Identification and Estimation of Average Partial Effects in Semiparametric Binary Response Panel Models*

Laura Liu Alexandre Poirier Ji-Liang Shiu

September 11, 2022

Abstract

Average partial effects (APEs) are generally not point-identified in binary response panel models with unrestricted unobserved heterogeneity. We show their point-identification under an index sufficiency assumption on the unobserved heterogeneity, even when the error distribution is unspecified. This assumption does not impose parametric restrictions on the unobserved heterogeneity. We then construct a three-step semiparametric estimator for the APE. In the first step, we estimate the common parameters using either a conditional logit or smoothed maximum score estimator. In the second step, we estimate the conditional expectation of the outcomes given the indices and a generated regressor that depends on first-step estimates. In the third step, we average derivatives of this conditional expectation to obtain a partial mean that estimates the APE. We show that this proposed three-step APE estimator is consistent and asymptotically normal. We evaluate its finite-sample properties in Monte Carlo simulations. We then illustrate our estimator in a study of determinants of married women's labor supply.

Keywords: Average partial effects, panel data, binary response models, semiparametric estimation, unobserved heterogeneity

JEL classification: C13, C14, C23, C25

*Liu: Department of Economics, Indiana University, lauraliu@iu.edu. Poirier: Department of Economics, Georgetown University, alexandre.poirier@georgetown.edu. Shiu: Institute for Economic and Social Research, Jinan University, jishiu.econ@gmail.com. We thank Stéphane Bonhomme, Iván Fernández-Val, Jiaying Gu, Louise Laage, Simon Lee, Ying-Ying Lee, Oliver Linton, Francis Vella, Martin Weidner, and participants at various seminars and conferences for helpful comments and discussions. We also thank Iván Fernández-Val for making the data publicly available on his website. The authors are solely responsible for any remaining errors.

1 Introduction

Binary response panel models with unobserved individual heterogeneity are commonly used in empirical research. This paper is concerned with panels where the binary outcome is generated from the underlying latent variable model

$$Y_{it} = \mathbb{1}\left(X_{it}'\beta_0 + C_i - U_{it} \geq 0\right), \tag{1.1}$$

for units $i = 1, \ldots, N$ and time periods $t = 1, \ldots, T$. Here, X_{it} are covariates, C_i is scalar unobserved heterogeneity, and U_{it} are unobserved idiosyncratic errors. We assume N is large, but that T is small and fixed, as is the case in many microeconomic datasets. This class of models includes fixed effects models, random effects models, and models that impose various levels of structure on the conditional distribution of $C_i|\mathbf{X}_i$, where $\mathbf{X}_i = (X_{i1}, \ldots, X_{iT})$. See Wooldridge (2010) Chapter 15.8 for an exposition of such models.

Identification results for β_0 are well-known and go back to the work of Rasch (1960) in the case where U_{it} is assumed to be logistic and Manski (1987) when its distribution is not specified. However, in this paper we focus on the *average structural function* (ASF) and *average partial effects* (APE) in model (1.1). Introduced in Blundell and Powell (2003), the ASF at potential value $\underline{x}$ is the conditional response probability $\mathbb{P}(Y_{it} = 1|X_{it} = \underline{x}, C_i = c)$ averaged over the marginal unobserved heterogeneity distribution F_C. It is used to assess the average impact of interventions in which the value of X_{it} is manipulated. The APE is a derivative of the ASF with respect to one covariate, hence measuring the partial effect of this covariate on the conditional response probability averaged over the marginal distribution of C_i. Both the ASF and APE are commonly used to evaluate the causal impact of policies in binary outcome models.

When $C_i|\mathbf{X}_i$ is unrestricted, i.e., fixed effects are assumed, the ASF and APE are generally not point-identified. In this paper, we show the point-identification of the ASF and APE under an index sufficiency assumption on the unobserved heterogeneity that we propose. This assumption restricts this conditional distribution to depend on covariates only through $v(\mathbf{X}_i)$, a (multiple) index of $\mathbf{X}_i$. It is related to an assumption of Altonji and Matzkin (2005) and Bester and Hansen (2009) that they use to show the identification of the *local average response* (LAR). The LAR is different from the APE since it conditions on covariate values rather than average over the population: see the discussion below and in Section 2.2 for a comparison of these estimands. Under our assumption, $v(\mathbf{X}_i)$ acts as a control function that does not require the specification of a first-stage or the existence of an instrument. As in Imbens and Newey (2009), the support of this index variable plays an important role, which we study in detail.

Note that the identification results in this paper do not rely on parametric assumptions on the conditional

distribution of $C_i|\mathbf{X}_i$, nor on the distribution of U_{it}. Both the ASF and APE depend directly on the distribution of C_i, which is not specified. We show that even when this distribution is not identified, the ASF and APE can be identified despite their dependence on F_C.

In our second main contribution, we construct three-step semiparametric estimators for the ASF and APE. In particular, we show the ASF is the partial mean of the conditional expectation of Y_{it} given $(X_{it}'\beta_0, v(\mathbf{X}_i))$, integrated over the marginal distribution of $v(\mathbf{X}_i)$. The APE can be expressed as a derivative of the ASF. We use this partial mean structure to construct semiparametric estimators of the ASF and APE. In a preliminary step, we estimate β_0 using one of the many available estimators in the literature: e.g., the conditional maximum likelihood estimator (CMLE) of Rasch (1960) when U_{it} is logistic, the smoothed panel maximum score of Charlier, Melenberg, and van Soest (1995) and Kyriazidou (1995) when U_{it} is nonparametric, or other estimators we list in Section 3.1. In a second step, we estimate the above conditional expectation, replacing unobserved $X_{it}'\beta_0$ by generated regressor $X_{it}'\widehat{\beta}$. We then use a local polynomial regression to recover this conditional mean. In a final step, we average this estimated conditional mean over the empirical distribution of $v(\mathbf{X}_i)$. The APE estimator is analogous, replacing the conditional expectation estimate with an estimate of its derivative, which is obtained directly via the local polynomial regression.

Next, we provide rate conditions on bandwidths, the convergence rate of $\widehat{\beta}$, and on the order of the polynomial regression. These conditions allow us to establish the consistency and asymptotic normality of both the ASF and APE estimators. Their rates of convergence do not depend on the dimension of $\mathbf{X}_i$ but instead depend on the dimension of the index $v(\mathbf{X}_i)$ in the sufficiency assumption. Moreover, their rates of convergence are fast relative to other nonparametric estimators since, after integrating over the distribution of $v(\mathbf{X}_i)$, the ASF and APE are functions of one-dimensional $X_{it}'\beta_0$. For example, when the index is one-dimensional, the ASF's and APE's rates of convergence are similar to standard rates of convergence of univariate nonparametric kernel regression estimators, which are fast within the class of nonparametric estimators.[1]

In an extension of our main results, we show that under additional support assumptions on $(X_{it}'\beta_0, v(\mathbf{X}_i))$, the marginal distribution of C_i is nonparametrically identified when U_{it} is logistic. Under slightly stronger support assumptions, it can also be identified without specifying the distribution of U_{it}. This allows for the identification of functionals of the conditional response probability beyond its mean, such as its quantiles. We then discuss the connection between the local average response and the APE, and show the LAR is identified under weaker versions of our main assumptions. We also discuss an extension of our identification results to the ASF and APE in dynamic panel models (see, e.g., Honoré and Kyriazidou (2000)) in Section

[1] In particular, we show the ASF can converge at a rate faster than $N^{2/5}$ when the index is one-dimensional.

4.3.

In the Monte Carlo simulation experiments, we compare the proposed semiparametric estimator with a random effects (RE) estimator and a correlated random effects (CRE) estimator. Both of them are commonly used parametric approaches which assume that $C_i|V_i$ is Gaussian and that U_{it} is logistic (see the definitions at the beginning of Section 5). Results show that the semiparametric estimator yields smaller biases and larger standard deviations, and the former channel dominates when the true distribution of the unobserved heterogeneity is non-Gaussian and the true distribution of the idiosyncratic errors is non-logistic.

In the empirical illustration, we study women's labor force participation using our semiparametric approach. We see that the semiparametric APE estimates are closer to zero for lower husband's incomes and more negative for higher ones, while their parametric counterparts vary less with respect to husband's incomes. Additionally, the effects of the husband's income are no longer significant once we allow for the flexibility in the distributions of the unobserved heterogeneity and of the idiosyncratic errors.

Related Literature

We now review the related literature. While we focus on the ASF and APE, our work builds on a large literature on the identification and estimation of β_0 in model (1.1). This literature can further be subdivided based on its distributional assumptions on $C_i|\mathbf{X}_i$, and on those on U_{it}. The case where both $C_i|\mathbf{X}_i$ and U_{it} are parametrized is studied in Chamberlain (1980). In this case, the distribution of $Y_i|\mathbf{X}_i$ is fully parametrized and β_0 can be estimated via maximum likelihood. To compute the likelihood function, one integrates the distribution of $Y_i|\mathbf{X}_i, C_i$ over the parametric distribution of $C_i|\mathbf{X}_i$. This case includes random effects, where $C_i|\mathbf{X}_i \overset{d}{=} C_i$ and C_i follows a parametric distribution. See Chapter 15.8 in Wooldridge (2010) for a review of this approach.

With fixed effects, the identification of β_0 in a binary panel with logit errors goes back to the work of Rasch (Rasch, 1960, 1961). In this work, he finds a sufficient statistic for C_i that allows β_0 to be estimated by maximizing the observations' likelihood conditional on this statistic. Andersen (1970) derives the asymptotic properties of the $\sqrt{N}$-consistent conditional logit estimator. See also Chamberlain (1980). In the case where errors are not logistic, Manski (1987) shows the identification of β_0 up to scale. He also presents a consistent maximum score estimator for β_0 that does not converge at a $\sqrt{N}$-rate. Charlier, Melenberg, and van Soest (1995) and Kyriazidou (1995) propose a smoothed version of Manski's estimator which converges at a faster rate. The impossibility of $\sqrt{N}$-estimation when the errors' distribution are not specified is described in Magnac (2004) and Chamberlain (2010). This impossibility can be overcome by making additional assumptions though. For example, in the presence of a special regressor, Honoré and

Lewbel (2002) show the $\sqrt{N}$-estimation of β_0 without specifying U_{it}'s distribution. See also Lee (1999) and Chen, Si, Zhang, and Zhou (2017) for alternative assumptions that restore $\sqrt{N}$-consistency.

In our paper, we consider an intermediate assumption on $C_i|\mathbf{X}_i$, where we assume this distribution depends on $\mathbf{X}_i$ only through a potentially multivariate index. We do not parametrize the distribution of $C_i|\mathbf{X}_i$, nor do we restrict how it depends on this index. As mentioned above, our primary focus is on the ASF and APE rather than β_0.

Due to our conditional independence assumption between the heterogeneity and covariates conditional on an index, our work is also related to a large literature on control functions. Newey, Powell, and Vella (1999) show the identification of structural functions in a triangular model, where a control variable V_i is identified from a first stage. Blundell and Powell (2004) consider a binary response model with endogeneity, and focus on the identification of the ASF. Imbens and Newey (2009) consider a nonseparable triangular model and, like us, focus on the identification of functionals of the structural function, such as the ASF. For results on the ASF in binary panels, Maurer, Klein, and Vella (2011) use a semiparametric maximum likelihood approach and a control function assumption to identify and estimate the ASF. Also see Laage (2020) for a panel data model with triangular endogeneity.

Although they consider a different estimand, the work of Altonji and Matzkin (2005) and Bester and Hansen (2009) is closely related to ours. In Altonji and Matzkin (2005), they consider an exchangeability assumption, where $F_{C|X_1,\ldots,X_T}$ is invariant to relabeling of the time indices on the regressors. They then assume that $C_i|X_{it}, v(\mathbf{X}_i) \overset{d}{=} C_i|v(\mathbf{X}_i)$ where $v(\mathbf{X}_i)$ are known symmetric functions of $(X_{i1},\ldots,X_{iT})$. They consider a nonparametric outcome equation, and show the identification of the LAR, an object which averages changes in the conditional response probability over the *conditional* distribution of the heterogeneity. This object differs from the APE since it integrates over the conditional distribution of $C_i|X_{it}$ rather than its marginal distribution. We discuss in more detail in Remark 2.3 the difference in estimands, and related differences in assumptions on the support of the index are discussed in Section 4.2. Unlike Altonji and Matzkin (2005), our structural equation (1.1) depends on index $X_{it}'\beta_0$ which allows for much faster rates of convergence for our APE when compared to the rates obtained for the LAR in their nonparametric outcome equation. In particular, their rate of convergence for their LAR estimator decreases with the dimension of X_{it} while the rate of convergence of our APE estimator is related to the dimension of $X_{it}'\beta_0$, which is fixed. In Bester and Hansen (2009) they also consider an index sufficiency assumption where $v(\mathbf{X}_i) = (v_1(\mathbf{X}_i^{(1)}),\ldots,v_{d_X}(\mathbf{X}_i^{(d_X)}))$, but where the indices $\{v_j(\cdot)\}_{j=1}^{d_X}$ are allowed to be unknown.[2] Their identification results are for the LAR of continuous regressors, which we view as complementary to our

[2]The dimension of X_{it} is denoted by d_X, and $\mathbf{X}_i^{(k)}$ denotes a $T \times 1$ vector with the kth components of X_{it} for $t = 1,\ldots,T$.

results on APEs.

Other identification approaches in these models have also been proposed. Chernozhukov, Fernández-Val, Hahn, and Newey (2013) derive bounds on the ASF in fixed effects binary response models with nonparametric distributions of $C_i|\mathbf{X}_i$ and of U_{it}. Davezies, D'Haultfoeuille, and Laage (2021) derive bounds for average marginal effects when U_{it} is assumed to be logistic. Fernández-Val (2009) proposes bias-corrected estimators of marginal effects when T is large and when U_{it} follows a normal distribution.

Our estimator is a three-step semiparametric estimator. The first step involves the estimation of the common parameters β_0, the second step is a nonparametric regression including a generated regressor, and the third step marginalizes over a subset of the regressors. Such estimators are called partial means. See Newey (1994) for seminal work on the estimation of partial means without generated regressors. The estimation of partial means with generated regressors is studied in Mammen, Rothe, and Schienle (2012), Mammen, Rothe, and Schienle (2016), and Lee (2018).

Finally, while we focus on the static case, there is a large literature on dynamic binary response models going back to Cox (1958). In particular, see Chamberlain (1985), Magnac (2000), Honoré and Kyriazidou (2000) and, more recently, Honoré and Weidner (2020) and Kitazawa (2021). These papers all focus on the identification and estimation of common coefficients, rather than for the ASF or APE. In recent work, Aguirregabiria and Carro (2020) show that the average marginal effect of changes in the lagged outcome are point-identified when U_{it} follows a logistic distribution. Also under a logit assumption, Dobronyi, Gu, and Kim (2021) characterize the identified set for the underlying distribution of individual effects and some of its functionals.

The remainder of this paper is organized as follows. In Section 2 we present the baseline model and provide our main identification results. In Section 3 we establish the asymptotic properties of our proposed ASF and APE estimators. Section 4 extends some of our identification results to more general settings. In Section 5 we conduct Monte Carlo experiments to study the finite-sample properties of our estimators. Section 6 applies our APE estimator to an empirical illustration on female labor force participation. Finally, Section 7 concludes. The appendix contains the proofs for all propositions and theorems, as well as supplemental tables and graphs.

2 Model and Identification

In this section, we describe the general binary response panel model of interest. We will consider two sets of assumptions: one in which errors are logistic, and the other in which the error distribution is not

parametrized. In both cases, the coefficients on regressors are known to be identified under standard conditions. Then, we show the identification of the ASF and APE under an index sufficiency assumption on the distribution of the heterogeneity.

2.1 General Model

Recall the baseline model in equation (1.1)

$$Y_{it} = \mathbb{1}\left(X_{it}'\beta_0 + C_i - U_{it} \geq 0\right),$$

where $i = 1, \ldots, N$, $t = 1, \ldots, T$. Here, $X_{it} \in \mathcal{X}_t \subseteq \mathbb{R}^{d_X}$ is a d_X-vector of covariates and $\beta_0 \in \mathcal{B} \subseteq \mathbb{R}^{d_X}$ is a d_X-vector of unknown parameters. Let $\mathbf{X}_i \in \mathcal{X} \subseteq \mathbb{R}^{T \times d_X}$ denote the observed covariate matrix which has X_{it}' as its tth row. Let $C_i \in \mathcal{C} \subseteq \mathbb{R}$ denote the unobserved individual-specific heterogeneity, and U_{it} are idiosyncratic errors. Let $Y_i = (Y_{i1}, \ldots, Y_{iT})$ denote the vector of outcomes for unit i. The i subscript is suppressed in the remainder of this section and when there is no confusion.

