}_{\ddot{Z}\ddot{X}}(x)]_s\iota_{s+1}'\widetilde{\mathbf{X}}_1'(x)\mathbf{K}(x)\widetilde{\mathbf{X}}_1(x)\widehat{\mathbf{M}}^{-1}(x)\iota_1\\
&\quad + N^{-1}[\mathbf{L}_{\ddot{Z}\ddot{X}}(x)]_{\ell+1}\iota_{\ell+1}'\widetilde{\mathbf{X}}_1'(x)\mathbf{H}(x)\mathbf{K}(x)\widetilde{\mathbf{X}}_1(x)\widehat{\mathbf{M}}^{-1}(x)\iota_1\\
&= N^{-1}[\mathbf{L}_{\ddot{Z}\ddot{X}}(x)]_{\ell+1}\iota_{\ell+1}'\mathbf{D}\mathbf{D}^{-1}\widetilde{\mathbf{X}}_1'(x)\mathbf{H}(x)\mathbf{K}(x)\widetilde{\mathbf{X}}_1(x)\mathbf{D}^{-1}\left(\mathbf{D}^{-1}\widehat{\mathbf{M}}(x)\mathbf{D}^{-1}\right)^{-1}\iota_1\\
&= h^{\ell+1}[\mathbf{L}_{\ddot{Z}\ddot{X}}(x)]_{\ell+1}\iota_{\ell+1}'\left(\mathbf{M}_1 f_{X_1}(x) + h\mathbf{M}_2 f_{X_1}^{(1)}(x) + o_p(h)\right)\\
&\qquad\qquad \times\left(f_{X_1}^{-1}(x)\mathbf{M}^{-1} + h\mathsf{c}_f(x)f_{X_1}^{-1}(x)\mathbf{M}^{-1}\mathbf{M}_1\mathbf{M}^{-1} + o_p(h)\right)\iota_1.
\end{aligned}
$$

The third equality holds because $\mathbf{H}(x)\widetilde{\mathbf{X}}_1(x)\iota_s = \widetilde{\mathbf{X}}(x)\iota_{s+1}$ for $s=1,\ldots,\ell$. The fourth equality holds because $N^{-1}\iota_{s+1}'\widetilde{\mathbf{X}}_1'(x)\mathbf{K}(x)\widetilde{\mathbf{X}}_1(x)\widehat{\mathbf{M}}^{-1}(x)\iota_1 = \iota_{s+1}'\widehat{\mathbf{M}}(x)^{-1}\widehat{\mathbf{M}}(x)\iota_1 = \iota_{s+1}'\iota_1 = 0$ for all $s=1,...,\ell$. The last equality holds from standard kernel derivations.

Further, given symmetry of the kernel function $\kappa(.)$, the (i,k)-th element of $\mathbf{M}$ is zero if $i+k$ is odd. This implies that the adjoint matrix of the $(1,k)$-th element of $\mathbf{M}$ is singular if k is even and, therefore, $\mathbf{M}^{-1}\iota_1$ is a $(\ell+1)\times 1$ vector with all even elements equal to zero. On the other hand, $\iota_{\ell+1}'\mathbf{M}_j = (\int u^{\ell+j}\kappa(u)du\ \ldots\ \int u^{2\ell+j}\kappa(u)du)$ have zero odd elements when $\ell+j$ is odd and zero even elements if $\ell+j$ is even. Therefore, $\iota_{\ell+1}'\mathbf{M}_1\mathbf{M}^{-1}\iota_1 = 0$ when ℓ is even. Using similar arguments, we know that $\iota_{\ell+1}'\mathbf{M}_1\mathbf{M}^{-1}\mathbf{M}_1\mathbf{M}^{-1}\iota_1 = 0$ when ℓ is even and $\iota_{\ell+1}'\mathbf{M}_2\mathbf{M}^{-1}\iota_1 = 0$ when ℓ is odd.

Therefore, we have

$$
\text{(D.10)} = h^{\ell+1}[\mathbf{L}_{\ddot{Z}\ddot{X}}(x)]_{\ell+1}\iota_{\ell+1}'\mathbf{M}_1\mathbf{M}^{-1}\iota_1 + O_p(h^{\ell+2}), \qquad \text{(D.11)}
$$

if ℓ is odd and

$$(\text{D.10}) = h^{\ell+2}[\mathbf{L}_{\ddot{Z}\ddot{X}}(x)]_{\ell+1}\mathsf{c}_f(x)\iota_{\ell+1}'\mathbf{M}_2\mathbf{M}^{-1}\iota_1 + O_p(h^{\ell+3}), \tag{D.12}$$

if ℓ is even. Similarly, the other component of $\mathbb{E}[N^{-1}\sum_i \widehat{b}_{i1}(x)|X_{i1}]$ follows

$$\begin{aligned}
&N^{-1}\sum_i h^{\ell+2}\frac{1}{(\ell+1)!}\text{vec}(\mathbf{\Lambda}^{(\ell+1)}(x))\left(\frac{X_{i1}-x}{h}\right)^{\ell+2}\widetilde{X}_{i1}'\widehat{\mathbf{M}}^{-1}(x)\iota_1\kappa_h(X_{i1}-x)\\
&= h^{\ell+2}\frac{1}{(\ell+1)!}\text{vec}(\mathbf{\Lambda}^{(\ell+1)}(x))\iota_{\ell+1}'\mathbf{M}_2\mathbf{M}^{-1}\iota_1 + O_p(h^{\ell+3}),
\end{aligned}$$

if ℓ is even and $O_p(h^{\ell+3})$ if ℓ is odd.

For the term $\widetilde{\mathbf{B}}_2(x)$, we know that

$$\begin{aligned}
&\mathbb{E}[N^{-1}\sum_i \widehat{b}_{i2}(x)|X_{i1}]\\
=&N^{-1}\sum_{s=1}^{\ell+1}[\mathbf{L}_{\ddot{Z}\ddot{X}}(x)]_s\iota_s'\widetilde{\mathbf{X}}_1'(x)\mathbf{H}^2(x)\mathbf{K}(x)\widetilde{\mathbf{X}}_1(x)\mathbf{D}^{-1}\left(\mathbf{D}^{-1}\widehat{\mathbf{M}}(x)\mathbf{D}^{-1}\right)^{-1}\iota_1 + O_p(h^{\ell+3})\\
=&N^{-1}[\mathbf{L}_{\ddot{Z}\ddot{X}}(x)]_\ell\iota_{\ell+1}'\mathbf{D}\mathbf{D}^{-1}\widetilde{\mathbf{X}}_1'(x)\mathbf{H}(x)\mathbf{K}(x)\widetilde{\mathbf{X}}_1(x)\mathbf{D}^{-1}\left(\mathbf{D}^{-1}\widehat{\mathbf{M}}(x)\mathbf{D}^{-1}\right)^{-1}\iota_1\\
&+ N^{-1}[\mathbf{L}_{\ddot{Z}\ddot{X}}(x)]_{\ell+1}\iota_{\ell+1}'\mathbf{D}\mathbf{D}^{-1}\widetilde{\mathbf{X}}_1'(x)\mathbf{H}_2(x)\mathbf{K}(x)\widetilde{\mathbf{X}}_1(x)\mathbf{D}^{-1}\left(\mathbf{D}^{-1}\widehat{\mathbf{M}}(x)\mathbf{D}^{-1}\right)^{-1}\iota_1\\
&+ O_p(h^{\ell+3}).
\end{aligned}$$

The first term in the RHS is not relevant for $\ell = 0$. When relevant, it reduces to $h^{\ell+1}[\mathbf{L}_{\ddot{Z}\ddot{X}}(x)]_\ell\iota_{\ell+1}'\mathbf{M}_1\mathbf{M}^{-1}\iota_1 + O_p(h^{\ell+2})$ when ℓ is odd and $h^{\ell+2}[\mathbf{L}_{\ddot{Z}\ddot{X}}(x)]_\ell\mathsf{c}_f(x)\iota_{\ell+1}'\mathbf{M}_2\mathbf{M}^{-1}\iota_1 + O_p(h^{\ell+3})$ when ℓ is even. The second term reduces to $h^{\ell+2}[\mathbf{L}_{\ddot{Z}\ddot{X}}(x)]_{\ell+1}\iota_{\ell+1}'\mathbf{M}_2\mathbf{M}^{-1}\iota_1 + O_p(h^{\ell+3})$ when ℓ is even and $O_p(h^{\ell+3})$ when ℓ is odd.

Using similar derivations, one can show that $\mathbb{V}[N^{-1}\sum_i(\widehat{b}_{i1}(x)+\widehat{b}_{i2}(x))|X_{i1}] = O_p(h/N) = O_p(h^4)$ if $Nh^3\to\infty$ and that $\mathbf{R}_{\mathbf{B}}(x) = O_p(h^3)$ under the uniform boundedness conditions.

Summing up, we conclude that

$$\begin{aligned}
\widehat{\mathbf{B}}(x) &= (\theta_2^{(1)}(x)\otimes\mathbf{I}_{d_{\ddot{Z}}})'\cdot h^2\text{vec}(\mathbf{\Lambda}_{\ddot{Z}\ddot{X}}^{(1)}(x))\iota_2'\mathbf{M}_1\mathbf{M}^{-1}\iota_1\\
&\quad + (\theta_2^{(2)}(x)/2\otimes\mathbf{I}_{d_{\ddot{Z}}})'\cdot h^2\text{vec}(\mathbf{\Lambda}_{\ddot{Z}\ddot{X}}(x))\iota_2'\mathbf{M}_1\mathbf{M}^{-1}\iota_1 + o_p(h^2)\\
&= h^2\mu_2\left(\mathbf{\Lambda}_{\ddot{Z}\ddot{X}}^{(1)}(x)\theta_2^{(1)}(x) + \mathbf{\Lambda}_{\ddot{Z}\ddot{X}}(x)\theta_2^{(2)}(x)/2\right) + O_p(h^3),
\end{aligned}$$

when $\ell = 1$ and

$$\widehat{\mathbf{B}}(x) = h^2\mu_2\left[\left(\mathsf{c}_f(x)\mathbf{\Lambda}_{\ddot{Z}\ddot{X}}(x) + \mathbf{\Lambda}_{\ddot{Z}\ddot{X}}^{(1)}(x)\right)\theta_2^{(1)}(x) + \mathbf{\Lambda}_{\ddot{Z}\ddot{X}}(x)\theta_2^{(2)}(x)/2\right] + O_p(h^3)$$

when $\ell = 0$, which corresponds to the statement in the lemma.

Proof of Lemma D.2

The first part of the lemma can be proven by showing that $\sup_x \left\|\mathbf{D}^{-1}\widehat{\mathbf{M}}(x)\mathbf{D}^{-1} - \mathbb{E}[\mathbf{D}^{-1}\widehat{\mathbf{M}}(x)\mathbf{D}^{-1}]\right\| = O_p(c_h)$, while $\sup_x \left\|\mathbb{E}[\mathbf{D}^{-1}\widehat{\mathbf{M}}(x)\mathbf{D}^{-1}] - \mathbf{M}f_{X_1}(x) - h\mathbf{M}_1 f_{X_1}^{(1)}(x)\right\| = O_p(h^2)$. The former follows from a trivial extension of the uniform convergence result stated in Mack and Silverman (1982) for local constant estimation to local polynomial estimation. The latter follows from standard kernel bias derivation and uniform boundedness conditions stated in the assumptions.

The second part of the lemma is obtained as a consequence of Lemma A.1 in Fan and Huang (2005). Specifically, it can be shown by using $\sup_{x\in\mathcal{X}_1} |f_{X_1}^{-1}(x)| < \infty$ (Assumption D.1.(b)) and the fact that

$$\sup_{x\in\mathcal{X}_1} \|N^{-1}\sum_i \widetilde{\varepsilon}_{i2}\ddot{Z}_{i2}\widetilde{X}_{i1}'(x)\mathbf{D}^{-1}\kappa_h(X_{i1}-x)\| = O_p(c_h). \tag{D.13}$$

Then, (D.5) is obtained as a consequence of (D.3) and (D.13); specifically, we have

$$\begin{aligned}
&\|\widehat{\boldsymbol{\Psi}}(x) - \boldsymbol{\Psi}(\mathbf{x})\| \\
&\leq \|N^{-1}\sum_i \widetilde{\varepsilon}_{i2}\ddot{Z}_{i2}\widetilde{X}_{i1}'(x)\mathbf{D}^{-1}\kappa_h(X_{i1}-x)\|\|(\mathbf{D}^{-1}\widehat{\mathbf{M}}(x)\mathbf{D}^{-1})^{-1} - \mathbf{M}^{-1}f_{X_1}^{-1}(x)\| \\
&\leq O_p(c_h)\|(\mathbf{D}^{-1}\widehat{\mathbf{M}}(x)\mathbf{D}^{-1})^{-1}\|\|\mathbf{D}^{-1}\widehat{\mathbf{M}}(x)\mathbf{D}^{-1} - \mathbf{M}f_{X_1}(x)\|\|\mathbf{M}^{-1}f_{X_1}^{-1}(x)\| \\
&= O_p(c_h)O_p(c_h + h),
\end{aligned} \tag{D.14}$$

uniformly in $x \in \mathcal{X}_1$. The second inequality follows from $A^{-1} - B^{-1} = B^{-1}(A-B)A^{-1}$ for matrices A and B.

To prove (D.6), we define $\upsilon_i = \mathrm{vec}(\ddot{Z}_{i2}\ddot{X}_{i2}' - \Lambda_{\ddot{Z}\ddot{X}}(X_{i1}))$ and then, given that $\boldsymbol{\Lambda}_{\ddot{Z}\ddot{X}}(.)$ is twice continuously differentiable, we have

$$\begin{aligned}
&\mathrm{vec}(\widehat{\boldsymbol{\Lambda}}_{\ddot{Z}\ddot{X}}(x) - \boldsymbol{\Lambda}_{\ddot{Z}\ddot{X}}(x)) \\
=&N^{-1}\widetilde{\mathbf{X}}_2'\mathbf{K}(x)\widetilde{\mathbf{X}}_1(x)\widehat{\mathbf{M}}(x)^{-1}\iota_1 - \mathbf{L}_{\ddot{Z}\ddot{X}}(x)N^{-1}\widetilde{\mathbf{X}}_1'(x)\mathbf{K}(x)\widetilde{\mathbf{X}}_1(x)\widehat{\mathbf{M}}^{-1}(x)\iota_1 \\
=&N^{-1}(\widetilde{\mathbf{X}}_2 - \widetilde{\mathbf{X}}_1(x)\mathbf{L}_{\ddot{Z}\ddot{X}}'(x))'\mathbf{K}(x)\widetilde{\mathbf{X}}_1(x)\mathbf{D}^{-1}\left(\mathbf{D}^{-1}\widehat{\mathbf{M}}(x)\mathbf{D}^{-1}\right)^{-1}\mathbf{D}^{-1}\iota_1.
\end{aligned}$$