We make the following assumptions on the baseline model.

Assumption A1 (Model assumptions).

(i) Y_t is generated according to equation (1.1);

(ii) $(\mathbf{X}, C) \perp\!\!\!\perp U_1 \perp\!\!\!\perp \cdots \perp\!\!\!\perp U_T$;

(iii) For $t = 1, \ldots, T$, U_t has a continuous density with respect to the Lebesgue measure that we denote by f_{U_t}. This density is bounded.

Besides assuming model equation (1.1) holds, A1 also imposes that unobserved variables $\{U_t\}_{t=1}^T$ are independent from covariates and the individual-specific heterogeneity. This is a strict exogeneity assumption on $\mathbf{X}$ given C. This rules out the presence of lagged dependent variables in $\mathbf{X}$. We relax this assumption and consider models with lagged dependent variables in Section 4.3. The relationship between C and $\mathbf{X}$ is unrestricted by A1. We also assume $\{U_t\}_{t=1}^T$ to be mutually independent, ruling out serial correlation. Assumption A1.(iii) is a standard regularity condition.

Let F_{U_t} denote the cumulative distribution function of U_t. The *conditional response probability* is

$$\mathbb{P}(Y_t = 1 | X_t = \underline{x}, C = c) = \mathbb{P}(U_t \leq \underline{x}'\beta_0 + c | X_t = \underline{x}, C = c) = F_{U_t}(\underline{x}'\beta_0 + c).$$

For $t = 1, \ldots, T$ and $\underline{x} \in \mathcal{X}_t$, let the ASF be the conditional response probability integrated over the

marginal distribution of the unobserved effect C:

$$\text{ASF}_t(\underline{x}) = \int_{\mathcal{C}} \mathbb{P}\left(Y_t = 1 | X_t = \underline{x}, C = c\right) \, dF_C(c) = \int_{\mathcal{C}} F_{U_t}(\underline{x}'\beta_0 + c) \, dF_C(c). \tag{2.1}$$

As defined in Blundell and Powell (2003), the ASF is the average conditional response probability, where the averaging occurs over the marginal distribution of C. The ASF differs from the identified conditional probability $\mathbb{P}(Y_t = 1 | X_t = \underline{x})$ due to the dependence between C and $\mathbf{X}$. Note that under $C \perp\!\!\!\perp X$, a random effects assumption, they are equal.

Our main object of interest is the APE, which measures the partial effect of changing one covariate, averaged over the marginal distribution of C. If this covariate is continuously distributed, the APE will be defined as the derivative of the ASF with respect to this covariate. If the covariate has a discrete distribution, the APE will be defined as the difference between two ASFs evaluated at different support points.

More concretely, define the APE of the kth element of $\underline{x} \in \mathcal{X}_t$, denoted by $\underline{x}^{(k)}$, as follows in the case where $\underline{x}^{(k)}$ is continuously distributed:

$$\text{APE}_{k,t}(\underline{x}) = \frac{\partial}{\partial \underline{x}^{(k)}} \int_{\mathcal{C}} \mathbb{P}\left(Y_t = 1 | X_t = \underline{x}, C = c\right) \, dF_C(c) = \beta_0^{(k)} \cdot \int_{\mathcal{C}} f_{U_t}(\underline{x}'\beta_0 + c) \, dF_C(c), \tag{2.2}$$

where $\beta_0^{(k)}$ is the kth element of β_0.

In the case where $X_t^{(k)}$ is discretely distributed, we let the APE be the difference between the ASF at two values, which can be viewed as an average treatment effect. We denote these values by $\underline{x}$ and $\widetilde{\underline{x}}_k = \underline{x} + e_k(\widetilde{\underline{x}}^{(k)} - \underline{x}^{(k)})$, where e_k denotes a vector of zeros with a 1 in position k, and where $\widetilde{\underline{x}}^{(k)}$ is another value in the support of $X_t^{(k)}$. We let

$$\text{APE}_{k,t}(\underline{x}, \widetilde{\underline{x}}_k) = \text{ASF}_t(\widetilde{\underline{x}}_k) - \text{ASF}_t(\underline{x}). \tag{2.3}$$

2.1.1 Identification of β_0

Before discussing the estimation of the ASF and APE, we first consider assumptions that lead to their point identification. As a preliminary step, we focus on the identification of the coefficients β_0, which itself relies on previously established results. We now present two cases in which β_0 is identified.

Logit Case. The leading special case we consider is one where U_t follows a standard logistic distribution, as in Rasch (1960). Under minimal support constraints on $\mathbf{X}$, assuming logistic errors entails the point identification and $\sqrt{N}$-consistent estimation of β_0, even with fixed effects. In fact, as shown in Chamberlain (2010), the logistic error distribution is the only error family that allows for $\sqrt{N}$-consistent estimation

without requiring further assumptions. Here we present assumptions which, together with A1, yield the point identification of β_0.

Assumption A2 (Logit case).

(i) For $t = 1, \ldots, T$, U_t follows a standard Logistic distribution: $F_{U_t}(u) = \frac{e^u}{1+e^u} \equiv \Lambda(u)$;

(ii) There exist $s, s' \in \{1, \ldots, T\}$ such that $X_s - X_{s'}$ does not lie in a proper linear subspace of $\mathbb{R}^{d_X}$.

Assumption A2.(ii) makes weak assumptions on the within-unit variation in covariates but does rule out time-invariant regressors.

General Case: Nonparametric Errors. The most general case we consider is one where the distribution of U_t is not specified. This case does not nest the logit case, since it requires the presence of continuous regressors while the logit case does not. Hence we consider it separately. This case was studied in Manski (1987), where the following assumption is made.

Assumption A2′ (General case) There exist $s, s' \in \{1, \ldots, T\}$ such that

(i) U_s and $U_{s'}$ are continuously distributed with positive density everywhere on $\mathbb{R}$;

(ii) The support of $X_s - X_{s'}$ does not lie in a proper linear subspace of $\mathbb{R}^{d_X}$;

(iii) For some $k \in \{1, \ldots, d_X\}$, $\beta_0^{(k)} \neq 0$ and $X_s^{(k)} - X_{s'}^{(k)}$ has positive density on $\mathbb{R}$ conditional on $\{X_s^{(1)} - X_{s'}^{(1)}, \ldots, X_s^{(k-1)} - X_{s'}^{(k-1)}, X_s^{(k+1)} - X_{s'}^{(k+1)}, \ldots, X_s^{(d_X)} - X_{s'}^{(d_X)}\}$ with probability one.

Like in Assumption A2.(i), the support of U_t is the entire real line. Assumption A2′.(iii) imposes that one regressor has support on the entire real line, which was not required in A2. This full support assumption on $X_s^{(k)} - X_{s'}^{(k)}$ is key to this semiparametric identification result.

Before presenting a result on the identification of the ASF and APE, we first present a lemma that shows the identification of β_0 in the above two cases.

Lemma 2.1 (Identification of β_0)**.** Suppose A1 holds. Suppose the distribution of $(Y, \mathbf{X})$ is known. If A2 holds, then β_0 is point identified. If A2′ holds, then β_0 is identified up to scale.

We mainly state this lemma for completeness, since it is a combination of results obtained in earlier work. In particular, derivations for the first part of this lemma can be found in Chamberlain (1980). The second part of this lemma is a restatement of Lemma 2 in Manski (1987). The identification of β_0 is used in a preliminary step to identify the ASF and APE, as we show below.

2.2 Identification of the ASF and APE

Without further assumptions, it is generally impossible to point identify the ASF or the APE, even under parametric assumptions on U_t.[3] Point identification is obtained if random effects are assumed: $C \perp\!\!\!\perp \mathbf{X}$. Under fixed effects, the ASF and APE are partially identified. To see this, denote by $G_t(\underline{x}'\beta_0, \mathbf{x})$ the counterfactual conditional probability

$$G_t(\underline{x}'\beta_0, \mathbf{x}) \equiv \int_{\mathcal{C}} F_{U_t}(\underline{x}'\beta_0 + c)\, dF_{C|\mathbf{X}}(c|\mathbf{x}).$$

Note that we observe the following conditional probabilities:

$$\mathbb{P}(Y_t = 1|\mathbf{X} = \mathbf{x}) \equiv G_t(x_t'\beta_0, \mathbf{x}), \qquad \text{for } \mathbf{x} \in \mathcal{X},\ t \in \{1, \ldots, T\}.$$

By the law of total probability, the ASF for covariate value $\underline{x}$ is

$$\begin{aligned}
\text{ASF}_t(\underline{x}) &= \int_{\mathcal{C}} F_{U_t}(\underline{x}'\beta_0 + c)\, dF_C(c) \\
&= \int_{\mathcal{X}} G_t(\underline{x}'\beta_0, \mathbf{x})\, dF_{\mathbf{X}}(\mathbf{x}) \qquad (2.4) \\
&= \int_{\mathcal{X}} \underbrace{G_t(\underline{x}'\beta_0, \mathbf{x})\mathbb{1}(x_t'\beta_0 = \underline{x}'\beta_0)}_{=\mathbb{P}(Y_t=1|\mathbf{X}=\mathbf{x})\mathbb{1}(x_t'\beta_0=\underline{x}'\beta_0)} dF_{\mathbf{X}}(\mathbf{x}) + \int_{\mathcal{X}} \underbrace{G_t(\underline{x}'\beta_0, \mathbf{x})\mathbb{1}(x_t'\beta_0 \neq \underline{x}'\beta_0)}_{\text{not point-identified}}\, dF_{\mathbf{X}}(\mathbf{x}). \qquad (2.5)
\end{aligned}$$

We can see from equation (2.4) that the ASF is an average over the distribution of $\mathbf{X}$ of conditional probability $G_t(\underline{x}'\beta_0, \mathbf{X})$. In order for the ASF to be point-identified, we need $G_t(\underline{x}'\beta_0, \mathbf{x})$ to be identified for $x \in \mathcal{X}$, but this generally fails since, given $X_t'\beta_0 = \underline{x}'\beta_0$, the support of $\mathbf{X}$ does not equal its marginal support. In equation (2.5), the $G_t(\underline{x}'\beta_0, \mathbf{x})$ where $x_t'\beta_0 \neq \underline{x}'\beta_0$ are counterfactual probabilities that are not point-identified from the data since they do not correspond to any conditional probability of Y_t given $\mathbf{X} = \mathbf{x}$. Unless restrictions are imposed on the distribution of $C|\mathbf{X}$, this causes the ASF, and therefore the APE too, to be partially identified. For the logit case, see Davezies, D'Haultfoeuille, and Laage (2021) for partial identification results for average marginal effects. In the nonparametric case, bounds on the ASF are obtained in Chernozhukov, Fernández-Val, Hahn, and Newey (2013).

To achieve point identification, in this paper we instead consider an index sufficiency restriction that imposes additional structure on the conditional distribution of the heterogeneity. Specifically, we make assumptions such that $G_t(\underline{x}'\beta_0, \mathbf{x})$ depends on indices of $\mathbf{x}$ that have common support given $X_t'\beta_0 = \underline{x}'\beta_0$.

Assumption A3 (Index sufficiency).

(i) Let $V \equiv v(\mathbf{X})$, where $v : \mathbb{R}^{T \times d_X} \to \mathbb{R}^{d_V}$ is known. Let $C \mid \mathbf{X} \overset{d}{=} C \mid V$;

[3] One exception is the point identification of the average marginal effects associated with the lagged dependent variable in the fixed effect logit AR(1) model, shown in Aguirregabiria and Carro (2020).

(ii) Let $\underline{x} \in \text{supp}(X_t)$. One of the following two assumptions holds:

(a) Let $X_t^{(k)}$ be continuously distributed in a neighborhood of $\underline{x}^{(k)}$ given $\{X_t^{(-k)} = \underline{x}^{(-k)}\}$.[4] There exists a neighborhood $\mathcal{N}$ of $\underline{x}'\beta_0$ such that $\mathcal{N} \times \text{supp}(V) \subseteq \text{supp}(X_t'\beta_0, V)$.

(b) Let $X_t^{(k)}$ be discretely distributed given $\{X_t^{(-k)} = \underline{x}^{(-k)}\}$. Let $\widetilde{\underline{x}}_k'\beta_0 = \underline{x}'\beta_0 + (\widetilde{\underline{x}}^{(k)} - \underline{x}^{(k)})\beta_0^{(k)}$. Let $\{\underline{x}'\beta_0, \widetilde{\underline{x}}_k'\beta_0\} \times \text{supp}(V) \subseteq \text{supp}(X_t'\beta_0, V)$.

Before proceeding with our identification result, we offer a discussion of this assumption.

Discussion of Assumption A3.(i)

Part (i) of this assumption is a correlated random effects assumption which restricts the conditional distribution of $C|\mathbf{X}$ to depend solely on $v(\mathbf{X})$, an index of $\mathbf{X}$. The conditional distribution of $C|v(\mathbf{X})$ remains nonparametric though. Such index assumption is considered in Altonji and Matzkin (2005), and in Bester and Hansen (2009).

In Altonji and Matzkin (2005), the exchangeability of $f_{C|\mathbf{X}}(c|x_1, \ldots, x_T)$ in $(x_1, \ldots, x_T)$ is assumed. In the case where X_t is scalar, they consider symmetric polynomials as candidates for the index function, e.g., $v(\mathbf{X}) = \left(\sum_{t=1}^T X_t, \sum_{1 \leq t_1 < t_2 \leq T} X_{t_1} X_{t_2}\right)$ when the indices are the first two elementary symmetric functions. Unlike us, Bester and Hansen (2009) do not assume $v(\cdot)$ is known, but they do not allow for the indices to be arbitrary functions of $\mathbf{X}$: each component on the index may only depend on one component of $\mathbf{X}_t$. The focus of these two papers is also different from ours: they identify the LAR rather than the ASF or APE, and its identification is shown for continuous covariates. See Remark 2.3 for an explicit comparison between the APE and LAR.

Treatment assignment models in panel data can be used to find candidate indices. This is explored in Arkhangelsky and Imbens (2019). For example, consider the following treatment assignment model when X_t is binary: Assume that $X_t = \mathbb{1}(E_t \leq \nu(C))$ where E_t are iid and independent of C, and ν is an arbitrary function. In this case, $(X_1, \ldots, X_T)|C$ are iid Bernoulli variables, hence $v(\mathbf{X}) = \sum_{t=1}^T X_t$ is a sufficient statistic that satisfies $\mathbf{X}|C, v(\mathbf{X}) \overset{d}{=} \mathbf{X}|v(\mathbf{X})$ which implies A3.(i) holds. This fact can be easily derived from the Fisher-Neyman factorization theorem. This result is generalized in Arkhangelsky and Imbens (2019) to cases where the distribution of $\mathbf{X}|C$ is from an exponential family with known sufficient statistic. For example, if $(X_1, \ldots, X_T)|C$ are assumed iid Gaussian, then $v(\mathbf{X}) = (\sum_{t=1}^T X_t, \sum_{t=1}^T X_t^2)$ would also form a sufficient statistic.

[4] The notation $X_t^{(-k)}$ is used to denote the X_t vector with its kth entry removed.

In a special case where the indices are time-averages, i.e., $v(\mathbf{X}) = \frac{1}{T}\sum_{t=1}^{T} X_t$, the index assumption is consistent with $C = \zeta\left(\left(\frac{1}{T}\sum_{t=1}^{T} X_t\right)'\gamma_0, \eta\right)$ where $\eta \perp\!\!\!\perp \mathbf{X}$ and $\zeta(\cdot,\cdot)$ is any function. This is related to the specification of the conditional distribution of C given $\mathbf{X}$ in Mundlak (1978), although we do not specify the distribution of η, nor restrict the functional form of $\zeta(\cdot,\cdot)$. In this specification for C, the one-dimensional index $\tilde{v}(\mathbf{X}) = \sum_{t=1}^{T} X_t'\gamma_0$ also satisfies A3.(i), but is unknown due to its dependence on unknown γ_0. Lower-dimensional indices usually lead to faster rates of convergence, but we show in Section 3 that the ASF rate of convergence can be the standard nonparametric rate of $N^{2/5}$ by letting the polynomial order in the local polynomial regression be large enough. Therefore, we can see that multidimensional indices obviate the need to identify and later estimate such γ_0. Estimated and unknown indices could in principle be accommodated by using the work of Ichimura and Lee (1991) on the identification and estimation of multiple index models. We leave extensions to unknown $v(\cdot)$ for future research.