By (D.3) and Assumption D.1.(b) on the density function, we know that $\left(\mathbf{D}^{-1}\widehat{\mathbf{M}}(x)\mathbf{D}^{-1}\right)^{-1} = O_p(1)$ uniformly in $x \in \mathcal{X}_1$. In addition, we have

$$N^{-1}(\widetilde{\mathbf{X}}_2 - \widetilde{\mathbf{X}}_1(x)\mathbf{L}_{\ddot{Z}\ddot{X}}'(x))'\mathbf{K}(x)\widetilde{\mathbf{X}}_1(x)\mathbf{D}^{-1} = \mathtt{r}_{11}(x) + \mathtt{r}_{12}(x), \tag{D.15}$$

$$\text{where } \mathtt{r}_{11}(x) = N^{-1}\sum_i \upsilon_i \widetilde{X}'_{i1}(x)\mathbf{D}^{-1}\kappa_h(X_{i1}-x),$$

$$\mathtt{r}_{12}(x) = 1_{\{\ell=0\}} N^{-1}h\sum_i \mathrm{vec}(\mathbf{\Lambda}^{(1)}_{\ddot{Z}\ddot{X}}(x))\left(\frac{X_{i1}-x}{h}\right)\kappa_h(X_{i1}-x)$$
$$+ N^{-1}h^2\sum_i \mathrm{vec}(\mathbf{\Lambda}^{(2)}_{\ddot{Z}\ddot{X}}(\xi_{i1}))\left(\frac{X_{i1}-x}{h}\right)^2 \widetilde{X}'_{i1}(x)\mathbf{D}^{-1}\kappa_h(X_{i1}-x),$$

by the Taylor expansion for some ξ_{i1} between X_{i1} and x. Because $\mathbb{E}[\upsilon_i|X_{i1}] = 0$ and $\sup_{x\in\mathcal{X}_1}\mathbb{E}[|\nu_i\|^s|X_{i1}=x] < \infty$, we follow the uniform convergence results in Mack and Silverman (1982) and Fan and Huang (2005, Lemma A.1) and obtain that $\mathtt{r}_{11}(.) = O_p(c_h)$ uniformly on $\mathcal{X}_1$. Similarly,

$$\mathtt{r}_{12}(x) = h1_{\{\ell=0\}}\mathrm{vec}(\mathbf{\Lambda}^{(1)}_{\ddot{Z}\ddot{X}}(x))\mathbb{E}\left[\left(\frac{X_{i1}-x}{h}\right)\kappa_h(X_{i1}-x)\right]$$
$$+ h^2\mathbb{E}\left[\left(\frac{X_{i1}-x}{h}\right)^2 \mathrm{vec}(\mathbf{\Lambda}^{(2)}_{\ddot{Z}\ddot{X}}(\xi_{i1}))\widetilde{X}'_{i1}(x)\mathbf{D}^{-1}\kappa_h(X_{i1}-x)\right] + O_p(c_h)$$
$$= h^2(\mathrm{vec}(\mathbf{\Lambda}^{(2)}_{\ddot{Z}\ddot{X}}(x)) + o_p(1))(f_{X_1}(x)\iota'_1\mathbf{M}_2 + O_p(h)) + O_p(c_h+h^2),$$

uniformly in $x \in \mathcal{X}_1$. The last equality follows from uniform boundedness conditions in Assumptions D.1.(b) and D.1.(e). Then, the desired result is followed.

$$N^{-1}(\widetilde{\mathbf{X}}_2 - \widetilde{\mathbf{X}}_1(x)\mathbf{L}'_{\ddot{Z}\ddot{X}}(x))'\mathbf{K}(x)\widetilde{\mathbf{X}}_1(x)\mathbf{D}^{-1}\left(\mathbf{D}^{-1}\widehat{\mathbf{M}}(x)\mathbf{D}^{-1}\right)^{-1}\mathbf{D}^{-1}\iota_1 = O_p(c_h+h^2).$$

We then focus on (D.7). Because of (D.6) and Assumption 3.1, we have $\|\widehat{\mathbf{\Omega}}(x) - \mathbf{\Omega}(x)\| = O_p(c_h+h^2)$ uniformly in $x \in \mathcal{X}_1$. Then, the desired result is obtained by combining this with (D.4).

To study the last part, we recall that $\widehat{\mathbf{B}}(x)$ can be decomposed into $\widetilde{\mathbf{B}}_1(x)$, $\widetilde{\mathbf{B}}_2(x)$ and $\mathbf{R}_{\mathbf{B}}(x)$, each of which is defined in the proof of Lemma D.1. We then let

$$N^{-1}\sum_i b_{i1}(x) = N^{-1}\sum_i h\left(\frac{X_{i1}-x}{h}\right)\widetilde{X}_{i2}\iota'_1\mathbf{M}^{-1}\mathbf{D}^{-1}\widetilde{X}_{i1}(x)\kappa_h(X_{i1}-x)f^{-1}_{X_1}(x),$$
$$N^{-1}\sum_i b_{i2}(x) = N^{-1}\sum_i h^2\left(\frac{X_{i1}-x}{h}\right)^2\widetilde{X}_{i2}\iota'_1\mathbf{M}^{-1}\mathbf{D}^{-1}\widetilde{X}_{i1}(x)\kappa_h(X_{i1}-x)f^{-1}_{X_1}(x).$$

Because of the uniform boundedness conditions in Assumptions D.1.(b), D.1.(c) and D.1.(f) and the uniform convergence result in Lemma A.1 in Fan and Huang (2005), we find that

$$N^{-1}\sum_i\left(\frac{X_{i1}-x}{h}\right)\widetilde{X}_{i2}\widetilde{X}_{i1}'\mathbf{D}^{-1}\kappa_h(X_{i1}-x)$$

$$=\text{vec}\left(\mathbf{\Lambda}_{\ddot{Z}\ddot{X}}(x)\right)\iota_1'\mathbf{M}_1 f_{X_1}(x)$$

$$+hf_{X_1}(x)\left(\mathsf{c}_f(x)\text{vec}\left(\mathbf{\Lambda}_{\ddot{Z}\ddot{X}}(x)\right)\iota_1'\mathbf{M}_1+\text{vec}\left(\mathbf{\Lambda}^{(1)}_{\ddot{Z}\ddot{X}}(x)\right)\iota_1'\mathbf{M}_2\right)+O_p(c_h+h^2),$$

uniformly in $x\in\mathcal{X}_1$. By using similar arguments in proving (D.11) and (D.12), we have $\iota_1'\mathbf{M}_1\mathbf{M}^{-1}\iota_1=0$ and thus $N^{-1}\sum_i b_{i1}(x)$ satisfies the following:

$$N^{-1}\sum_i b_{i1}(x)=h^2\text{vec}\left(\mathbf{\Lambda}^{(1)}_{\ddot{Z}\ddot{X}}(x)\right)\iota_1'\mathbf{M}_2\mathbf{M}^{-1}\iota_1+O_p(h(c_h+h^2))=O_p(hc_h+h^2). \quad \text{(D.16)}$$

Similarly, the following holds uniformly in $x\in\mathcal{X}_1$:

$$N^{-1}\sum_i\left(\frac{X_{i1}-x}{h}\right)^2\widetilde{X}_{i2}\widetilde{X}_{i1}'\mathbf{D}^{-1}\kappa_h(X_{i1}-x)=\text{vec}\left(\mathbf{\Lambda}_{\ddot{Z}\ddot{X}}(x)\right)\iota_1'\mathbf{M}_2 f_{X_1}(x)+O_p(c_h+h),$$

from which it is deduced that

$$N^{-1}\sum_i b_{i2}(x)=h^2\text{vec}\left(\mathbf{\Lambda}_{\ddot{Z}\ddot{X}}(x)\right)\iota_1'\mathbf{M}_2\mathbf{M}^{-1}\iota_1+O_p(h^2(c_h+h))=O_p(h^2c_h+h^2). \quad \text{(D.17)}$$

Then, the followings are deduced from Assumption D.1.(c), (D.3), (D.16), and (D.17):

$$\widetilde{\mathbf{B}}_1(x)=(\theta_1^{(1)}(x)\otimes\mathbf{I}_{d_{\ddot{Z}}})'N^{-1}\sum_i b_{i1}(x)+o_p(hc_h+h^2)=O_p(hc_h+h^2), \quad \text{(D.18)}$$

$$\widetilde{\mathbf{B}}_2(x)=(\theta_1^{(2)}(x)/2\otimes\mathbf{I}_{d_{\ddot{Z}}})'N^{-1}\sum_i b_{i2}(x)+o_p(h^2c_h+h^2)=O_p(h^2c_h+h^2), \quad \text{(D.19)}$$

uniformly in $\mathcal{X}_1$. Using similar derivations, one can show that $\mathbf{R}_{\mathbf{B}}(.)=o_p(hc_h+h^2)$. By combining these results, the last part of the lemma is obtained.