Discussion of Assumption A3.(ii)

Assumption A3.(ii).(a) implies that $X_t'\beta_0$ is continuously distributed in a neighborhood $\mathcal{N}$ of $\underline{x}'\beta_0$. This allows for some components of X_t to be discretely distributed. Note that we also do not require $X_t'\beta_0$ to be supported on the entire real line. We do require that the support of the sufficient statistic is independent of the value of $X_t'\beta_0$ in a neighborhood of $\underline{x}'\beta_0$. Letting $\mathcal{V} = \text{supp}(V)$ and $\mathcal{V}_t(u) = \text{supp}(V|X_t'\beta_0 = u)$, this is stated as $\mathcal{V}_t(u) = \mathcal{V}$ for $u \in \mathcal{N}$. This support assumption is related to the common support assumption of Imbens and Newey (2009), although we only restrict the support of $V|X_t'\beta_0$ rather than the support of $V|\mathbf{X}$. As opposed to Imbens and Newey (2009), we do not posit the existence of a first stage or of exogenous excluded variables, since our indices are functions of $\mathbf{X}$ only.

Altonji and Matzkin (2005) do not require this support assumption to show the identification of the LAR, since this object is conditional on $X_t = \underline{x}$. To better see this difference, in their nonparametric setting, identification of the ASF or APE would require $\text{supp}(v(\mathbf{X})|X_t = \underline{x}) = \text{supp}(v(\mathbf{X}))$, which is significantly stronger than our condition whenever more than one covariate is present. To see this, assume $d_X = T = 2$, $X_t^{(1)}$ is continuously distributed on $\mathbb{R}$, and that $X_t^{(2)} \in \{0,1\}$ is binary. Let $v(\mathbf{X}) = \sum_{t=1}^{2} X_t = (\sum_{t=1}^{2} X_t^{(1)}, \sum_{t=1}^{2} X_t^{(2)})$. Then, under minimal assumptions, $\text{supp}(v(\mathbf{X})) = \mathbb{R} \times \{0,1,2\}$ but $\text{supp}(v(\mathbf{X})|X_1 = \underline{x}) = \mathbb{R} \times \{\underline{x}^{(2)}, \underline{x}^{(2)} + 1\} \neq \text{supp}(v(\mathbf{X}))$. On the other hand, the conditional support of $v(\mathbf{X})$ given $\{X_t'\beta_0 = x_t'\beta_0\}$ equals $\text{supp}(v(\mathbf{X}))$ when $\beta_0^{(1)} \neq 0$. Therefore, in this example the ASF/APE will be identified under our assumptions in the semiparametric model.

This important condition also has implications on the dimension of $v(\mathbf{X})$. For example, this condition is violated when $v(\mathbf{X}) = \mathbf{X}$, i.e., no index restriction are imposed and, equivalently, we have fixed effects. This

is because the support of $\mathbf{X}$ does not equal its conditional support given $X_t'\beta_0$: $\text{supp}(\mathbf{X}|X_t'\beta_0 = \underline{x}'\beta_0) \neq \text{supp}(\mathbf{X})$. In the case where $v(\mathbf{X}) = \sum_{t=1}^T X_t \in \mathbb{R}^{d_X}$, this condition is written as $\text{supp}(\sum_{t=1}^T X_t|X_t'\beta_0 = u) = \text{supp}(\sum_{t=1}^T X_t)$. For example, we can see that this holds in the simple case where $(X_1, \ldots, X_T)$ are jointly normally distributed. In Remark 2.2 below we show that while this support condition may not always be warranted, the ASF and APE are partially identified when it fails.

Assumption A3.(ii).(b) is used in when the covariate of interest is discrete, and ensures that the ASF is identified at the two values $\underline{x}$ and $\widetilde{\underline{x}}_k$. When these ASFs are identified, the APE, which is a difference in ASF, is also identified.

Finally, we note that the validity of these assumptions depends on the value of unknown β_0. However, β_0 is point identified from the data in the logit case, and point identified up to scale in the non-logistic case. Assumption A3.(ii) does not depend on the scale normalization, therefore it only depends on observable random variables and identified parameters. Hence, it is a falsifiable assumption.

With these assumptions, we can show that the ASF and APE are point identified.

Theorem 2.1. Suppose A1 and A3 hold. Suppose either A2 or A2′ hold. Suppose the distribution of $(Y, \mathbf{X})$ is known. Then, $\text{ASF}_t(\underline{x})$ and $\text{APE}_{k,t}(\underline{x})$ are point identified.

We now provide some intuition for this result. Since β_0 is identified, $\mathbb{P}(Y_t = 1|X_t'\beta_0 = \underline{x}'\beta_0, V = v)$ is identified for all $v \in \mathcal{V}$ by A3. Then, we can write

$$\begin{aligned}
\text{ASF}_t(\underline{x}) &= \int_{\mathcal{C}} F_{U_t}(\underline{x}'\beta_0 + c)\, dF_C(c) \\
&= \int_{\mathcal{V}} \int_{\mathcal{C}} F_{U_t}(\underline{x}'\beta_0 + c)\, dF_{C|V}(c|v)\, dF_V(v) \\
&= \int_{\mathcal{V}} \mathbb{P}(Y_t = 1|X_t'\beta_0 = \underline{x}'\beta_0, V = v)\, dF_V(v). \qquad (2.6)
\end{aligned}$$

The second equality follows from the law of total probability and the last one from the index restriction. Equation (2.6) depends only on $\{\mathbb{P}(Y_t = 1|X_t'\beta_0 = \underline{x}'\beta_0, V = v) : v \in \mathcal{V}\}$ and on the marginal distribution of V, which are both identified from the data.

To identify the APE for a continuous regressor, we note that A3.(ii).(a) implies the ASF is point identified for values of X_t near $\underline{x}$. Since the APE is a derivative of the ASF, we can identify the APE as a limit of finite differences between ASFs. Formally, we can write

$$\begin{aligned}
\text{APE}_{k,t}(\underline{x}) &= \frac{\partial}{\partial \underline{x}^{(k)}} \int_{\mathcal{V}} \mathbb{P}(Y_t = 1|X_t'\beta_0 = \underline{x}'\beta_0, V = v)\, dF_V(v) \\
&= \beta_0^{(k)} \cdot \int_{\mathcal{V}} \frac{\partial}{\partial u} \mathbb{P}(Y_t = 1|X_t'\beta_0 = u, V = v)|_{u=\underline{x}'\beta_0}\, dF_V(v). \qquad (2.7)
\end{aligned}$$

All quantities in equation (2.7) are identified, hence the APE is identified. Note that the identification of the ASF and APE bypasses the need to identify F_C, the distribution of the heterogeneity. In Section 4.1 below, we show that F_C is identified under stronger support assumptions on $(X_t'\beta_0, V)$.

Remark 2.1 (Excluded control variable)**.** A more general version of A3 is that we can *identify* a variable V such that $C|\mathbf{X}, V \overset{d}{=} C|V$. This is a control variable assumption, where V may be an unobservable that is not functionally related to X, as is the case in A3.(i). For example, it could be the residual in a first-stage equation relating X to some excluded instruments, whose existence we do not assume in this paper. See for example Imbens and Newey (2009) in the nonseparable cross-sectional case, or Laage (2020) for a panel model with triangular endogeneity and control functions. If this control variable satisfies Assumption A3, our identification results still apply. As for estimation, if V is identified from a first-stage equation, we should substitute $\widehat{V}_i$ for V_i, where $\widehat{V}_i$ is a suitable estimator for the control variable. This generated regressor's impact on the limiting distribution would then have to be taken into account.[5] In this paper, we focus on the case where no such V is observed or identified from a first-stage model, and instead where V is an index of $\mathbf{X}$.

Remark 2.2 (Relaxing support assumptions)**.** The validity of support assumption A3.(ii) depends intricately on the support of $\mathbf{X}$, and on the considered value $\underline{x}$. For a given value of $\underline{x}$, it is possible this assumption fails. When it does, we can show that the ASF and APE are partially identified instead. Recall that $\mathcal{V}_t(u) = \text{supp}(V|X_t'\beta_0 = u)$, and assume that $\mathcal{V}_t(\underline{x}'\beta_0) \subsetneq \mathcal{V}$. Then, the conditional probability $\mathbb{P}(Y_t = 1|X_t'\beta_0 = \underline{x}'\beta_0, V = v)$ is identified for all $v \in \mathcal{V}_t(\underline{x}'\beta_0)$. Therefore, following Theorem 4 in Imbens and Newey (2009), the identified set for the ASF is contained in

$$\begin{aligned}
\text{ASF}_t(\underline{x}) &\in \left[\underline{\text{ASF}}_t(\underline{x}), \overline{\text{ASF}}_t(\underline{x})\right] \\
&\equiv \left[\int_{\mathcal{V}_t(\underline{x}'\beta_0)} \mathbb{P}(Y_t = 1|X_t'\beta_0 = \underline{x}'\beta_0, V = v)\, dF_V(v),\right. \\
&\qquad \left. \int_{\mathcal{V}_t(\underline{x}'\beta_0)} \mathbb{P}(Y_t = 1|X_t'\beta_0 = \underline{x}'\beta_0, V = v)\, dF_V(v) + \mathbb{P}(V \in \mathcal{V} \setminus \mathcal{V}_t(\underline{x}'\beta_0))\right].
\end{aligned}$$

The width of these bounds depends only on the probability that V is outside of $\mathcal{V}_t(\underline{x}'\beta_0)$, which is small if V is continuously distributed and the measure of $\mathcal{V} \setminus \mathcal{V}_t(\underline{x}'\beta_0)$ is close to zero. Hence, small violations of the support condition yield a narrow identified set. These bounds can be used to construct bounds on the APE with discrete covariates as well:

$$\text{APE}_{k,t}(\underline{x}, \widetilde{x}_k) \in \left[\underline{\text{ASF}}_t(\widetilde{x}_k) - \overline{\text{ASF}}_t(\underline{x}), \overline{\text{ASF}}_t(\widetilde{x}_k) - \underline{\text{ASF}}_t(\underline{x})\right].$$

[5] Note that we take into account the generated regressor $X_t'\widehat{\beta}$'s impact in this paper.

To obtains bounds on the APE for a continuous covariate, an upper bound on the density f_{U_t} is needed. While it may not be easy to postulate such a bound, in the logit case this density attains a maximal value of 1 at the origin. Therefore, these bounds on f_{U_t} yield the following APE bounds:

$$\begin{aligned}\text{APE}_{k,t}(\underline{x}) \in \Bigg[&\beta_0^{(k)} \cdot \int_{\mathcal{V}} \frac{\partial}{\partial u}\mathbb{P}(Y_t = 1|X_t'\beta_0 = u, V = v)|_{u=\underline{x}'\beta_0}\, dF_V(v), \\ &\beta_0^{(k)} \cdot \int_{\mathcal{V}} \frac{\partial}{\partial u}\mathbb{P}(Y_t = 1|X_t'\beta_0 = u, V = v)|_{u=\underline{x}'\beta_0}\, dF_V(v) + \beta_0^{(k)} \cdot \mathbb{P}(V \in \mathcal{V} \setminus \mathcal{V}_t(\underline{x}'\beta_0))\Bigg].\end{aligned}$$

The estimation methods we provide below for the point-identified ASF or APE can easily be adapted to estimate these bounds under violations of Assumption A3.(ii).

Remark 2.3 (LAR)**.** The LAR is defined as

$$\text{LAR}_{k,t}(\underline{x}) = \int \frac{\partial}{\partial \underline{x}^{(k)}}\mathbb{P}(Y_t = 1|X_t = \underline{x}, C = c)\, dF_{C|X_t}(c|\underline{x}), \tag{2.8}$$

whereas the APE measures changes in the ASF, which is the average value of Y_t when X_t has been exogenously set to $\underline{x}$. Thus, the APE is analogous to average treatment effects (ATE) in the causal inference literature, which averages the difference between two potential outcomes over its unconditional distribution; meanwhile, the LAR is analogous to a local treatment effect, where the averaging occurs over the conditional distribution of the heterogeneity given $X_t = \underline{x}$. Note that neither estimand is more general, since knowledge of the LAR for all $\underline{x} \in \mathcal{X}_t$ does not imply knowledge of APEs, and vice-versa. See Abrevaya and Hsu (2021) for a comparison of various causal estimands in panel models with unobserved heterogeneity. Section 4.2 further shows that the LAR can be identified under weaker index support assumptions than the APE.

3 Estimation

We now propose estimators for the ASF and APE, and establish their limiting distributions. The estimators we construct are sample analogs of (2.6) for the ASF, and of (2.7) for the APE. We show that the rate of convergence of the ASF estimator is similar to that of a kernel regression estimator with one continuous regressor. The APE estimator converges at the same rate as a derivative of a kernel regression estimator with one continuous regressor. In particular, we show the ASF converges at the rate $\sqrt{Nb_N}$ and the APE at the rate $\sqrt{Nb_N^3}$ where b_N is a scalar bandwidth used in the estimation of the conditional expectation of Y_t. We describe below in Assumption B6 what assumptions b_N must satisfy. These rates of convergence are obtained from our estimator being a partial mean, where we average over all components of the conditional expectation of $\mathbb{E}[Y_t|X_t'\beta_0, V]$, except for one. The rate of convergence does not depend on d_X or T, the dimensions of $\mathbf{X}$.

Throughout this section, we assume that we observe a random sample of $(Y_i, \mathbf{X}_i)$ of size N.

Assumption B1 (IID). $\{(Y_i, \mathbf{X}_i)\}_{i=1}^N$ are iid.

We start by considering the estimation of β_0, which forms the first step of our semiparametric estimator.

3.1 Estimation of β_0

Here, we again consider two cases. In the first case, the distribution of U_t is assumed to be logistic and the CMLE is used to estimate β_0. In the second case, the distribution of U_t is not specified. In this case, β_0 can be estimated by the smoothed maximum score estimator, or using a variety of $\sqrt{N}$-consistent estimators if stronger assumptions are made. The convergence of $\widehat{\beta}$ needs to be relatively fast to establish the limiting distributions of the ASF and APE estimators. In particular, convergence rates equal to or slower than $N^{1/3}$ are not compatible with our rate assumption B6 below. This rules out the maximum score estimator of Manski (1987), as Kim and Pollard (1990) show its cube-root convergence. These estimators' rate of convergence and associated regularity conditions are analyzed in prior work, so we only offer a brief description here.

Logit Case. In the case where U_t follows a logistic distribution, we can use the conditional maximum likelihood estimator (CMLE) of Rasch (1960) and Andersen (1970). To define the estimator, let $n_i = \sum_{t=1}^T Y_{it}$. Then, under equation (1.1) and Assumption A1,

$$\mathbb{P}(Y_i = (y_1, \ldots, y_T) \mid \mathbf{X}_i, n_i) = \frac{\exp(\sum_{t=1}^T y_t X_{it}' \beta_0)}{\sum_{\mathbf{d} \in D_{n_i}} \exp(\sum_{t=1}^T d_t X_{it}' \beta_0)},$$

where

$$D_{n_i} = \left\{ \mathbf{d} \in \{0,1\}^T : \sum_{t=1}^T d_t = n_i \right\}.$$

We define the CMLE as follows:

$$\widehat{\beta} = \underset{\beta \in \mathcal{B}}{\operatorname{argmax}} \prod_{i=1}^N \frac{\exp(\sum_{t=1}^T Y_{it} X_{it}' \beta)}{\sum_{\mathbf{d} \in D_{n_i}} \exp(\sum_{t=1}^T d_t X_{it}' \beta)}.$$

As is well known (Andersen, 1970), this estimator is $\sqrt{N}$-consistent for β_0 under standard regularity conditions.

General Case. When U_t is not assumed to be logistic, the CMLE is inconsistent. If U_t instead satisfies Assumption A2′, there exist alternative estimators for β_0. Without imposing further assumptions, these estimators' rate of convergence is slower than $\sqrt{N}$: See Chamberlain (2010).

One such estimator is the conditional smoothed maximum score estimator of Charlier, Melenberg, and van Soest (1995) and Kyriazidou (1995), a generalization of Horowitz (1992) to the panel model of Manski (1987). The estimator is defined by

$$\widehat{\beta} = \underset{\beta \in \mathcal{B}}{\operatorname{argmax}} \frac{1}{N} \sum_{i=1}^{N} \sum_{1 \leq s < t \leq T} (Y_{is} - Y_{it}) \cdot \widetilde{I}\left(\frac{(X_{is} - X_{it})'\beta}{h_N} \right),$$

where $\widetilde{I}(\cdot/h_N)$ is a smooth function that approximates the indicator function as $h_N \to 0$. Since β_0 is only identified up to scale, we can normalize the absolute value of the first element of $\widehat{\beta}$ to be unity. As in Horowitz (1992), if $\widetilde{I}$ can be represented as the integral of a νth order kernel, $\nu \geq 2$, the rate of convergence of $\widehat{\beta}$ is $N^{\frac{\nu}{2\nu+1}}$.