Appendix E: Proofs of Theorems and Propositions

Proof of equality (3.3)

The first equality in equation (3.3) holds because

$$\mathbb{E}\left[\omega_t(Z_{it})\left(Y_{it}-\left(\beta_t^0(x)X_{it}+H_{it}'\widetilde{\gamma}_t(x)\right)\right)|X_{i(t-1)}=x\right]$$

$$=\mathbb{E}\left[\omega_t(Z_{it})\left(\varepsilon_{it}-H_{it}'(\widetilde{\gamma}_t(x)-\gamma_t(x))\right)|X_{i(t-1)}=x\right]$$

$$=\mathbb{E}\left[\omega_t(Z_{it})\varepsilon_{it}|X_{i(t-1)}=x\right]$$

$$-\mathbb{E}\left[\omega(Z_{it})\mathbb{E}[\mathbb{L}[\varepsilon_{it}|X_{i(t-1)}=x,\widetilde{H}_{it}]|Z_{it},X_{i(t-1)}=x]|X_{i(t-1)}=x\right]$$

$$=\mathbb{E}\left[\omega_t(Z_{it})\varepsilon_{it}|X_{i(t-1)}=x\right]$$

$$-\mathbb{E}\left[\omega_t(Z_{it})\mathbb{E}[\mathbb{L}[\varepsilon_{it}|X_{i(t-1)}=x,\widetilde{H}_{it}]|X_{i(t-1)}=x]|X_{i(t-1)}=x\right],$$

where the first two equalities hold respectively from the outcome equation in (3.1) and by the law of iterated expectations. The third equality holds because $\mathbb{E}[\widetilde{H}_{it}|X_{i(t-1)}, Z_{it}] = \mathbb{E}[\widetilde{H}_{it}|X_{i(t-1)}]$ by Assumption 3.2 while the fourth holds because $\mathbb{E}[\mathbb{L}[\varepsilon_{it}|X_{i(t-1)} = x, \widetilde{H}_i]|X_{i(t-1)} = x] = \mathbb{E}[\varepsilon_{it}|X_{i(t-1)} = x]$ by the definition of linear projection. To see this, one just needs to show by block matrix inversion that $\mathbb{E}[\mathbb{L}[Y|X]] = \mathbb{E}[Y]$, where $\mathbb{L}[Y|X] = (1\ X')(\mathbb{E}[(1\ X')'(1\ X')])^{-1}(\mathbb{E}[(1\ X')'Y])$ for any scalar random variable Y and random vector X.

The second equality in equation (3.3) is explained in the main text.

Proof of Theorem 4.1

We prove the theorem without loss of generality with an identity GMM weighting matrix. Recall that

$$\widehat{\theta}_2(x) = \left(\widehat{\Lambda}_{\ddot{Z}\ddot{X}}(x)'\widehat{\Lambda}_{\ddot{Z}\ddot{X}}(x)\right)^{-1}\widehat{\Lambda}_{\ddot{Z}\ddot{X}}(x)'\widehat{\Lambda}_{\ddot{Z}Y}(x).$$

Let $\widehat{\mathbf{\Omega}}(x) = \widehat{\Lambda}_{\ddot{Z}\ddot{X}}(x)\left(\widehat{\Lambda}_{\ddot{Z}\ddot{X}}(x)'\widehat{\Lambda}_{\ddot{Z}\ddot{X}}(x)\right)^{-1}$. Then, $\widehat{\theta}_2(x) = \widehat{\mathbf{\Omega}}'(x)\widehat{\Lambda}_{\ddot{Z}Y}(x)$, and $\theta_2(x) = \widehat{\mathbf{\Omega}}'(x)\widehat{\Lambda}_{\ddot{Z}\ddot{X}}(x)\theta_2(x)$.

Since $\widehat{\mathbf{\Lambda}}_{\ddot{Z}Y}(x) = N^{-1}\sum_i \kappa_h(X_{i1}-x)\ddot{Z}_{i2}\ddot{X}'_{i2}\theta_2(X_{i1})\widetilde{X}'_{i1}(x)\widehat{\mathbf{M}}^{-1}(x)\iota_1 + N^{-1}\sum_i \kappa_h(X_{i1}-x)\ddot{Z}_{i2}\widetilde{\varepsilon}_{i2}\widetilde{X}'_{i1}(x)\widehat{\mathbf{M}}^{-1}(x)\iota_1$. If we define $\widehat{\mathbf{\Psi}}(x) = N^{-1}\sum_i \kappa_h(X_{i1}-x)\ddot{Z}_{i2}\widetilde{\varepsilon}_{i2}\widetilde{X}'_{i1}(x)\widehat{\mathbf{M}}^{-1}(x)\iota_1$. Then, the difference between $\widehat{\theta}_2(x)$ and $\theta_2(x)$ can be written as:

$$\widehat{\theta}_2(x) - \theta_2(x) = \widehat{\mathbf{\Omega}}'(x)\widehat{\mathbf{B}}(x) + \widehat{\mathbf{\Omega}}'(x)\widehat{\mathbf{\Psi}}(x). \tag{E.1}$$

From (D.2), we have

$$\widehat{\mathbf{\Omega}}(x) - \mathbf{\Omega}(x) = O_p((Nh)^{-1/2}) + o_p(h_\ell), \tag{E.2}$$

where $\mathbf{\Omega}(x) = \mathbf{\Lambda}_{\ddot{Z}\ddot{X}}(x)\left(\mathbf{\Lambda}'_{\ddot{Z}\ddot{X}}(x)\mathbf{\Lambda}_{\ddot{Z}\ddot{X}}(x)\right)^{-1}$ as is defined in Theorem 4.1.

Next, rewrite $\widehat{\mathbf{\Psi}}(x)$ as follows.

$$\begin{aligned}\widehat{\mathbf{\Psi}}(x) &= \left(\iota'_1(\mathbf{D}^{-1}\widehat{\mathbf{M}}\mathbf{D}^{-1})^{-1} \otimes \mathbf{I}_{d_{\ddot{Z}}}\right) N^{-1}\sum_i \widetilde{\varepsilon}_{i2}\kappa_h(X_{i1}-x)(\mathbf{D}^{-1}\widetilde{X}_{i1}(x) \otimes \mathbf{I}_{d_{\ddot{Z}}})\ddot{Z}_{i2} \\ &\equiv \left(\iota'_1(\mathbf{D}^{-1}\widehat{\mathbf{M}}\mathbf{D}^{-1})^{-1} \otimes \mathbf{I}_{d_{\ddot{Z}}}\right) N^{-1}\sum_i \varphi_i(x),\end{aligned}$$

where $\varphi_i(x)$ is a $d_{\ddot{Z}} \times d_{\ddot{X}}$ dimensional vector. Note that

$$
\begin{aligned}
&(\iota_1'(\mathbf{M} f_{X_1}(x))^{-1} \otimes \mathbf{I}_{\ddot{Z}})\mathbb{V}(\sqrt{h}\varphi_i(x))((\mathbf{M} f_{X_1}(x))^{-1}\iota_1 \otimes \mathbf{I}_{\ddot{Z}}) \\
&= f_{X_1}^{-2}(x)\mathbb{E}\left[h\kappa_h^2(X_{i1}-x)\iota_1'\mathbf{M}^{-1}\mathbf{D}^{-1}\widetilde{X}_{i1}(x)\widetilde{X}_{i1}'(x)\mathbf{D}^{-1}\mathbf{M}^{-1}\iota_1\mathbb{E}[\widetilde{\varepsilon}_{i2}^2\ddot{Z}_{i2}\ddot{Z}_{i2}'|X_{i1}]\right] \\
&= f_{X_1}^{-1}(x)\mathbb{E}[\widetilde{\varepsilon}_{i2}^2\ddot{Z}_{i2}\ddot{Z}_{i2}'|X_{i1}=x]\iota_1'\mathbf{M}^{-1}\mathbf{M}_\kappa\mathbf{M}^{-1}\iota_1 + o(1),
\end{aligned} \tag{E.3}
$$

where $\mathbf{M}_\kappa$ is a $(\ell+1)\times(\ell+1)$ matrix whose (i,j)-th component is ν_{i+j-2}. Note that $\iota_1'\mathbf{M}^{-1}\mathbf{M}_\kappa\mathbf{M}^{-1}\iota_1 = \nu_0$ for both $\ell = 0, 1$.