While $\sqrt{N}$-estimation of β_0 is generally not possible without specifying U_t's distribution, there are alternative assumptions and estimators that allow for it. In particular, Lee (1999) considers an "index increment sufficiency" assumption: $(X_t'\beta_0, C)|X_t - X_s \overset{d}{=} (X_t'\beta_0, C)|(X_t - X_s)'\beta_0$. Honoré and Lewbel (2002) assume the presence of a special regressor among X_t. Chen, Si, Zhang, and Zhou (2017) assume that $C = v(\mathbf{X}) + \zeta$, where ζ satisfies $(U_1, \ldots, U_T, \zeta) \perp\!\!\!\perp \mathbf{X}$. In all three papers, $\sqrt{N}$-consistent estimators for β_0 are proposed.

Both Cases. To accommodate the above two cases, as well as the variety of estimators considered within the general case, we make the following high-level assumption on the preliminary estimator $\widehat{\beta}$.

Assumption B2 (First-stage estimator). The estimator $\widehat{\beta}$ satisfies

$$a_N \|\widehat{\beta} - \beta_0\| = O_p(1),$$

where $a_N = O(N^{\epsilon})$ for some $\epsilon > 0$.

This assumption holds with $\epsilon = 1/2$ in the case where U_t has a logistic distribution and the CMLE is used to estimate β_0. When the smoothed maximum score estimator is used to estimate β_0, this assumption holds with $\epsilon = \nu/(2\nu + 1)$, where ν is the order of the kernel used to estimate $\widehat{\beta}$. The rate of convergence of this preliminary estimator will play a role in Assumption B6 below.

3.2 A Semiparametric Estimator of the ASF

We now present the ASF estimator and show its consistency and asymptotic normality under our assumptions. As mentioned earlier, this estimator is a three-step estimator. Section 3.1 examined the first step, which estimates the common parameters using either a conditional logit or smoothed maximum score estimator. We now describe the second and third steps, which estimate the ASF using a sample

analog of equation (2.6). In the second step, we nonparametrically estimate the conditional expectation $\mathbb{E}[Y_t|X_t'\beta_0 = \underline{x}'\beta_0, V = v]$ using a local polynomial regression of Y_t on generated regressor $X_t'\widehat{\beta}$ and V. In the final step, we evaluate the estimated conditional expectation at $(\underline{x}'\widehat{\beta}, V_i)$ for $i = 1, \ldots, N$, and then average over the empirical marginal distribution of V_i. To define this estimator, let $Z_t(\beta) = (X_t'\beta, V) \in \mathbb{R}^{1+d_V}$ and denote $Z_t = Z_t(\beta_0)$. Throughout the paper, we use z to denote $z = (u, v) \in \mathbb{R}^{1+d_V}$ where $u \in \mathbb{R}$ and $v \in \mathbb{R}^{d_V}$. In the rest of this section, we assume that V's components are all continuously distributed and that V has dimension d_V. In our analysis, the number of discrete components of V does not affect the rate of convergence. When the number of support points for the discrete components is small, we can handle these discrete components by performing a cell-by-cell analysis. Alternatively, they can be accommodated through a discrete kernel, for example as in Racine and Li (2004) equation (2.3). We omit these cases for notational simplicity.

We consider a local polynomial regression of order $\ell \geq 0$. The notation that follows is similar to that in Masry (1996). For $s \in \{0, 1, \ldots, \ell\}$, let $N_s = \binom{s+d_V}{d_V}$ be the number of distinct $(1 + d_V)$-tuples $r \in \mathbb{R}^{1+d_V}$ such that $|r| \equiv \sum_{k=1}^{1+d_V} |r_k| = s$. We arrange these $(1 + d_V)$-tuples in a lexicographical order with the highest priority given to the last position, so that $(0, \ldots, 0, s)$ is the first element and $(s, 0, \ldots, 0)$ is the last element in this sequence. We let τ_s denote this one-to-one mapping. This mapping satisfies $\tau_s(1) = (0, \ldots, 0, s), \ldots, \tau_s(N_s) = (s, 0, \ldots, 0)$. For each $s \in \{0, 1, \ldots, \ell\}$, define $N_s \times 1$ vector $\xi_s(a)$ by its kth element $a^{\tau_s(k)}$, where $k \in \{1, \ldots, 1 + d_V\}$ and $a \in \mathbb{R}^{1+d_V}$. Here we used the notation $a^b = a_1^{b_1} \times \cdots \times a_{d_V}^{b_{d_V}}$. Let

$$\xi(a) = (1, \xi_1(a)', \ldots, \xi_\ell(a)')' \in \mathbb{R}^{\bar{N}},$$

where $\bar{N} = \sum_{s=0}^{\ell} N_s$.

Let $\mathcal{K} : \mathbb{R}^{1+d_V} \to \mathbb{R}$ denote a $(1 + d_V)$-dimensional kernel. Let $\mathcal{K}_b(z) = b^{-(1+d_V)}\mathcal{K}(z)$, where $b > 0$ is a scalar bandwidth. Let b_N denote a sequence of bandwidths that converges to zero.

Let

$$\begin{aligned}\widehat{h}(z; \widehat{\beta}) &= \underset{h \in \mathbb{R}^{\bar{N}}}{\operatorname{argmin}} \sum_{j=1}^{N} \left(Y_{jt} - \sum_{0 \leq |r| \leq \ell} \left(\frac{Z_{jt}(\widehat{\beta}) - z}{b_N} \right)^r h_r \right)^2 \mathcal{K}_{b_N}\left(\frac{Z_{jt}(\widehat{\beta}) - z}{b_N} \right) \\ &= \underset{h \in \mathbb{R}^{\bar{N}}}{\operatorname{argmin}} \sum_{j=1}^{N} \left(Y_{jt} - \xi\left(\frac{Z_{jt}(\widehat{\beta}) - z}{b_N} \right)' h \right)^2 \mathcal{K}_{b_N}\left(\frac{Z_{jt}(\widehat{\beta}) - z}{b_N} \right).\end{aligned}$$

As $\widehat{\beta} \xrightarrow{p} \beta_0$, the vector $\widehat{h}(z; \widehat{\beta})$ estimates coefficients in a Taylor expansion of degree ℓ of the conditional expectation of Y_t given $Z_t(\beta_0) = z$. In particular, the first component of this vector, denoted by $\widehat{h}_1(z; \widehat{\beta}) = e_1'\widehat{h}(z; \widehat{\beta})$, is an estimator of the conditional mean of Y_t given $(X_t'\beta_0, V)$. This estimator is a least-squares

estimator and can be written as

$$\widehat{h}(z;\widehat{\beta}) = S_N(z;\widehat{\beta})^{-1} T_N(z;\widehat{\beta}),$$

where

$$S_N(z;\beta) = \frac{1}{N}\sum_{j=1}^{N} \xi\left(\frac{Z_{jt}(\beta)-z}{b_N}\right)\xi\left(\frac{Z_{jt}(\beta)-z}{b_N}\right)' \mathcal{K}_{b_N}\left(\frac{Z_{jt}(\beta)-z}{b_N}\right)$$
$$T_N(z;\beta) = \frac{1}{N}\sum_{j=1}^{N} \xi\left(\frac{Z_{jt}(\beta)-z}{b_N}\right) Y_{jt}\mathcal{K}_{b_N}\left(\frac{Z_{jt}(\beta)-z}{b_N}\right).$$

In analogy to equation (2.6), we average this conditional mean over the empirical marginal distribution of V_i to obtain the ASF estimator:

$$\widehat{\text{ASF}}_t(\underline{x}) = \frac{1}{N}\sum_{i=1}^{N} \widehat{h}_1(\underline{x}'\widehat{\beta}, V_i;\widehat{\beta})\hat{\pi}_{it},$$

where $\widehat{h}_1(z;\widehat{\beta}) = e_1'\widehat{h}(z;\widehat{\beta})$ is the first component in $\widehat{h}(z;\widehat{\beta})$, $\hat{\pi}_{it} = \mathbb{1}((\underline{x}'\widehat{\beta}, V_i) \in \mathcal{Z}_t)$ is a trimming function, and $\mathcal{Z}_t$ is an appropriately selected compact set in which the density $f_{Z_t(\beta)}(z)$ is bounded away from zero. This trimming function prevents issues with the invertibility of $S_N(z;\widehat{\beta})$. Since $\mathcal{Z}_t$ is a fixed compact set, the parameter that is consistently estimated by $\widehat{\text{ASF}}_t$ is a trimmed ASF defined by

$$\begin{aligned}\text{ASF}_t^{\pi}(\underline{x}) &\equiv \mathbb{E}[\mathbb{E}[Y_t|X_t'\beta_0 = \underline{x}'\beta_0, V]\pi_t] \\ &= \int_{\mathcal{C}} F_{U_t}(\underline{x}'\beta_0 + c)\mathbb{P}((\underline{x}'\beta_0, V) \in \mathcal{Z}_t|C = c)\, dF_C(c).\end{aligned}$$

Here we let $\pi_{it} = \mathbb{1}((\underline{x}'\beta_0, V_i) \in \mathcal{Z}_t)$. Note that if $(\underline{x}'\beta_0, V) \in \mathcal{Z}_t$ with probability 1, $\text{ASF}_t^{\pi}(\underline{x}) = \text{ASF}_t(\underline{x})$ and the trimming does not alter the estimand. By expanding $\mathcal{Z}_t$ along with the sample size at a slow enough rate,[6] it is likely that $\text{ASF}_t(\underline{x})$ is consistently estimated by $\widehat{\text{ASF}}_t(\underline{x})$. Since fixed trimming is often employed in the partial mean literature (see, for example, Newey (1994) or more recently Lee (2018)) we focus on this approach.

To understand the effect of trimming on the estimand, note that $F_{U_t}(\underline{x}'\beta_0 + c) \in (0,1)$ is bounded. This means that if $\mathbb{P}((\underline{x}'\beta_0, V) \in \mathcal{Z}_t|C = c)$ is close to 1, the trimmed ASF will be close to the ASF. In particular, if $\mathbb{P}((\underline{x}'\beta_0, V) \in \mathcal{Z}_t|C) \in [1-\underline{\varepsilon}, 1]$ with probability 1, then

$$\text{ASF}_t(\underline{x}) \in \left[\text{ASF}_t^{\pi}(\underline{x}), \frac{\text{ASF}_t^{\pi}(\underline{x})}{1-\underline{\varepsilon}}\right] \tag{3.1}$$

are bounds on the true ASF that collapse to a point as $\underline{\varepsilon}$ approaches zero.

To obtain the limiting distribution of the ASF, we make the following assumptions. We begin with a

[6] This is sometimes called a vanishing, or random, trimming approach.

standard assumption on the kernel.

Assumption B3 (Kernel). The kernel $\mathcal{K}$ satisfies $\mathcal{K}(z) = K(u) \cdot \prod_{k=1}^{d_V} K(v_k)$ where $K : \mathbb{R} \to \mathbb{R}_{\geq 0}$ is such that (i) $K(u)$ is equal to zero for all u outside of a compact set, (ii) K is twice continuously differentiable on $\mathbb{R}$ with all these derivatives being Lipschitz continuous, (iii) $\int_{-\infty}^{\infty} K(u)\, du = 1$, (iv) K is symmetric.

Note that we do not require the use of higher-order kernels in this local polynomial regression.

To state the next assumption precisely, let $\mathcal{C}_m(\mathcal{A})$ denote the set of m-times continuously differentiable functions $f : \mathcal{A} \to \mathbb{R}$. Here m is an integer and $\mathcal{A}$ is a subset of $\mathbb{R}^{1+d_V}$. Denote the differential operator by

$$\nabla^\lambda = \frac{\partial^{|\lambda|}}{\partial z_1^{\lambda_1} \cdots \partial z_{1+d_V}^{\lambda_{1+d_V}}},$$

where $\lambda = (\lambda_1, \ldots, \lambda_{1+d_V}) \in \{0, 1, \ldots\}^{1+d_V}$ is comprised of nonnegative integers such that $\sum_{k=1}^{1+d_V} \lambda_k = |\lambda|$. For a given set $\mathcal{A}$, let

$$\|f\|_m^{\mathcal{A}} = \max_{|\lambda| \leq m} \sup_{z \in \text{int}(\mathcal{A})} \|\nabla^\lambda f(z)\|.$$

We omit the $\mathcal{A}$ superscript when it does not cause confusion. Next, we impose smoothness and regularity conditions on the distribution of $(Y_t, Z_t(\beta))$ for β in a neighborhood of β_0.

Assumption B4 (Smoothness). Let $\mathcal{B}_\varepsilon = \{\beta \in \mathcal{B} : \|\beta - \beta_0\| \leq \varepsilon\}$.

(i) There exists $\varepsilon > 0$ such that for all $\beta \in \mathcal{B}_\varepsilon$, $Z_t(\beta)$ has a density $f_{Z_t(\beta)}(z)$ with respect to the Lebesgue measure;

(ii) $f_{Z_t(\beta)}(z)$ and $\left\| \frac{\partial}{\partial \beta} f_{Z_t(\beta)}(z) \right\|$ are uniformly bounded and uniformly bounded away from zero for $z \in \mathcal{Z}_t$ and $\beta \in \mathcal{B}_\varepsilon$, where $\mathcal{Z}_t$ is a compact set;

(iii) $\left\| f_{Z_t(\beta_0)}(z) \right\|_{\ell+2}^{\mathcal{Z}_t} < \infty$ and $\|\mathbb{E}[Y_t | Z_t(\beta_0) = z]\|_{\ell+2}^{\mathcal{Z}_t} < \infty$;

(iv) $\underline{x}'\beta_0$ is in the interior of $\mathcal{Z}_{1t} \equiv \{e_1' z : z \in \mathcal{Z}_t\}$;

(v) $f_{Z_t(\beta_0)|Y_t}(z|y)$ exists and is bounded for $y \in \{0, 1\}$.

Assumptions (i) and (ii) ensure the boundedness and sufficient smoothness of the distribution of $f_{Z_t(\beta)}$ as a function of β in a neighborhood of β_0. Assumption (iii) ensures additional smoothness in z for the distribution of $Z_t(\beta_0)$. The degree of smoothness is linked to the degree of the polynomial in the local polynomial regression. Assumptions (iv) and (v) are standard technical assumptions. We also impose the following moment existence condition.

Assumption B5 (Moment existence). Let $\mathbb{E}[\|X_t\|^2] < \infty$.

The following rate conditions govern the rate of convergence of the bandwidth b_N to zero.

Assumption B6 (Bandwidth). For some $\kappa, \delta > 0$, let $b_N = \kappa \cdot N^{-\delta}$. When ℓ is odd, δ satisfies

$$\max\left\{\frac{1}{2\ell+3}, 1-2\epsilon\right\} < \delta < \min\left\{\frac{2\epsilon}{3+2d_V}, \frac{1}{1+2d_V}\right\}.$$

When ℓ is even, δ satisfies

$$\max\left\{\frac{1}{2\ell+5}, 1-2\epsilon\right\} < \delta < \min\left\{\frac{2\epsilon}{3+2d_V}, \frac{1}{1+2d_V}\right\}.$$

This assumption has joint implications on these four quantities: δ, the bandwidth's rate of convergence to zero; ϵ, the rate of convergence of $\widehat{\beta}$ to β_0; d_V, the dimension of V; and ℓ, the order of the local polynomial regression. One of these implications is on the minimal value of ϵ, the rate of convergence of $\widehat{\beta}$. The constraint $1-2\epsilon < \delta < \frac{2\epsilon}{3+2d_V}$ implies that $\epsilon > \frac{3+2d_V}{8+4d_V}$. This is satisfied for all d_V when $\epsilon = 1/2$, e.g., when $\widehat{\beta}$ is estimated by the CMLE. In the case when $d_V = 1$, it requires that $\epsilon > 5/12$. These rate considerations motivate our earlier discussion of the smoothed maximum score estimator's convergence rate. This estimator satisfies the rate assumption when $\nu > \frac{3}{2} + d_V$. Hence, a kernel of order $\nu = 4$ for the smoothed maximum score estimator satisfies our assumptions when $d_V \leq 2$.

Another consequence of this assumption is that ℓ must increase as d_V increases. In particular, we require $\ell > d_V$ when $\epsilon = 1/2$, which corresponds to the rate of convergence of the CMLE. Thus we must consider higher-order local polynomials if the dimension of V is large.

We can now state the main convergence result for the ASF.

Theorem 3.1 (ASF asymptotics)**.** Suppose the assumptions of Theorem 2.1 hold. Suppose Assumptions B1–B6 hold. Then,

$$\sqrt{Nb_N}\left(\widehat{\text{ASF}}_t(\underline{x}) - \text{ASF}_t^{\pi}(\underline{x})\right) \xrightarrow{d} \mathcal{N}(0, \sigma^2_{\text{ASF}_t}(\underline{x}'\beta_0)),$$

where

$$\sigma^2_{\text{ASF}_t}(u) = \mathbb{E}\left[\text{Var}(Y_t|X_t'\beta_0 = u, V)\frac{f_V(V)}{f_{Z_t(\beta_0)}(u,V)}\mathbb{1}((u,V)\in\mathcal{Z}_t)\right]$$
$$\cdot e_1'\left(\int \xi(z)\xi(z)'\mathcal{K}(z)\,dz\right)^{-1}\int\left(\int \mathcal{K}(z)\,\xi(z)\,dv\right)\left(\int \mathcal{K}(z)\,\xi(z)\,dv\right)'du\left(\int \xi(z)\xi(z)'\mathcal{K}(z)\,dz\right)^{-1}e_1.$$

To understand the limiting distribution of this estimator, we break down its sampling variation into four separate sources. The terms associated with three of these are asymptotically negligible under our

assumptions. We can write

$$\begin{aligned}\sqrt{Nb_N}\left(\widehat{\mathrm{ASF}}_t(\underline{x})-\mathrm{ASF}_t^{\pi}(\underline{x})\right) &= \sqrt{Nb_N}\left(\frac{1}{N}\sum_{i=1}^{N}\left(\widehat{h}_1(\underline{x}'\widehat{\beta},V_i;\widehat{\beta})-\widehat{h}_1(\underline{x}'\widehat{\beta},V_i;\beta_0)\right)\hat{\pi}_{it}\right)\\ &+\sqrt{Nb_N}\left(\frac{1}{N}\sum_{i=1}^{N}\left(\widehat{h}_1(\underline{x}'\widehat{\beta},V_i;\beta_0)-\widehat{h}_1(\underline{x}'\beta_0,V_i;\beta_0)\right)\hat{\pi}_{it}\right)\\ &+\sqrt{Nb_N}\left(\frac{1}{N}\sum_{i=1}^{N}\widehat{h}_1(\underline{x}'\beta_0,V_i;\beta_0)(\hat{\pi}_{it}-\pi_{it})\right)\\ &+\sqrt{Nb_N}\left(\frac{1}{N}\sum_{i=1}^{N}\widehat{h}_1(\underline{x}'\beta_0,V_i;\beta_0)\pi_{it}-\mathbb{E}[h_1(\underline{x}'\beta_0,V;\beta_0)\pi_t]\right).\end{aligned}$$

The first term reflects the impact of the generated regressors $X_t'\widehat{\beta}$ being used instead of $X_t'\beta_0$. The bandwidth constraints involving ϵ—the rate of convergence of $\widehat{\beta}$ to β_0—ensure this term is asymptotically negligible. The second term reflects the impact of the approximation of the evaluation point $\underline{x}'\beta_0$ by $\underline{x}'\widehat{\beta}$. Once again, ϵ plays a crucial role and this term is asymptotically negligible as it is of asymptotic order $O_p(\sqrt{Nb_N}a_N^{-1})=o_p(1)$ by our assumptions. The third term pertains to the estimation of the trimming function π_{it} by $\hat{\pi}_{it}$. This term is asymptotically dominated due to the superconsistency of $\hat{\pi}_{it}$ to π_{it} uniformly in $i=1,\ldots,N$. The fourth and final term asymptotically dominates the other three and converges in distribution to a mean-zero Gaussian variable at the $\sqrt{Nb_N}$ rate. Some of the technical tools we use to show this convergence in distribution build on Masry (1996) and Kong, Linton, and Xia (2010).

The rate of convergence of $\widehat{\mathrm{ASF}}_t(\underline{x})$ when $\epsilon=1/2$ is $N^{\tilde{\delta}_{\mathrm{ASF}}}$, where $\tilde{\delta}_{\mathrm{ASF}}$ ranges in the interval $\left(\frac{1+d_V}{3+2d_V},\frac{1+\ell}{3+2\ell}\right)$. In the case where $d_V=1$ and $\ell=2$, this range corresponds to $\left(\frac{2}{5},\frac{3}{7}\right)$. Recall that 2/5 is the standard rate of convergence of univariate kernel estimation when using second-order kernels. We again note that this rate of convergence does not depend on either T or d_X, the dimension of X_t.

3.3 Semiparametric Estimation of the APE

We focus here on the case where $X_t^{(k)}$ is continuously distributed. When $X_t^{(k)}$ is discretely distributed, the APE is a difference between two ASFs, in which case Theorem 3.1 can be used to obtain its limiting distribution.

Let $\widehat{h}_2(z;\widehat{\beta})=\frac{1}{b_N}e_{2+d_V}'\widehat{h}(z;\widehat{\beta})$ denote the $(2+d_V)$-th component of the local polynomial regression coefficient vector. By the definition of the above lexicographical order, this is an estimator of the derivative of the conditional mean of Y_t given $(X_t'\beta_0,V)=(u,v)$ with respect to u. This estimated derivative is used

in the APE estimator, which is defined as

$$\widehat{\text{APE}}_{k,t}(\underline{x}) = \widehat{\beta}^{(k)} \cdot \frac{1}{N} \sum_{i=1}^{N} \widehat{h}_2(\underline{x}'\widehat{\beta}, V_i; \widehat{\beta})\hat{\pi}_{it},$$

where $\widehat{\beta}^{(k)}$ denotes the kth component of $\widehat{\beta}$.

As for the ASF, we use a trimming function in the estimator for technical reasons. Therefore, the estimator is consistent for a trimmed APE. This trimmed APE is defined as

$$\begin{aligned}
\text{APE}^{\pi}_{k,t}(\underline{x}) &\equiv \mathbb{E}\left[\frac{\partial}{\partial \underline{x}^{(k)}} \mathbb{E}[Y_t | X_t'\beta_0 = \underline{x}'\beta_0, V] \cdot \pi_t\right] \\
&= \beta_0^{(k)} \cdot \mathbb{E}\left[\frac{\partial}{\partial u} \mathbb{E}[Y_t | X_t'\beta_0 = u, V]|_{u=\underline{x}'\beta_0} \cdot \pi_t\right] \\
&= \beta_0^{(k)} \cdot \int_{\mathcal{C}} f_{U_t}(\underline{x}'\beta_0 + c)\mathbb{P}((\underline{x}'\beta_0, V) \in \mathcal{Z}_t | C = c)\, dF_C(c).
\end{aligned}$$

The difference between the trimmed and untrimmed APE depends solely on the probability $\mathbb{P}((\underline{x}'\beta_0, V) \in \mathcal{Z}_t | C = c) \in [0, 1]$. Therefore, the trimmed APE is attenuated relative to its untrimmed counterpart, and the size of the bias depends on the discrepancy between the above probability and 1. An analogous result to the ASF bounds in equation (3.1) can be easily constructed when knowledge of f_{U_t} is assumed, e.g., when U_t has a logistic distribution.

The following theorem shows that the APE is $\sqrt{Nb_N^3}$-consistent, where b_N is a bandwidth satisfying Assumption B6. Like the ASF, the APE's rate of convergence does not depend on the dimension of $\mathbf{X}$.

Theorem 3.2 (APE asymptotics)**.** Suppose the assumptions of Theorem 2.1 hold. Suppose Assumptions B1–B6 hold. Suppose $X_t^{(k)}$ is continuously distributed. Then,

$$\sqrt{Nb_N^3}\left(\widehat{\text{APE}}_{k,t}(\underline{x}) - \text{APE}^{\pi}_{k,t}(\underline{x})\right) \xrightarrow{d} \mathcal{N}\left(0, (\beta_0^{(k)})^2 \cdot \sigma^2_{\text{APE}_t}(\underline{x}'\beta_0)\right),$$

where

$$\begin{aligned}
&\sigma^2_{\text{APE}_t}(u) = \mathbb{E}\left[\text{Var}(Y_t | X_t'\beta_0 = u, V)\frac{f_V(V)}{f_{Z_t(\beta_0)}(u, V)}\mathbb{1}((u, V) \in \mathcal{Z}_t)\right] \\
&\cdot e_{2+d_V}'\left(\int \xi(z)\xi(z)'\mathcal{K}(z)\, dz\right)^{-1} \int \left(\int \mathcal{K}(z)\,\xi(z)\, dv\right)\left(\int \mathcal{K}(z)\,\xi(z)\, dv\right)' du \left(\int \xi(z)\xi(z)'\mathcal{K}(z)\, dz\right)^{-1} e_{2+d_V}.
\end{aligned}$$

We can decompose the APE's sample variation into five components. The first four components are analogous to those in the earlier ASF decomposition. In particular, the fourth component is

$$\widehat{\beta}^{(k)} \cdot \sqrt{Nb_N^3}\left(\frac{1}{N}\sum_{i=1}^{N} \widehat{h}_2(\underline{x}'\beta_0, V_i; \beta_0)\pi_{it} - \mathbb{E}[h_2(\underline{x}'\beta_0, V; \beta_0)\pi_t]\right)$$

and converges in distribution to a mean-zero Gaussian distribution while dominating the other components.

The fifth component is due to the presence of $\widehat{\beta}^{(k)}$ and is of the same order as $\sqrt{Nb_N^3}(\widehat{\beta}^{(k)} - \beta_0^{(k)}) = O_p(\sqrt{Nb_N^3}a_N) = o_p(1)$ by B6.

The rate of convergence of $\widehat{\text{APE}}_{k,t}(\underline{x})$ in the logit case is $N^{\tilde{\delta}_{\text{APE}}}$, where $\tilde{\delta}_{\text{APE}}$ ranges in the interval $\left(\frac{d_V}{3+2d_V}, \frac{\ell}{3+2\ell}\right)$. When $d_V = 1$ and $\ell = 2$, this range equals $\left(\frac{1}{5}, \frac{2}{7}\right)$. Recall that 2/7 is the standard rate of convergence of for derivatives of univariate kernel estimators when using second-order kernels. Our estimator can approach and achieve this rate whenever $\ell \geq 2$, i.e., the local polynomial contains quadratic terms.

3.4 Implementation Details

Here are a few practical concerns related to the implementation of the ASF and APE estimators. We explore some of these in more detail in our simulations (Section 5) and empirical illustration (Section 6).

Local polynomial regression. First, a common practice in kernel-based methods is the standardization and orthogonalization of the conditioning variables, in our case $Z_t(\widehat{\beta}) = (X_t'\widehat{\beta}, V)$, before the nonparametric estimation step. The standardization leads to more comparable scales across different components of $Z_t(\widehat{\beta})$. The orthogonalization, which can be done via a Cholesky decomposition, is performed on V alone rather than all of $Z_t(\widehat{\beta})$.[7] This orthogonalization makes it sensible to use a product of one-dimensional kernels as our joint kernel, as is done in Assumption B3.

Second, according to Assumption B6, the required polynomial order increases with d_V, the number of continuous index variables. When d_V is 1 or 2, as in our Monte Carlo and empirical illustration, any $\ell \geq 2$ is sufficient. Larger values of ℓ improve the accuracy of the nonparametric approximation but may cause overfitting, especially in small samples. In general, our estimates are not sensitive to ℓ around 2 to 4 in our Monte Carlo simulations and empirical illustration. We use $\ell = 3$ in the Monte Carlo simulations and $\ell = 2$ in the empirical illustration. The smaller ℓ is adopted for the latter because there are discrete index variables dividing the observations into cells, resulting in fewer observations in each cell: See Section 6.1.

Third, we modified the Gaussian kernel as follows to satisfy Assumption B3:

$$K(u) = \begin{cases} \frac{1}{\sqrt{2\pi}} \exp(-u^2/2) & \text{for } |u| \leq 5, \\ \frac{1}{\sqrt{2\pi}} \exp(-5^2/2) \cdot \left(4(6-|u|)^5 - 6(6-|u|)^4 + 3(6-|u|)^3\right) & \text{for } 5 < |u| \leq 6, \\ 0 & \text{for } |u| > 6. \end{cases}$$

This kernel is equivalent to the Gaussian kernel for $|u| \leq 5$ and their results are generally indistinguishable. The truncation at ± 6 ensures the compact support assumption B3.(i) holds. The quintic polynomial for $5 < |u| \leq 6$ guarantees the twice continuous differentiability assumed in B3.(ii).

[7] This is for technical reasons that ensure that $\underline{x}'\widehat{\beta}$ and V enter in the kernel as a product since the latter is averaged out based on its empirical distribution: see the proofs in Appendix B, such as the proof of Lemma B.1.

Bandwidth selection. In practice, one needs to select a bandwidth $b_N = \kappa \cdot N^{-\delta}$. First, we choose δ^* that satisfies our rate conditions in Assumption B6. Then, we find the scaling constant κ^* using leave-one-out cross-validation over a finite grid. In our simulations and empirical illustration, κ^* usually ranges from 1 to 2, and the estimated ASF and APE are generally stable for scaling constants κ ranging in $[\kappa^* - 0.2,\ \kappa^* + 0.2]$.

Trimming set. The compact set $\mathcal{Z}_t$ in the trimming function $\hat{\pi}_{it} = \mathbb{1}((\underline{x}'\widehat{\beta}, V_i) \in \mathcal{Z}_t)$ helps bound $f_{Z_t(\widehat{\beta})}(z)$ away from zero. Candidate criteria could be: a lower bound directly on $\widehat{f}_{Z_t(\widehat{\beta})}(z) = \frac{1}{N}\sum_{j=1}^{N} \mathcal{K}_{b_N}\left(\frac{Z_{jt}(\widehat{\beta}) - z}{b_N}\right)$, an upper bound on the condition number of $S_N(z;\widehat{\beta})$ to ensure the accuracy of matrix inversion and, in a similar sense, a lower bound on the determinant of $S_N(z;\widehat{\beta})$. We combine all three criteria simultaneously to construct the trimming set in our Monte Carlo simulations and empirical illustration.

Asymptotic variance estimation. To conduct inference on the ASF and APE, one could in principle estimate $\sigma_{\text{ASF}_t}(\underline{x}'\beta_0)$ and $\sigma_{\text{APE}_t}(\underline{x}'\beta_0)$ analytically. This can be done by estimating $\text{Var}(Y_t | X_t'\beta_0 = \underline{x}'\beta_0, V_i)$ by $\widehat{h}_1(\underline{x}'\widehat{\beta}, V_i; \widehat{\beta})(1 - \widehat{h}_1(\underline{x}'\widehat{\beta}, V_i; \widehat{\beta}))$, $f_{Z_t}(\underline{x}'\beta_0, V_i)$ by $\frac{1}{N}\sum_{j=1}^{N} \mathcal{K}_{b_N}\left(\frac{Z_{jt}(\widehat{\beta}) - (\underline{x}'\widehat{\beta}, V_i)}{b_N}\right)$, and $f_V(V_i)$ by $\frac{1}{N}\sum_{j=1}^{N} \mathcal{K}^V_{b_N}\left(\frac{V_j - V_i}{b_N}\right)$. For simplicity, we focus on bootstrap-based inference instead. Another benefit of the bootstrap is that it may better capture higher-order terms in the asymptotic expansion of our estimator.

Multiple time periods. Finally, note that the above estimator is for the ASF (or APE), at period t, which may vary with t in the population. Under a stationarity assumption, i.e., $F_{U_t} = F_{U_{t'}}$ for all $t, t' \in \{1, \ldots, T\}$, then $\text{ASF}_t(\underline{x}) = \text{ASF}_{t'}(\underline{x})$ for any pair of time periods. For example, this assumption was made in A2.(i) in the logit case. When the ASF does not depend on t, we can combine ASF estimators from multiple time periods to obtain a more precise estimator. A particularly simple combination consists of averaging the estimated ASFs over time:

$$\overline{\text{ASF}}(\underline{x}) = \frac{1}{T}\sum_{t=1}^{T} \widehat{\text{ASF}}_t(\underline{x}).$$

The asymptotic variance of the equally weighted average can be reduced by selecting weights that depend on t. Weights that minimize the asymptotic variance of the weighted ASF depend on the inverse of an estimate of the joint asymptotic covariance matrix of all T ASF estimators. For simplicity, we propose the simple time average as our rule of thumb.

4 Extensions

4.1 Identification of the Heterogeneity Distribution

In the logit case, we show that full support assumptions lead to the identification of the marginal distribution of the unobserved heterogeneity. In the general case, we also show that further support assumptions point-identify F_C even when U_t is not specified. Therefore, we can recover additional functionals of this unconditional distribution. For example, the entire distribution of C or other nonlinear functionals of the conditional response probability $F_{U_t}(x_t'\beta_0 + C)$ could be of interest, such as its variance or its quantiles: see Chernozhukov, Fernández-Val, and Luo (2018). The following proposition formalizes the logit result.

Proposition 4.1 (Logit case)**.** Suppose A1–A3 hold. Suppose $\text{supp}(X_t'\beta_0, V) = \mathbb{R} \times \mathcal{V}$. Suppose the distribution of $(Y, \mathbf{X})$ is known. Then, $F_{C|V}$ and F_C are point-identified.