Then, because of (E.3), the convergence of $\mathbf{D}^{-1}\widehat{\mathbf{M}}\mathbf{D}^{-1}$ in (D.1), the Cramér-Wold device, the Lyapunov's central limit theorem, and Slutsky's theorem, we have the weak convergence result that

$$
\sqrt{Nh}\widehat{\mathbf{\Psi}}(x) \to_d \mathcal{N}\left(0, \Sigma(x)\right).
$$

Together with (E.1) and (E.2), the weak convergence result of the theorem is proven.

Proof of Proposition 4.1

The pointwise consistency of $\widehat{\mathbf{\Omega}}(x)$ has been established in (E.2). In this proof, we focus on the consistency of $\widehat{\mathbf{\Sigma}}(x)$. Let $\widehat{f}_{X_1}(x)$ be a consistent estimator of $f_{X_1}(x)$. Then,

$$
\begin{aligned}
\widehat{\mathbf{\Sigma}}(x) &= \widehat{f}_{X_1}^{-2}(x)\frac{h}{N}\sum_i \widetilde{\varepsilon}_{i2}^2\ddot{Z}_{i2}\ddot{Z}_{i2}'\kappa_h^2(X_{i1}-x) + \mathbf{R}_\Sigma(x) \\
&= f_{X_1}^{-2}(x)\mathbb{E}[\widetilde{\varepsilon}_{i2}^2\ddot{Z}_{i2}\ddot{Z}_{i2}'h\kappa_h^2(X_{i1}-x) + \mathbf{R}_\Sigma(x) + o_p(1) \\
&= f_{X_1}^{-2}(x)\int \mathbb{V}(\widetilde{\varepsilon}_{i2}\ddot{Z}_{i2}|X_{i1}=x+uh)f_{X_1}(x+uh)\kappa^2(u)du + \mathbf{R}_\Sigma(x) + o_p(1) \\
&= f_{X_1}^{-1}(x)\mathbb{V}(\widetilde{\varepsilon}_{i2}\ddot{Z}_{i2}|X_{i1}=x)\int \kappa^2(u)du + \mathbf{R}_\Sigma(x) + o_p(1) \\
&= \mathbf{\Sigma}(x) + \mathbf{R}_\Sigma(x) + o_p(1),
\end{aligned} \tag{E.4}
$$

where $\mathbf{R}_\Sigma(x)$ is the sum of two terms, $\mathbf{R}_{\Sigma,1}(x)$ and $\mathbf{R}_{\Sigma,2}(x)$, defined by

$$
\begin{aligned}
\mathbf{R}_{\Sigma,1}(x) &:= \frac{h}{N}\sum_i \left(\ddot{X}_{i2}'(\widehat{\theta}_2(X_{i1}) - \theta_2(X_{i1}))\right)^2 \ddot{Z}_{i2}\ddot{Z}_{i2}'\kappa_h^2(X_{i1}-x), \\
\mathbf{R}_{\Sigma,2}(x) &:= \frac{2h}{N}\sum_i \left(\widetilde{\varepsilon}_{i2}\ddot{X}_{i2}'(\widehat{\theta}_2(X_{i1}) - \theta_2(X_{i1}))\right) \ddot{Z}_{i2}\ddot{Z}_{i2}'\kappa_h^2(X_{i1}-x).
\end{aligned}
$$

The second and third equalities of (E.4) are obtained from the standard arguments on the pointwise consistency of the kernel estimator and kernel estimation derivations.

The fourth equality is obtained from Assumptions D.1.(b) and D.1.(f). The rest of the proof then focuses on showing that $\mathbf{R}_\Sigma(x)$ is $o_p(1)$.

The first remainder $\mathbf{R}_{\Sigma,1}(x)$ is bounded above as follows.

$$\begin{aligned}\|\mathbf{R}_{\Sigma,1}(x)\| &\le 2N^{-1}\sum_i \|\ddot{X}_{i2}\|^2\|\ddot{Z}_{i2}\|^2\|\widehat{\theta}_2(X_{i1})-\theta_2(X_{i1})\|^2 h\kappa_h^2(X_{i1}-x)\\ &\le \sup_{x\in\mathcal{X}_1}\|\widehat{\theta}_2(x)-\theta_2(x)\|^2 N^{-1}\sum_i \|\ddot{X}_{i2}\|^2\|\ddot{Z}_{i2}\|^2 h\kappa_h^2(X_{i1}-x).\end{aligned}$$

Then, because of the results in (E.1), (E.2), and (D.5), and Lemma D.2, we find that

$$\sup_{x\in\mathcal{X}_1}\|\widehat{\theta}_2(x)-\theta_2(x)\|+o_p(1)\le \sup_{x\in\mathcal{X}_1}\|\widehat{\mathbf{\Omega}}(x)\|\sup_{x\in\mathcal{X}_1}\left(\|\widehat{\mathbf{B}}(x)\|+\|\widehat{\mathbf{\Psi}}(x)\|\right)=o_p(1). \tag{E.5}$$

Moreover, because of the Markov's inequality and Assumptions D.1.(a), D.1.(d), and D.1.(f), we have that $N^{-1}\sum_i \|\ddot{X}_{i2}\|^2\|\ddot{Z}_{i2}\|^2 h\kappa_h^2(X_{i1}-x)=O_p(1)$. Hence, by combining these, we have uniformly in $x\in\mathcal{X}_1$

$$\|\mathbf{R}_{\Sigma,1}(x)\|=o_p(1). \tag{E.6}$$

By using similar arguments, we find that uniformly in $x\in\mathcal{X}_1$,

$$\|\mathbf{R}_{\Sigma,2}(x)\|\le o_p(1)N_s^{-1}\sum_{i:X_{i1}\in\mathcal{X}_1}\|\widetilde{\varepsilon}_{i2}\|\|\ddot{X}_{i2}\|\|\ddot{Z}_{i2}\|^2 h\kappa_h^2(X_{i1}-x)=o_p(1). \tag{E.7}$$

Thus, the desired result is given from (E.2), (E.4), (E.6) and (E.7).