Note that this proposition applies only to the logit case. This result relies on a deconvolution, which requires stronger support conditions than required by Theorem 2.1. In particular, the support restriction implies that $X_t'\beta_0$ has full support on the real line. This is the case when at least one regressor, say $X_t^{(k)}$, has full support conditional on $X_t^{(-k)}$, and when $\beta_0^{(k)} \neq 0$, which is related to A2′.(iii). Under this support assumption, it is possible to recover the conditional distribution of $U_t - C|\{V = v\}$ over its entire support, for all $v \in \mathcal{V}$. Given that the distribution of $U_t|V \overset{d}{=} U_t$ is known, a conditional deconvolution argument shows the point identification of the conditional distribution of $C|V$.

Since the distribution of V is identified, the marginal distribution of C is also identified. This implies the identification of all functionals of $(F_C, F_{Y,\mathbf{X}}, \beta_0)$, such as quantiles of the conditional response probability $\Lambda(\underline{x}'\beta_0 + C)$. The estimation of F_C and its functionals is beyond the scope of this paper.

Going further, under stronger support restrictions, we can also show the identification of the distributions of U_t and C when neither are specified. This is described in the following proposition.

Proposition 4.2 (General case)**.** Suppose A1, A2', and A3 hold. For $1 \leq s < t \leq T$, assume that $\text{supp}(X_s'\beta_0, X_t'\beta_0, V) = \mathbb{R}^2 \times \mathcal{V}$. Assume the conditional characteristic functions of U_s, U_t, and C have no zeros. Suppose the distribution of $(Y, \mathbf{X})$ is known. Then, F_C, F_{U_s} and F_{U_t} are point-identified.

This result uses Kotlarski's lemma (Kotlarski, 1967), which has also been used in panel data models by Evdokimov (2009). It requires stronger support assumptions on the joint distribution of indices $X_s'\beta_0$ and $X_t'\beta_0$ to ensure the nonparametric identification of the distribution of $(U_s - C, U_t - C)|V$. The conditions for the application of Kotlarski's lemma, such as the characteristic function restrictions, can be relaxed following Evdokimov and White (2012).

4.2 Identification of the LAR

We now show that the LAR from equation (2.8) is identified under weaker assumptions than required for the identification of the APE.

First, to obtain the identification of the LAR, we replace A3 with the following assumption. For simplicity, we only consider the case where the covariate of interest, $X_t^{(k)}$, is continuously distributed.

Assumption A3* (Local average response)

(i) Let $V \equiv v(\mathbf{X})$, where $v : \mathbb{R}^{T \times d_X} \to \mathbb{R}^{d_V}$ is known. Let $C \mid \mathbf{X} \overset{d}{=} C \mid V$;

(ii) Let $\underline{x} \in \text{supp}(X_t)$. Let $X_t^{(k)}$ be continuously distributed in a neighborhood of $\underline{x}^{(k)}$ given $\{X_t^{(-k)} = \underline{x}^{(-k)}\}$. There exists a neighborhood $\mathcal{N}$ of $\underline{x}'\beta_0$ such that $\text{supp}(V|X_t'\beta_0 = u) \neq \emptyset$ for all $u \in \mathcal{N}$.

Contrary to Assumption A3, $\text{supp}(V|X_t'\beta_0 = u)$ does not need to equal the unconditional support $\mathcal{V}$ for all $u \in \mathcal{N}$ to identify the LAR. We only require it to be nonempty for all u. This assumption allows us to identify $\frac{\partial}{\partial u}\mathbb{E}[Y_t|X_t'\beta_0 = u, V = v]$ from the data. Similar to equation (2.4) in Altonji and Matzkin (2005), the LAR is identified as

$$\text{LAR}_{k,t}(\underline{x}) = \beta_0^{(k)} \cdot \int_{\text{supp}(V|X_t=\underline{x})} \frac{\partial}{\partial u}\mathbb{P}(Y_t = 1|X_t'\beta_0 = u, V = v)|_{u=\underline{x}'\beta_0} \, dF_{V|X_t}(v|\underline{x}).$$

Note that A3* only constrains the support of $V|X_t'\beta_0$, and not of $V|X_t$ as is the case in Altonji and Matzkin (2005). This is a weaker constraint since we condition on a single index $X_t'\beta_0$ rather than all regressors.

Estimation of the LAR could proceed by replacing the components of the above equation with sample analogs. One such estimator is proposed in Altonji and Matzkin (2005) for the nonparametric case. One could also propose an estimator based on a local polynomial regression, which is used in this paper, and which uses the single-index structure of the outcome equation (1.1). We leave the asymptotic analysis of the LAR estimator for future work.

4.3 Extension to a dynamic panel model

We now present an extension of our identification results to a dynamic panel model.[8] Aguirregabiria and Carro (2020) establish the identification of the average marginal effects for a change in Y_{t-1}, and Dobronyi, Gu, and Kim (2021) consider the (partial) identification of functionals of the underlying distribution of individual effects. Both papers utilize the logistic error distribution. Relaxing the logistic assumption, but making the index sufficiency assumption on the unobserved heterogeneity, we obtain identification results

[8] See Arellano and Bonhomme (2017) for a review of nonlinear dynamic panel data models.

for the ASF and APE corresponding to exogenous changes in the components of either X_t or Y_{t-1}. For $t = 1, \ldots, T$, assume that

$$Y_{it} = \mathbb{1}(X_{it}'\beta_0 + \gamma_0 Y_{i\,t-1} + C_i - U_{it} \geq 0). \tag{4.1}$$

Assume the distribution of $(\{Y_{it}\}_{t=0}^T, \{X_{it}\}_{t=1}^T)$ is observed from the data. Also, assume that U_t is stationary: $F_{U_t} = F_{U_s}$ for $1 \leq s < t \leq T$. Let $\theta_0 = (\beta_0', \gamma_0)'$ and $W_t = (X_t', Y_{t-1})'$. Let $\underline{w} \equiv (\underline{x}, \underline{y})$ be a value in $\mathcal{X}_t \times \{0, 1\}$, and consider the following modified index assumption.

Assumption A3† (Index restriction for dynamic panels)

(i) Let $V \equiv v(\mathbf{X})$, where $v : \mathbb{R}^{T \times d_X} \to \mathbb{R}^{d_V}$ is known. Let $C \mid (\mathbf{X}, Y_0) \overset{d}{=} C \mid (V, Y_0)$;

(ii) Let $\underline{x} \in \text{supp}(X_1)$. One of the following two assumptions holds:

(a) let $X_1^{(k)}$ be continuously distributed in a neighborhood of $\underline{x}^{(k)}$ given $\{X_1^{(-k)} = \underline{x}^{(-k)}, Y_0 = \underline{y}\}$. There exists a neighborhood $\mathcal{N}$ of $\underline{w}'\theta_0$ such that $\mathcal{N} \times \text{supp}(V, Y_0) \subseteq \text{supp}(W_1'\theta_0, V, Y_0)$.

(b) let $X_1^{(k)}$ be discretely distributed given $\{X_1^{(-k)} = \underline{x}^{(-k)}, Y_0 = \underline{y}\}$. Let $\widetilde{\underline{x}}_k'\beta_0 = \underline{x}'\beta_0 + (\widetilde{\underline{x}}^{(k)} - \underline{x}^{(k)})\beta_0^{(k)}$. Let $\{\underline{w}'\theta_0, (\widetilde{\underline{x}}_k'\beta_0, \underline{y}\gamma_0)\} \times \text{supp}(V, Y_0) \subseteq \text{supp}(W_1'\theta_0, V, Y_0)$.

In dynamic panel models, this assumption replaces A3. The main difference is that we let the conditional index restriction hold when conditioning on Y_0, the initial value of the outcome variable. We make the assumption that θ_0 is point-identified. This is justified by the work of Chamberlain (1985) and Honoré and Kyriazidou (2000), which show the identification of θ_0 in this model. Identification of θ_0 generally requires the presence of units whose covariate values do not change over time, known as "stayers". As shown in Honoré and Kyriazidou (2000), identification of θ_0 can be achieved even when U_t does not follow a logistic distribution. Under these assumptions, we can point identify the average structural function:

$$\begin{aligned}
\text{ASF}(\underline{w}) &= \int_{\mathcal{C}} \mathbb{P}(Y_1 = 1 | X_t = \underline{x}, Y_0 = \underline{y}, C = c)\, dF_C(c) \\
&= \int_{\mathcal{C}} F_{U_1}(\underline{w}'\theta_0 + c)\, dF_C(c) \\
&= \int_{\mathcal{V} \times \{0,1\}} \int_{\mathcal{C}} F_{U_1}(\underline{w}'\theta_0 + c)\, dF_{C|V,Y_0}(c|v, y_0)\, dF_{V,Y_0}(v, y_0) \\
&= \int_{\mathcal{V} \times \{0,1\}} \mathbb{P}(U_1 \leq \underline{w}'\theta_0 + C | (V, Y_0) = (v, y_0))\, dF_{V,Y_0}(v, y_0) \\
&= \int_{\mathcal{V} \times \{0,1\}} \mathbb{P}(U_1 \leq \underline{w}'\theta_0 + C | (X_1'\beta_0, V, Y_0) = (\underline{w}'\theta_0 - \gamma_0 y_0, v, y_0))\, dF_{V,Y_0}(v, y_0) \\
&= \int_{\mathcal{V} \times \{0,1\}} \mathbb{P}(U_1 \leq \underline{w}'\theta_0 + C | (X_1'\beta_0 + \gamma_0 Y_0, V, Y_0) = (\underline{w}'\theta_0, v, y_0))\, dF_{V,Y_0}(v, y_0)
\end{aligned}$$

$$= \int_{\mathcal{V}\times\{0,1\}} \mathbb{P}(Y_1 = 1 | W_1'\theta_0 = \underline{w}'\theta_0, V = v, Y_0 = y_0)\, dF_{V,Y_0}(v, y_0).$$

The first equality follows from stationarity of U_t.[9] The fifth equality follows from the index restriction $C|(\mathbf{X}, Y_0) \overset{d}{=} C|(V, Y_0)$, which implies $C|(X_1'\beta_0, V, Y_0) \overset{d}{=} C|(V, Y_0)$. The identification of the last line follows from the support assumption A3†.(ii). The above derivation shows that the ASF is identified under the above assumption, and under assumptions that are sufficient for the identification of θ_0. The identification of the APE in the continuous case follows from the identification of the ASF in a neighborhood of $\underline{w}$, as per A3†.(ii).(a). The discrete case follows from support assumption A3†.(ii).(b), guaranteeing the identification of the ASF at values $\underline{w}$ and $(\underline{\widetilde{x}}_k', \underline{y})$.

5 Monte Carlo Simulations

This section conducts two sets of Monte Carlo simulation experiments featuring the logit case and the general case. We focus on APE estimation and defer corresponding ASF results to Appendix 3.2.[10]

We compare the proposed semiparametric estimator with two commonly used parametric alternatives: the RE and CRE. See, for example, Wooldridge (2010). Both assume standard logistic distribution for the error term U_t, which is correctly specified in the logit case but misspecified in the general case. They are characterized by different assumptions on the distribution of individual effects C. For the RE,

$$C \sim \mathcal{N}(\mu_c, \sigma_c^2)$$

and is independent of V. For the CRE,

$$C|V \sim \mathcal{N}(\mu_{c0} + \mu_{c1}'V, \sigma_c^2).$$

Then, the CRE is equivalent to an augmented RE with V being additional regressors.

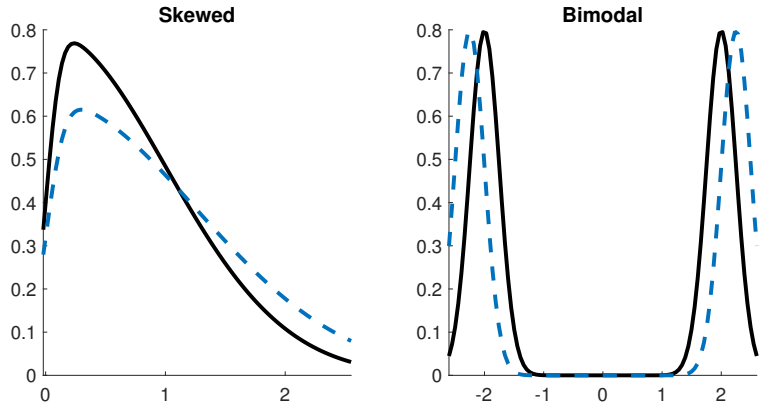
Following the standard practice in the literature, we use the MLE to jointly estimate β_0 and the distribution parameters (μ_c, σ_c^2) or $(\mu_{c0}, \mu_{c1}, \sigma_c^2)$. In the same spirit as the semiparametric estimator in Sections 3.2 and 3.3, we allow the marginal distribution of V to be unrestricted. The conditional expectation of the binary outcome and its derivative are calculated based on the MLE estimates, and the ASF and APE are obtained by averaging out V.

[9] If stationarity fails, the ASF and APE at $t = 1$ remain point-identified.

[10] The ASF could be interesting by itself as a counterfactual probability, and the APE of discrete covariates could be obtained by differencing corresponding ASFs.

Table 1: Monte Carlo Design - Logit Case

Binary outcome:	$Y_{it} = \mathbb{1}(X_{it}\beta_0 + C_i - U_{it} \geq 0)$
Common param.:	$\beta_0 = \begin{cases} 1, & \text{for DGP L.1} \\ 2, & \text{for DGP L.2} \end{cases}$
Covariate:	$X_{it} \sim \mathcal{N}(0,1)$
Index:	$V_i = \frac{1}{T}\sum_{t=1}^{T} X_{it}$
Error term:	$U_{it} \sim \sqrt{3}/\pi \cdot$ standard logistic, so $\text{Var}(U_{it}) = 1$
Sample Size:	$N = 1500$, $T = 10$
# Repetitions:	$N_{sim} = 100$
$f_{C\|V}$:	
DGP L.1, skewed:	$C_i \| V_i \sim (V_i^2 + 1) \cdot \mathcal{SN}(0, 1, 10)$
DGP L.2, bimodal:	$C_i \| V_i \sim \frac{1}{2}\mathcal{N}\left(V_i^2 + 2, \frac{1}{4^2}\right) + \frac{1}{2}\mathcal{N}\left(-V_i^2 - 2, \frac{1}{4^2}\right)$



Notes: $\mathcal{SN}(\xi, \omega, \alpha)$ denotes a skewed normal distribution with location parameter ξ, scale parameter ω, and shape parameter α, and its pdf is given by $f(x) = \frac{2}{\omega}\phi\left(\frac{x-\xi}{\omega}\right)\Phi\left(\alpha\left(\frac{x-\xi}{\omega}\right)\right)$, where $\phi(\cdot)$ and $\Phi(\cdot)$ denote the pdf and cdf of a standard normal distribution. The graphs depict $f_{C|V}$. Black solid and blue dashed lines are conditional on $V = 0$ and $|V| = 0.5$, respectively. Since $f_{C|V}(c|v)$ is symmetric with respect to v, $f_{C|V}(c|0.5) = f_{C|V}(c|-0.5)$.

5.1 Logit Case

The Monte Carlo design is summarized in Table 1. Covariate X_t is drawn from a standard normal distribution, which satisfies the support conditions in Assumptions A2 and A3. Our choices of $N = 1500$ and $T = 10$ are directly comparable with the dataset in our empirical illustration on female labor force participation in which $N = 1461$ and $T = 9$. There are two experiments with different true distributions of $C|V$, where $f_{C|V}$ is skewed in data-generating process (DGP) L.1 and bimodal in DGP L.2.[11]

We evaluate the estimated ASF and APE based on a collection of $\underline{x}$ ranging from -1 to 1, which covers 68% of the distribution of X_t. For the semiparametric method, we first estimate $\widehat{\beta}$ via the CMLE approach,

[11] Many empirical applications feature skewed and/or multimodal distributions of unobserved individual heterogeneity. For example, Liu (2020) estimated the latent productivity distribution of young firms, which exhibits a long right tail and thus concurs with the intuition that good ideas are scarce. Also, Fisher, Jensen, and Tkac (2019) found three modes in the underlying skill distribution of mutual fund management—a primary mode with average ability, a secondary mode with poor performance, and a minor mode with exceptionally high skill.

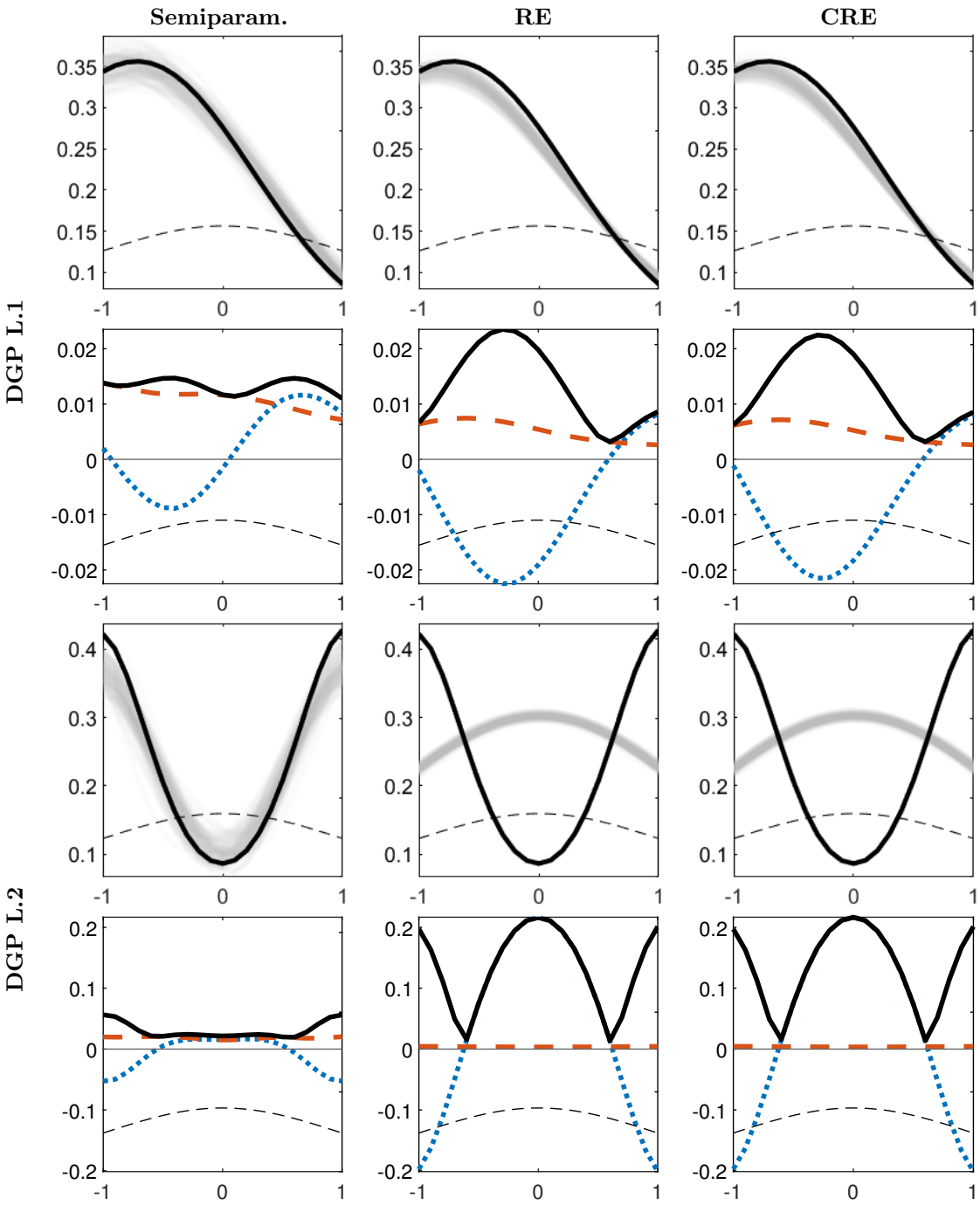
which simultaneously incorporates the information from a logistic error distribution and maintains the flexibility in the distribution of individual effects. We then employ a local cubic regression (i.e., polynomial order $\ell = 3$) to flexibly infer the conditional expectation of Y_t evaluated at $(\underline{x}'\widehat{\beta}, V)$.

Figure 1 compares the estimated APE versus the true APE in the first and third rows, and plots the bias, standard deviation, and root mean square error (RMSE) across 100 Monte Carlo repetitions in the second and fourth rows. Figure 5 in the appendix shows corresponding graphs for the ASF estimates. We see that the semiparametric estimator better captures the peak in the skewed case and the valley in the bimodal case, whereas the RE and CRE completely reverse the valley in the bimodal case due to their parametric restrictions. As expected, the semiparametric estimator generates smaller biases and larger standard deviations than the RE and CRE. The improvement in bias dominates the deterioration in standard deviation around the peak in the skewed setup and most of the time in the bimodal setup. The difference between the RE and CRE is relatively negligible—their parametric assumptions in $f_{C|V}$ seem too restrictive and lead to considerable misspecification biases given current DGPs.

In Table 2, the first three columns summarize the APE estimation performance by computing weighted average performance measures across the collection of evaluation points $\underline{x} \in [-1, 1]$ with weights proportional to $f_{X_t}(\underline{x})$. Similar to what we observed in Figure 1, all three estimators provide similar RMSEs in the skewed case, and the semiparametric estimator yields the smallest RMSE in the bimodal case. The last three columns present the minimum, median, and maximum of the ratios of RMSE($\underline{x}$) to the true APE($\underline{x}$). The minimum, median, and maximum are taken over $\underline{x}$. We see that the ratios range between 2% and 13% in the skewed case and between 5% and 250% in the bimodal case. The differences across estimators are relatively small in the skewed case. In the bimodal case, the RE and CRE have lower *minimal* ratios, which occurs at $\underline{x} = \pm 0.6$ where the grey bands "intersect" with true APE curve; at the same time, the semiparametric estimator reduces the *median* ratio from 47% to 13%, and the *maximal* ratio from 250% to 25%.

We also examine the performance for the common parameter and ASF in Table 6 in the appendix. The structure of the ASF part of the table is the same as Table 2 for the APE. The ratios of RMSE($\underline{x}$) to the true ASF($\underline{x}$) are generally smaller than their APE counterparts, and the semiparametric estimator dominates the RE and CRE. For $\widehat{\beta}$, compared with the RE and CRE, the correctly-specified yet flexible CMLE provides an estimator with a smaller bias, larger standard deviation, and smaller RMSE. Both a more precisely estimated $\widehat{\beta}$ and a flexibly characterized conditional expectation of Y_t contribute to the better performance of the proposed semiparametric method for estimating the ASF and APE.

Figure 1: APE Estimation - Logit Case



Notes: X-axes are potential values $\underline{x}$. In the first and third rows, black solid lines are the true APE, gray bands are collections of lines where each line corresponds to the estimated APE based on one simulation repetition. In the second and fourth rows, black solid / blue dotted / red dashed lines represent the RMSEs / biases / standard deviations of the APE estimates. Thin dashed lines at the bottom of all panels show $f_{X_t}(\underline{x})$.

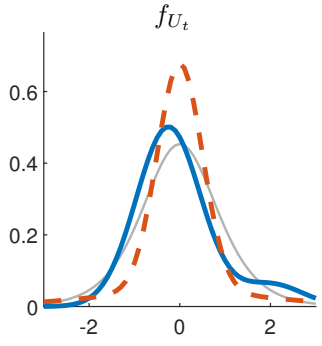
Table 2: APE Estimation - Logit Case

		|Bias|	SD	RMSE	Min	Med.	Max
DGP L.1	Semiparam.	0.011	0.011	**0.013**	3.8%	4.4%	12.8%
	RE	0.013	0.005	0.014	1.9%	5.3%	10.0%
	CRE	0.012	0.005	**0.013**	1.8%	5.2%	10.0%
DGP L.2	Semiparam.	0.025	0.018	**0.029**	7.4%	13.3%	25.0%
	RE	0.137	0.004	0.137	5.2%	46.7%	251.5%
	CRE	0.137	0.004	0.137	5.3%	46.7%	251.9%

Notes: |Bias| indicates the absolute value of the bias. The reported |Bias|, SD, and RMSE are weighted averages across the collection of evaluation points $\underline{x}$, where the weights are proportional to $f_{X_t}(\underline{x})$. Bold entries indicate the best estimator (i.e., with the smallest RMSE) for each DGP. The last three columns are the minimum/median/maximum of RMSE($\underline{x}$)/APE($\underline{x}$) $\times$ 100% over $\underline{x}$.

Table 3: Monte Carlo Design - General Case

Binary outcome:	$Y_{it} = \mathbb{1}\left(X_{it}'\beta_0 + C_i - U_{it} \geq 0\right)$
Common param.:	$\beta_0 = \begin{cases} (1,1)', & \text{for DGP G.1y} \\ (1,2)', & \text{for DGP G.2y} \end{cases}$
Covariate:	$X_{it} \sim \mathcal{N}\left(0_{2\times 1}, I_2\right)$
Index:	$V_i = \frac{1}{T}\sum_{t=1}^{T} X_{it}$
Sample Size:	$N = 1500$, $T = 10$
# Repetitions:	$N_{sim} = 100$
$f_{C\|V}$:	
DGP G.1y, skewed:	$C_i \| V_i \sim \left(\sum_{\iota=1}^{2} V_{\iota,i}^2 + 1\right) \cdot \mathcal{SN}(0,1,10)$
DGP G.2y, bimodal:	$C_i \| V_i \sim \frac{1}{2}\mathcal{N}\left(\sum_{\iota=1}^{2} V_{\iota,i}^2 + 2, 1\right) + \frac{1}{2}\mathcal{N}\left(-\sum_{\iota=1}^{2} V_{\iota,i}^2 - 2, 1\right)$
f_{U_t}, with $\mathbb{E}(U_{it}) = 0$ and $\mathrm{Var}(U_{it}) = 1$:	
DGP G.x1, skewed:	$U_{it} \sim \frac{1}{9}\mathcal{N}\left(2, \frac{1}{2}\right) + \frac{8}{9}\mathcal{N}\left(-\frac{1}{4}, \frac{1}{2}\right)$
DGP G.x2, fat-tailed:	$U_{it} \sim \frac{1}{5}\mathcal{N}\left(0, 4\right) + \frac{4}{5}\mathcal{N}\left(0, \frac{1}{4}\right)$



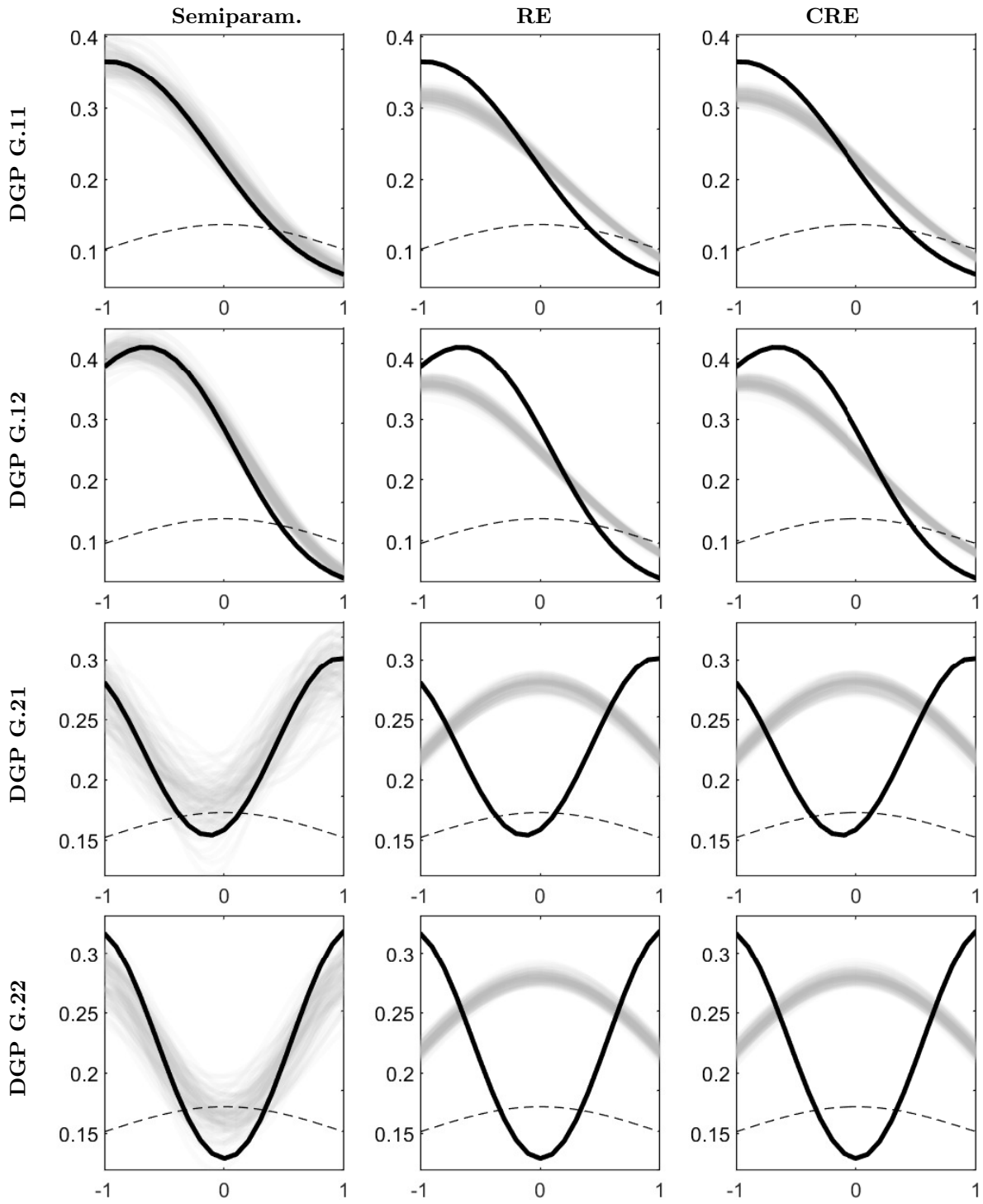
Notes: The blue solid and red dashed lines depict f_{U_t} in DGPs G.x1 (skewed) and G.x2 (fat-tailed), respectively. For reference, the thin gray line plots a rescaled logistic distribution with zero mean and unit variance.

5.2 General Case

The general case accounts for two key features: multidimensional index variables and general error distributions. The exact design is described in Table 3. Now, both X_t and V are 2-by-1 vectors. The distributions of individual effects, $f_{C|V}$, are modified from their counterparts in the logit case where V is a scalar. For the error term, we consider error distributions f_{U_t} that exhibit skewness or fat-tails. We use "DGP G.xy" to indicate the DGP with $f_{C|V}$ being type x and f_{U_t} being type y.

Regarding the evaluation points $\underline{x} = (\underline{x}^{(1)}, \underline{x}^{(2)})'$, we fix $\underline{x}^{(1)}$ at its population mean (i.e., $\underline{x}^{(1)} = 0$) and vary $\underline{x}^{(2)} \in [-1,\ 1]$ as in the logit case. Since $\underline{x}^{(1)}$ and $\underline{x}^{(2)}$ enter symmetrically into the DGPs, similar ASF and APE estimates obtain if we exchange $\underline{x}^{(1)}$ and $\underline{x}^{(2)}$. Given non-logistic error distributions, the

Figure 2: Estimated APE vs True APE - General Case



Notes: X-axes are potential values $\underline{x}^{(2)}$. Black solid lines are the true APE. Gray bands are collections of lines where each line corresponds to the estimated APE based on one simulation repetition. Thin dashed lines at the bottom of all panels show $f_{X_t^{(2)}}\left(\underline{x}^{(2)}\right)$.

Table 4: APE Estimation - General Case

		|Bias|	SD	RMSE	Min	Med.	Max
DGP G.11	Semiparam.	0.013	0.012	**0.016**	4.2%	8.4%	15.7%
	RE	0.028	0.005	0.029	2.7%	13.6%	39.1%
	CRE	0.028	0.005	0.029	2.7%	13.3%	39.1%
DGP G.12	Semiparam.	0.018	0.012	**0.020**	3.3%	6.0%	35.2%
	RE	0.047	0.006	0.047	2.6%	18.3%	107.5%
	CRE	0.046	0.006	0.047	2.5%	18.4%	107.3%
DGP G.21	Semiparam.	0.020	0.015	**0.023**	6.2%	9.4%	20.4%
	RE	0.071	0.004	0.071	3.1%	23.8%	81.5%
	CRE	0.071	0.004	0.071	3.0%	23.7%	81.7%
DGP G.22	Semiparam.	0.022	0.019	**0.026**	7.4%	9.3%	26.6%
	RE	0.086	0.004	0.086	6.3%	31.1%	116.9%
	CRE	0.086	0.004	0.086	6.2%	31.0%	117.2%

Notes: |Bias| indicates the absolute value of the bias. The reported |Bias|, SD, and RMSE are weighted averages across the collection of evaluation points $\underline{x}$, where the weights are proportional to $f_{X_t}(\underline{x})$. Bold entries indicate the best estimator (i.e., with the smallest RMSE) for each DGP. The last three columns are the minimum/median/maximum of RMSE($\underline{x}$)/APE($\underline{x}$) $\times$ 100% over $\underline{x}$.

semiparametric approach estimates β_0 using a smoothed maximum score estimator as in Charlier, Melenberg, and van Soest (1995) and Kyriazidou (1995), and adopts a fourth order cdf kernel to satisfy the bandwidth requirement in Assumption B6. We normalize $|\widehat{\beta}^{(1)}| = 1$ since the identification of β_0 is up to scale.