Proof of Theorem 4.2

Let $p=\mathbb{E}[1_{\{X_{i1}\in\mathcal{X}_1^*\}}]$ and $\widehat{p}=N_s/N$. Unless otherwise specified, all summations in the proof over i are with respect to $X_{i1}\in\mathcal{X}_1^*$, while all summations over j are with respect to the full sample, or $j=1,...,N$. Let $\widetilde{\vartheta}_2={N_s}^{-1}\sum_{i:X_{i1}\in\mathcal{X}_1^*}\theta_2(X_{i1})$. We first notice that

$$\sqrt{N_s}(\widehat{\vartheta}_2-\vartheta_2)=\sqrt{N_s}(\widehat{\vartheta}_2-\widetilde{\vartheta}_2)+\sqrt{N_s}(\widetilde{\vartheta}_2-\vartheta_2). \tag{E.8}$$

For notational convenience, we for the moment let the scalar-valued random variable $\varsigma(X_{i1},X_{j1})$ be defined as follows.

$$\varsigma(X_{i1},X_{j1})=\kappa_h(X_{j1}-X_{i1})f_{X_1}^{-1}(X_{i1})\cdot\widetilde{X}_{j1}'(X_{i1})\mathbf{D}^{-1}\mathbf{M}^{-1}\iota_1.$$

In our case of $\ell=0,1$, the random variable $\varsigma(X_{i1},X_{j1})$ reduces to $\kappa_h(X_{i1}-X_{j1})f_{X_1}^{-1}(X_{i1})$ for any i and j, because $\iota_1'\mathbf{M}^{-1}\mathbf{D}^{-1}\widetilde{X}_{j1}(.)=1$ for any j.

Then, the following holds because of results in Lemma D.2:

$$
\begin{aligned}
\sqrt{N_s}(\widehat{\vartheta}_2-\widetilde{\vartheta}_2) =& N_s^{-1/2}\sum_i\Big(\widehat{\theta}_2(X_{i1})-\theta_2(X_{i1})\Big)\\
=& N_s^{-1/2}\sum_i\Big(\widehat{\mathbf{\Omega}}'(X_{i1})\widehat{\mathbf{B}}'(X_{i1})+\widehat{\mathbf{\Omega}}'(X_{i1})\widehat{\mathbf{\Psi}}'(X_{i1})\Big)\\
=& N_s^{-1/2}\sum_i\Big(\widehat{\mathbf{\Omega}}'(X_{i1})\widehat{\mathbf{\Psi}}'(X_{i1})\Big)+O_p(\sqrt{N}(hc_h+h^2))\\
=& N_s^{-1/2}\sum_i\mathbf{\Omega}'(X_{i1})\mathbf{\Psi}(X_{i1})\\
&+N_s^{-1/2}\sum_i(\widehat{\mathbf{\Omega}}'(x)-\mathbf{\Omega}'(x))\mathbf{\Psi}(x)\\
&+N_s^{-1/2}\sum_i\widehat{\mathbf{\Omega}}'(X_{i1})(\widehat{\mathbf{\Psi}}(x)-\mathbf{\Psi}(x))+O_p(\sqrt{N}(hc_h+h^2))\\
=& N_s^{-1/2}\sum_i\mathbf{\Omega}'(X_{i1})\mathbf{\Psi}(X_{i1})+O_p(\sqrt{N}(c_h+h)^2)\\
=& N^{-1}N_s^{-1/2}\sum_j\sum_i\widetilde{\varepsilon}_{j2}\mathbf{\Omega}(X_{i1})'\ddot{Z}_{j2}\varsigma(X_{i1},X_{j1})+o_p(1)\\
=& N^{-1}N_s^{-1/2}\sum_j\sum_{i\neq j}\widetilde{\varepsilon}_{j2}\mathbf{\Omega}(X_{i1})'\ddot{Z}_{j2}\varsigma(X_{i1},X_{j1})+o_p(1).
\end{aligned}
$$

The second last equality holds from $\sqrt{N}(c_h+h)^2=o(1)$ under the rate condition in Assumption D.1 and the additional condition $Nh^4=o(1)$ stated in the theorem. The last equality holds because

$$
\begin{aligned}
&N^{-1}N_s^{-1/2}\sum_{j=1}^{N}\mathbf{\Omega}'(X_{j1})\ddot{Z}_{j2}\widetilde{\varepsilon}_{j2}\varsigma(X_{j1},X_{j1})\\
\leq&\sup_{x\in\mathcal{X}_1}\left\|f_{X_1}^{-1}(x)\right\|\sup_{x\in\mathcal{X}_1}\left\|\mathbf{\Omega}'(X_{j1})\right\|\iota_1'\mathbf{M}^{-1}\iota_1\kappa_h(0)N^{-1}N_s^{-1/2}\sum_{j=1}^{N}\ddot{Z}_{j2}\widetilde{\varepsilon}_{i2}=o_p(1).
\end{aligned}
$$

By letting $\zeta(X_{j1})=N_s^{-1}\sum_{i:i\neq j}\mathbf{\Omega}'(X_{i1})\varsigma(X_{i1},X_{j1})$, we have

$$
\sqrt{N_s}(\widehat{\vartheta}_2-\widetilde{\vartheta}_2)=(N_s/N)^{1/2}\cdot N^{-1/2}\sum_{j=1}^{N}\zeta(X_{j1})\widetilde{\varepsilon}_{j2}\ddot{Z}_{j2}+o_p(1). \tag{E.9}
$$

Let $\overline{\zeta}(x)=\mathbb{E}[\zeta(x)|X_{i1}\in\mathcal{X}_1^*]=\mathbb{E}[\mathbf{\Omega}'(X_{i1})\kappa_h(X_{i1}-x)f_{X_1}^{-1}(X_{i1})\widetilde{X}_{j1}(X_{i1})'\mathbf{D}^{-1}\mathbf{M}^{-1}\iota_1|X_{i1}\in\mathcal{X}_1^*]=\mathbb{E}[\mathbf{\Omega}'(X_{i1})\kappa_h(X_{i1}-x)f_{X_1}^{-1}(X_{i1})|X_{i1}\in\mathcal{X}_1^*]$ when $\ell=0,1$. It is easy to show that $\overline{\zeta}(x)=p^{-1}\int_{\mathcal{X}_1^*}\mathbf{\Omega}'(t)\kappa_h(t-x)dt$ and therefore $\overline{\zeta}(X_{j1})=p^{-1}\int_{\mathcal{X}_1^*}\kappa_h(t-X_{j1})\mathbf{\Omega}'(t)dt$.

Since $\mathbb{E}[\zeta(X_{j1})\widetilde{\varepsilon}_{j2}\ddot{Z}_{j2}]=0$ by the exclusion restriction, we know that $\mathbb{E}[\overline{\zeta}(X_{j1})\widetilde{\varepsilon}_{j2}\ddot{Z}_{j2}]=$

0 and $\mathbb{E}[(\zeta(X_{j1}) - \overline{\zeta}(X_{j1}))\widetilde{\varepsilon}_{j2}\ddot{Z}_{j2}] = 0$ as well. In addition,

$$\mathbb{E}[\|N^{-1/2}\sum_j(\zeta(X_{j1}) - \overline{\zeta}(X_{j1}))\widetilde{\varepsilon}_{j2}\ddot{Z}_{j2}\|^2] \leq \mathbb{E}\left[\|\zeta(X_{j1}) - \overline{\zeta}(X_{j1})\|^2\mathbb{E}[\widetilde{\varepsilon}_{j2}^2\|\ddot{Z}_{j2}\|^2|X_{j1}]\right]$$

$$\leq O(1)\mathbb{E}[\|\zeta(X_{j1}) - \overline{\zeta}(X_{j1})\|^2] = O((Nh)^{-1}). \tag{E.10}$$

Then, by the Markov's inequality, we have

$$N^{-1/2}\sum_{j=1}^N \zeta(X_{j1})\widetilde{\varepsilon}_{j2}\ddot{Z}_{j2} - N^{-1/2}\sum_{j=1}^N \overline{\zeta}(X_{j1})\widetilde{\varepsilon}_{j2}\ddot{Z}_{j2} = o_p(1). \tag{E.11}$$

Combining this with (E.9), we have

$$\sqrt{N_s}(\widehat{\vartheta}_2 - \widetilde{\vartheta}_2) = (N_s/N)^{1/2} \cdot N^{-1/2}\sum_{j=1}^N \overline{\zeta}(X_{j1})\widetilde{\varepsilon}_{j2}\ddot{Z}_{j2} + o_p(1). \tag{E.12}$$

with $\mathbb{E}[\overline{\zeta}(X_{j1})\widetilde{\varepsilon}_{j2}\ddot{Z}_{j2}] = 0$. In addition, when $\ell = 0, 1$,

$$\begin{aligned}
&\mathbb{E}[\overline{\zeta}(X_{j1})\widetilde{\varepsilon}_{j2}^2\ddot{Z}_{j2}\ddot{Z}_{j2}'\overline{\zeta}'(X_{j1})] \\
&= p^{-2}\int\int_{\mathcal{X}_1^*}\int_{\mathcal{X}_1^*}\kappa_h(t - X_{j1})\kappa_h(\widetilde{t} - X_{j1})\mathbf{\Omega}'(t)\mathbb{V}(\widetilde{\varepsilon}_{j2}\ddot{Z}_{j2}|X_{j1})\mathbf{\Omega}(\widetilde{t})dtd\widetilde{t}dF_{X_1}(X_{j1}) \\
&= p^{-2}\int\int\kappa(u)\kappa(u - s)u(u - s)'duds \\
&\quad\times\int_{\mathcal{X}_1^*}\mathbf{\Omega}'(w)\mathbb{V}(\widetilde{\varepsilon}_{j2}\ddot{Z}_{j2}|X_{j1} = w)\mathbf{\Omega}(w)dF_{X_1}(w) + o(1) \\
&= p^{-1}\mathbf{\Sigma}_1^* + o(1).
\end{aligned}$$

Moreover, we note that, because $\overline{\zeta}(.)$ is uniformly bounded,

$$\mathbb{E}[\|\overline{\zeta}(X_{j1})\|^{2+\delta}\|\varepsilon_{j2}\ddot{Z}_{j2}\|^{2+\delta}] \leq O(1)\mathbb{E}[\|\overline{\zeta}(X_{j1})\|^{2+\delta}] = O(1),$$

and by combining it with the boundedness of the variance of $\overline{\zeta}(X_{j1})\widetilde{\varepsilon}_{j2}\ddot{Z}_{j2}$, one can show that the Lyapunov's condition holds. Therefore, the following is obtained by applying the Lyapunov CLT to (E.12) and $N_s/N \to p$;

$$(N_s/N)^{1/2} \cdot N^{-1/2}\sum_j \overline{\zeta}(X_{j1})\widetilde{\varepsilon}_{j2}\ddot{Z}_{j2} \to_d N(0, \mathbf{\Sigma}_1^*) \tag{E.13}$$

In addition, because $\widetilde{\vartheta}_2 - \vartheta_2 = N_s^{-1}\sum_{i:X_{i1}\in\mathcal{X}_1}(\theta_2(X_{i1}) - \vartheta_2) = N_s^{-1}\sum_i(\theta_2(X_{i1}) - \mathbb{E}[\theta_2(X_{i1})|X_{i1} \in \mathcal{X}_1])$, and then by the CLT, we have

$$\sqrt{N_s}(\widetilde{\vartheta}_2 - \vartheta_2) \to_d N(0, \mathbb{V}[\theta_2(X_{i1})|X_{i1} \in \mathcal{X}_1^*]) \overset{d}{=} N(0, \mathbf{\Sigma}_2^*). \tag{E.14}$$

Then, we have $\sqrt{N_s}(\widehat{\vartheta}_2 - \vartheta_2) \to_d N(0, \mathbf{\Sigma}_1^* + \mathbf{\Sigma}_2^*)$ that follows from (E.13) and (E.14) and the fact that $\widetilde{\vartheta}_2 - \vartheta_2$ is a function of X_{i1} only and thus $\mathbb{E}[(\widetilde{\vartheta}_2 - \vartheta_2)'\zeta(X_{i1})\ddot{Z}_{i2}\widetilde{\varepsilon}_{i2}] = 0$.

Proof of Proposition 4.2

The suggested estimator of $\mathbf{\Sigma}_1^*$ is as follows,

$$\widehat{\mathbf{\Sigma}}_1^* = \widehat{p} \cdot N^{-1} \sum_j \widehat{\varepsilon}_{j2}^2 \widehat{\zeta}(X_{j1}) \ddot{Z}_{j2} \ddot{Z}_{j2}' \widehat{\zeta}'(X_{j1}), \tag{E.15}$$

where $\widehat{\zeta}(x) = N_s^{-1} \sum_{i:X_{i1} \in \mathcal{X}_1} \widehat{f}_{X_1}^{-1}(X_{i1}) \kappa_h(X_{i1} - x) \widehat{\mathbf{\Omega}}'(X_{i1})$ and $\widehat{\varepsilon}_{i2} = \widetilde{\varepsilon}_{i2} - \ddot{X}_{i2}'(\widehat{\theta}_2(X_{i1}) - \theta_2(X_{i1}))$. Then, because of (D.3) in Lemma D.2, (E.2), and the fact that $\iota_1' \mathbf{M}^{-1} \widetilde{X}_{j1}(.)$ reduces to 1 for all j the following holds uniformly in $x \in \mathcal{X}_1$:

$$\|\widehat{\zeta}(x) - \zeta(x)\| \leq O_p(1) \sup_{\widetilde{x} \in \mathcal{X}_1^*} \left(|\widehat{f}_{X_1}^{-1}(\widetilde{x}) - f_{X_1}^{-1}(\widetilde{x})| + \|\widehat{\mathbf{\Omega}}(\widetilde{x}) - \mathbf{\Omega}(\widetilde{x})\| \right) = o_p(1). \tag{E.16}$$

In addition, because of Lemma A.1 in Fan and Huang (2005), $\zeta(.)$ satisfies that

$$\sup_{x \in \mathcal{X}_1} \|\zeta(x) - \overline{\zeta}(x)\| = O_p(c_h), \tag{E.17}$$

The uniform convergence results in (E.16) and (E.17) imply that

$$\sup_{x \in \mathcal{X}_1} \|\widehat{\zeta}(x) - \overline{\zeta}(x)\| = o_p(1). \tag{E.18}$$

We for the moment define $\widetilde{\mathbf{\Sigma}}_1^*$ as follows.

$$\widetilde{\mathbf{\Sigma}}_1^* = \widehat{p} \cdot N^{-1} \sum_j \overline{\zeta}(X_{j1}) \widetilde{\varepsilon}_{j2}^2 \ddot{Z}_{j2} \ddot{Z}_{j2}' \overline{\zeta}'(X_{j1}).$$

Then, by the LLN for the i.i.d. data and the probability limit of $\widehat{p}$, we have

$$\widetilde{\mathbf{\Sigma}}_1^* \to_p \mathbf{\Sigma}_1^*. \tag{E.19}$$

Moreover, because of (E.5), (E.18), (E.19), Assumption D.1.(c), the uniform boundedness of $\overline{\zeta}(.)$, and $N^{-1} \sum_j \|\ddot{X}_{j2}\|^2 \|\ddot{Z}_{j2}\|^2 = O_p(1)$ (Assumption D.1.(f)), we find that

$$\|\widehat{\mathbf{\Sigma}}_1^* - \widetilde{\mathbf{\Sigma}}_1^*\| \leq O_p(1) \sup_{x \in \mathcal{X}_1} \left(\|\widehat{\zeta}(x) - \overline{\zeta}(x)\| + \|\widehat{\theta}_2(x) - \theta_2(x)\| \right) = o_p(1). \tag{E.20}$$

Therefore, the consistency of $\widehat{\mathbf{\Sigma}}_1^*$ can be obtained from (E.19) and (E.20).

The suggested estimator of $\mathbf{\Sigma}_2^*$ is given by

$$\widehat{\mathbf{\Sigma}}_2^* = N_s^{-1} \sum_{i:X_{i1} \in \mathcal{X}_1} (\widehat{\theta}_2(X_{i1}) - \widehat{\vartheta}_2)(\widehat{\theta}_2(X_{i1}) - \widehat{\vartheta}_2)'. \tag{E.21}$$

Let $\widetilde{\mathbf{\Sigma}}_2^*$ be defined by

$$\widetilde{\mathbf{\Sigma}}_2^* = N_s^{-1} \sum_{i:X_{i1} \in \mathcal{X}_1} (\theta_2(X_{i1}) - \vartheta_2)(\theta_2(X_{i1}) - \vartheta_2)', \tag{E.22}$$

and then because of the LLN, it satisfies that

$$\widetilde{\mathbf{\Sigma}}_2^* - \mathbf{\Sigma}_2^* = o_p(1). \tag{E.23}$$

Note that $\widehat{\theta}_2(.) - \widehat{\vartheta}_2 = \widehat{\theta}_2(.) - \theta_2(.) + \theta_2(.) - \vartheta_2 + \vartheta_2 - \widehat{\vartheta}_2$ and because of the uniform boundedness of $\theta_2(.)$, (E.5) and Theorem 4.2, we find that

$$\|\widehat{\mathbf{\Sigma}}_2^* - \widetilde{\mathbf{\Sigma}}_2^*\| \leq O_p(1) \left(\|\widehat{\vartheta}_2 - \vartheta_2\| + \sup_{x \in \mathcal{X}_1^*} \|\widehat{\theta}_2(x) - \theta_2(x)\| \right) = o_p(1). \tag{E.24}$$

Then the consistency of $\widehat{\mathbf{\Sigma}}_2^*$ can be established by using (E.23) and (E.24).

References

ACEMOGLU, D., D. AUTOR, D. DORN, G. H. HANSON, AND B. PRICE (2016): "Import competition and the great US employment sag of the 2000s," Journal of Labor Economics, 34, S141–S198.

ADÃO, R., M. KOLESÁR, AND E. MORALES (2019): "Shift-Share Designs: Theory and Inference," The Quarterly Journal of Economics, 134, 1949–2010.

AHN, S. C. AND P. SCHMIDT (1995): "Efficient estimation of models for dynamic panel data," Journal of Econometrics, 68, 5–27.

ANDERSON, T. AND C. HSIAO (1982): "Formulation and estimation of dynamic models using panel data," Journal of Econometrics, 18, 47–82.

ARELLANO, M. AND S. BOND (1991): "Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations," The Review of Economic Studies, 58, 277–297.

AUTOR, D., D. DORN, AND G. HANSON (2013): "The China syndrome: Local labor market effects of import competition in the United States," American Economic Review, 103, 2121–2168.

AUTOR, D., D. DORN, G. HANSON, AND K. MAJLESI (2020a): "Importing political polarization? The electoral consequences of rising trade exposure," American Economic Review, 110, 3139–3183.

AUTOR, D., D. DORN, AND G. H. HANSON (2021): "On the persistence of the China shock," Working Paper 29401, National Bureau of Economic Research.

AUTOR, D., D. DORN, G. H. HANSON, G. PISANO, AND P. SHU (2020b): "Foreign Competition and Domestic Innovation: Evidence from US Patents," American Economic Review: Insights, 2, 357–374.

——— (2020c): "Foreign competition and domestic innovation: Evidence from US patents," American Economic Review: Insights, 2, 357–374.

Autor, D. H., D. Dorn, G. H. Hanson, and J. Song (2014): "Trade adjustment: Worker-level evidence," Quarterly Journal of Economics, 129, 1799–1860.

Battisti, M., G. Peri, and A. Romiti (2022): "Dynamic effects of co-ethnic networks on immigrants' economic success," The Economic Journal, 132, 58–88.

Bernstein, S., R. Diamond, A. Jiranaphawiboon, T. McQuade, and B. Pousada (2022): "The contribution of high-skilled immigrants to innovation in the United States," Working Paper 30797, National Bureau of Economic Research.

Bloom, N., M. Draca, and J. Van Reenen (2016): "Trade induced technical change? The impact of Chinese imports on innovation, IT and productivity," The Review of Economic Studies, 83, 87–117.

Bojinov, I., A. Rambachan, and N. Shephard (2021): "Panel experiments and dynamic causal effects: A finite population perspective," Quantitative Economics, 12(4), 1171–1196.

Borusyak, K., P. Hull, and X. Jaravel (2022): "Quasi-Experimental Shift-Share Research Designs," The Review of Economic Studies, 89, 181–213.

Boustan, L. P. (2010): "Was postwar suburbanization "white flight"? Evidence from the black migration," The Quarterly Journal of Economics, 125, 417–443.

Bravo, F. (2023): "Local polynomial estimation of nonparametric general estimating equations," Statistics & Probability Letters, 197, 109805.

Burchardi, K. B., T. Chaney, T. A. Hassan, L. Tarquinio, and S. J. Terry (2020): "Immigration, innovation, and growth," Tech. rep., National Bureau of Economic Research.

Cai, Z., Y. Fang, M. Lin, and J. Su (2019): "Inferences for a partially varying coefficient model with endogenous regressors," Journal of Business & Economic Statistics, 37, 158–170.

Cai, Z. and Q. Li (2008): "Nonparametric estimation of varying coefficient dynamic panel data models," Econometric Theory, 24, 1321–1342.

Calonico, S., M. D. Cattaneo, M. H. Farrell, and R. Titiunik (2019): "Regression discontinuity designs using covariates," The Review of Economics and Statistics, 101, 442–451.

Cellini, S. R., F. Ferreira, and J. Rothstein (2010): "The value of school facility investments: Evidence from a dynamic regression discontinuity design," The Quarterly Journal of Economics, 125, 215–261.

Chernozhukov, V., S. Lee, and A. M. Rosen (2013): "Intersection bounds: Estimation and inference," Econometrica, 81, 667–737.

Fan, J. and I. Gijbels (1996): Local polynomial modelling and its applications, Monographs on statistics and applied probability 66, London: Chapman and Hall/CR, first edition. ed.

Fan, J., W. Härdle, and E. Mammen (1998): "Direct estimation of low-dimensional components in additive models," The Annals of Statistics, 26, 943 – 971.

Fan, J. and T. Huang (2005): "Profile likelihood inferences on semiparametric varying-coefficient partially linear models," Bernoulli, 11, 1031 – 1057.

Fan, Y. and Q. Li (2003): "A Kernel-based Method for Estimating Additive Partially Linear Models," Statistica Sinica, 13, 739–762.

Feenstra, R. C., H. Ma, and Y. Xu (2019): "US exports and employment," Journal of International Economics, 120, 46–58.

Flores, C. and Flores-Lagunes (2009): "Identification and estimation of causal mechanisms and net effects of a treatment under unconfoundedness," working paper.

——— (2010): "Nonparametric partial identification of causal net and mechanism average treatment effects," working paper.

Gallen, Y., J. S. Joensen, E. R. Johansen, and G. F. Veramendi (2023): "The Labor Market Returns to Delaying Pregnancy," SSRN working paper.

Han, S. (2021): "Identification in nonparametric models for dynamic treatment effects," Journal of Econometrics, 225, 132–147.

Heckman, J. J., J. E. Humphries, and G. Veramendi (2016): "Dynamic treatment effects," Journal of Econometrics, 191, 276–292, innovations in Measurement in Economics and Econometrics.

Heckman, J. J. and S. Navarro (2007): "Dynamic discrete choice and dynamic treatment effects," Journal of Econometrics, 136, 341–396.

Hsu, Y.-C. and S. Shen (2023): "Dynamic Regression Discontinuity Under Treatment Effect Heterogeneity," University of California, Davis, Working Paper.

Mack, Y. P. and B. W. Silverman (1982): "Weak and strong uniform consistency of kernel regression estimates," Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete, 61, 405–415.

Su, L., I. Murtazashvili, and A. Ullah (2013): "Local linear GMM estimation of functional coefficient IV models with an application to estimating the rate of return to schooling," Journal of Business & Economic Statistics, 31, 184–207.