Figure 2 shows the estimated APEs based on all Monte Carlo repetitions. Table 4 reports the bias, standard deviation, RMSE, and RMSE ratio statistics for the APE estimators. See Figures 6 to 8 and Table 7 in the appendix for supplemental results evaluating the estimation performance for the common parameter, ASF, and APE. For $\widehat{\beta}$, the nonparametric smoothed maximum score estimator produces less biased but noisier estimates, and their RMSEs are larger than those of the RE and CRE. Nevertheless, the semiparametric estimator still better traces the shape of the ASF and APE, and hence provides the most accurate ASF/APE estimates, whose RMSEs are around half of its RE/CRE counterparts.[12]

In terms of the RMSE ratios in the last three columns in Table 4, the message is similar to the logit case: the RMSEs are generally sizeable compared to the true APEs, so the more precise semiparametric estimator is preferable; the RE and CRE exhibit smaller *minimal* ratios at $\underline{x}$ values where the grey bands "intersect" with the true APEs, but the semiparametric estimator gives much smaller *median* and *maximal* ratios.

[12]To take a closer look at how the β_0 estimation affects the APE estimation, we further examine an infeasible semiparametric estimator with known β_0 (see Table 8 in the appendix). Results show that the smoothed maximum score estimates of β_0 slightly increase the absolute value of the bias, the standard deviation, and the RMSE, but the difference is minor—the flexible semiparametric estimator of the APE partially absorbs the effect of the slightly imprecisely estimated β_0.

6 Empirical Illustration

6.1 Background and Specification

In this empirical illustration, we examine women's participation in the labor market using our flexible approach. See the handbook chapter by Killingsworth and Heckman (1986) for an extensive review of the literature on female labor supply. For illustrative purposes, our analysis is based on the static setup in Fernández-Val (2009), where covariates X_t include numbers of children in three age categories, log husband's income, a quadratic function of age, as well as time dummies.[13]

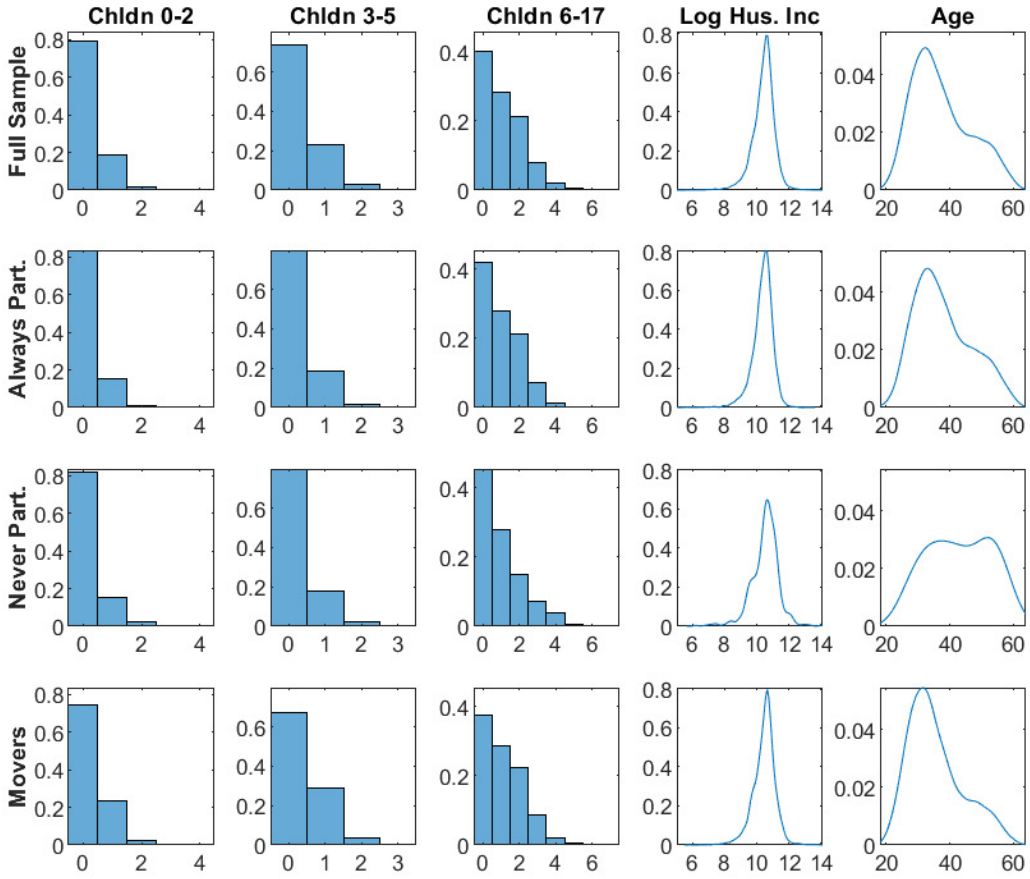
The sample consists of $N = 1461$ married women observed for $T = 9$ years from the PSID between 1980–1988. We use the dataset kindly made available on Iván Fernández-Val's website. Figure 3 plots the distributions of the covariates, and Table 9 in the appendix summarizes the corresponding descriptive statistics. Roughly 45% of the women in the sample always participated in the labor market, less than 10% never participated, and around 45% changed their status during the sample period. Movers tended to be younger and have more children in all children age categories. Never participants were relatively uniformly distributed between ages 30 to 50, whereas the women in other subgroups were generally younger. All subgroups exhibited heavy tails in log husband's income. If this kind of variation in the observables is also present in the unobservables, we suspect the proposed flexible semiparametric estimators might fare better than those requiring distributional assumptions, such as RE and CRE.

The unobserved individual effects C could be interpreted as an individual's willingness to work. A natural choice for the index V would be time-averages of covariates. Women's ages and numbers of children are discrete variables, and we consider a cell-by-cell analysis.[14] These covariates generate over 1000 cells in this sample, thus some cells do not contain sufficient observations to use a semiparametric estimator within them. Therefore, we collapse the discrete index variables as follows. First, we sum over children age categories and average the total number of children under 18 over time. Then, the total number of children is collapsed into a trinary variable depending on whether it is below the 33rd quantile, between the 33rd and 67th quantiles, or above the 67th quantile, and the initial age is collapsed into a binary indicator depending on whether it is above or below the median. This coarsening scheme results in 6 cells, and the number of observations in each cell ranges from 88 to 401. Thus, we have three index variables: a trinary fertility variable, a binary age variable, and a continuously distributed average log husband's income. The number

[13] Charlier, Melenberg, and van Soest (1995) and Chen, Si, Zhang, and Zhou (2017), among others, also considered female labor force participation in their empirical applications. They used similar model specifications, but most of these papers focused on the estimation of common parameters β_0 instead of the ASF or APE.

[14] For a more comprehensive empirical analysis, one could handle discrete index variables using a discrete kernel as suggested in Racine and Li (2004), which would be outside the scope of the current illustration.

Figure 3: Distributions of Observables - Female Labor Force Participation



Notes: The sample consists of $N = 1461$ married women observed for $T = 9$ years from the PSID between 1980–1988. See Fernández-Val (2009) for details.

of continuous index variables is $d_V = 1$. Alternative coarsening schemes are explored in Appendix D.2 and the semiparametric estimator is generally robust to variations in coarsening.

6.2 Results

Table 5 reports the estimated common coefficients on key covariates.[15] We see that women are more inclined to withdraw from the labor force when they have more children, especially younger ones, and when their husbands earn a higher income. Compared to the RE and CRE, the flexible smoothed maximum score

[15] Figure 9 in the appendix also plots the estimated coefficients on time dummies, which capture the time-variation in aggregate participation rates.

estimator provides slightly larger (in magnitude) estimates with larger standard errors.

In our illustration, we focus on the effects of the husband's income, which is linked to the wife's reservation wage. We select evaluation points $\underline{x}$ such that the log husband's income ranges from its 20th to 80th quantiles, and other variables are equal to their means (if continuous) or medians (if discrete). These choices correspond to hypothetical women who are 35 years old, have 0 children between 0 and 2, 0 children between 3 and 5, 1 child between 6 and 17, and whose husband's income ranges from 21K to 55K. All time dummies are set to zeros.

Figure 4 shows estimates of the ASF and APE across $\underline{x}$ together with the 95% bootstrap confidence intervals based on 500 bootstrap samples.[16] For the APE, the semiparametric estimates are closer to zero for lower husband's incomes and more negative for higher ones, while their RE and CRE counterparts are rather flat. Note that for continuous $\underline{x}^{(k)}$,

$$\text{APE}_{k,t}(\underline{x}) = \beta_0^{(k)} \cdot f_{U_t - C}\left(\underline{x}'\beta_0\right) = \beta_0^{(k)} \cdot \int_{\mathcal{C}} f_{U_t}(\underline{x}'\beta_0 + c)\, dF_C(c),$$

where $f_{U_t - C}$ denotes the pdf of $U_t - C$, i.e., a convolution of $-C$ and U_t. Then, the slope of the APE with respect to $\underline{x}'\beta_0$ reflects the shapes of f_C and f_{U_t} as well as the magnitude of $\beta_0^{(k)}$. In this sense, the flatter APE profile with respect to husband's incomes in the RE and CRE could be due to the following three sources:

(i) The RE and CRE feature a Gaussian $f_{C|V}$ and estimate the mean and variance of the Gaussian distribution. The estimated Gaussian variance could be fairly large to accommodate some non-Gaussian heterogeneity in $C|V$, and the resulting $\widehat{f}_{U_t - C}$ could be flatter (around the peak) than the true distribution.

(ii) The RE and CRE assume a logistic f_{U_t}, which could be incorrect.

(iii) The smaller magnitudes of $\widehat{\beta}_0^{(k)}$ for RE and CRE could be due to misspecification of the distributions of U_t and C, and in turn further lead to a milder slope of the APE profile.[17]

In contrast, the semiparametric estimator does not require the parametrization of $f_{C|V}$ or f_{U_t}, thus reducing potential biases due to misspecification.

When using our flexible semiparametric estimator which does not constrain the distributions of C or U_t,

[16] For the ASF, all bootstrap estimates are between 0 and 1, and so is the symmetric percentile-t confidence band based on bootstrap standard deviations. In our model, the condition $\frac{d}{du}\mathbb{P}(Y_t = 1|X_t'\beta_0 = u)|_{u=\underline{x}'\beta_0} \geq 0$ holds, and we impose its empirical counterpart in our estimation procedure. In the bootstrap, this constraint occasionally binds so we censor it at zero and employ the percentile bootstrap to account for the possible non-standard distribution due to censoring.

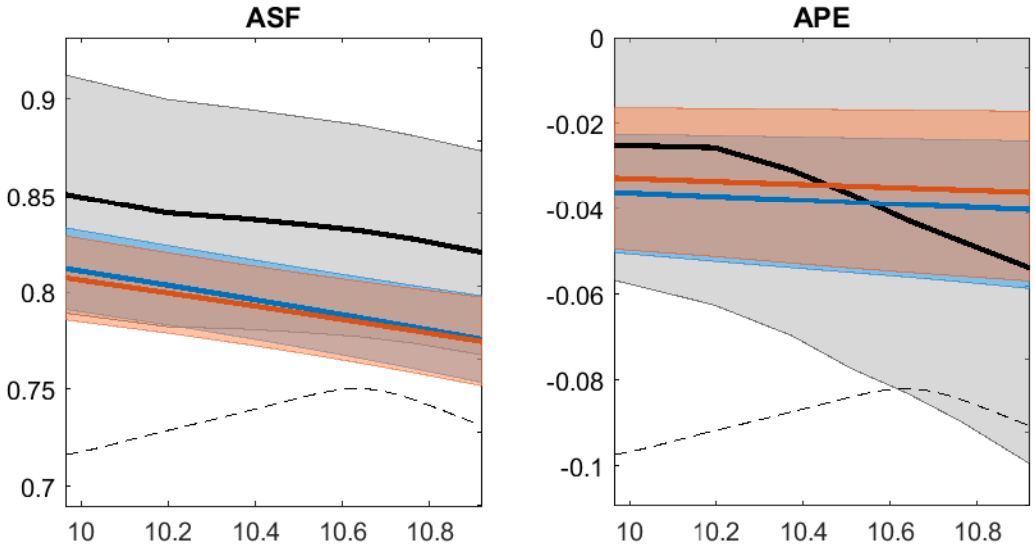
[17] The discrepancy in $\left|\widehat{\beta}_0^{(k)}\right|$ alone cannot explain all difference in the slopes of the APE profiles.

Table 5: Estimated β_0 - Female Labor Force Participation

	Smoothed Max. Score		RE		CRE	
	$\widehat{\beta}$	SD	$\widehat{\beta}$	SD	$\widehat{\beta}$	SD
Children 0–2	*-1*	*0*	*-1*	*0*	*-1*	*0*
Children 3–5	-0.83***	0.18	-0.60***	0.08	-0.60***	0.09
Children 6–17	-0.19	0.17	-0.18***	0.05	-0.17***	0.06
Log Husband's Income	-0.54**	0.22	-0.38***	0.09	-0.35***	0.10
Age/10	3.45*	1.98	2.34***	0.60	2.61***	0.65
$(\text{Age}/10)^2$	-0.51***	0.18	-0.35***	0.08	-0.37***	0.08

Notes: Standard deviations are calculated via the bootstrap. Significance levels are indicated by *: 10%, **: 5%, and ***: 1%. The first row follows from scale normalization $|\widehat{\beta}^{(1)}| = 1$, and we rescale the RE and CRE estimates to allow comparisons across estimators. $\widehat{\beta}^{(1)}$ is negative in all bootstrap samples for all three estimators so, after rescaling, their bootstrap standard deviations all equal to 0. Considering that the support of $\widehat{\beta}^{(1)}$ is ± 1, we do not put asterisks in the first row.

Figure 4: Estimated ASF and APE - Female Labor Force Participation



Notes: X-axes are potential values of log husband's income. Black/blue/orange solid lines represent point estimates of the ASF and APE using the semiparametric/RE/CRE estimators. Bands with corresponding colors indicate the 95% bootstrap confidence intervals. Thin dashed lines at the bottom of both panels show the distribution of log husband's income.

APEs with respect to the husband's income are no longer significant. Highly significant APEs estimated via RE and CRE could partly be an artifact of their parametric restrictions. This is consistent with the empirical observation that married women's labor supply choices became less sensitive to their husbands' income around 1980 when baby boomers started constituting a larger portion of the labor force, and both partners contribute to housework and earnings more equally. Hence fewer married women were at the margin of labor force participation that could be nudged by temporary fluctuations in husbands' income.

For the ASF, all point estimates are downward sloping with respect to the husband's income. The semiparametric estimator yields slightly higher participation probabilities compared to the RE and CRE, though differences across estimators are insignificant at the 5% level.

7 Conclusion

The distributions of the unobserved heterogeneity and of the idiosyncratic errors play a crucial role in the identification of the ASF and APE in binary response models. In this paper we first show the identification of the ASF and APE in semiparametric binary response models with potentially unspecified distributions of the unobserved heterogeneity and of the idiosyncratic errors. To achieve this identification, we assume that units with the same value of the index V have correspondingly similar distributions of their unobserved heterogeneity C. We then develop three-step semiparametric estimators for the ASF and APE, and show their consistency and asymptotic normality. After conducting simulation experiments, we illustrate our semiparametric estimator in a study of determinants of women's labor supply.

We explore in Section 4 several avenues for future research. In particular, we show the identification of the unobserved heterogeneity's distribution under stricter support conditions than those in Section 2.2. This result can be used to show the identification of additional measures of treatment effects beyond the APE. We also provide an identification result that applies to dynamic panel models in Section 4.3. The generalization of our results to a wider class of dynamic models appears promising. We hope to pursue these ideas further in future work.

References

Abrevaya, J., and Y.-C. Hsu (2021): "Partial effects in non-linear panel data models with correlated random effects," *The Econometrics Journal*, Forthcoming.

Aguirregabiria, V., and J. M. Carro (2020): "Identification of Average Marginal Effects in Fixed Effects Dynamic Discrete Choice Models," *Working Paper*.

Altonji, J., and R. Matzkin (2005): "Cross Section and Panel Data Estimators for Nonseparable Models with Endogenous Regressors," *Econometrica*, 73(4), 1053–1102.

Andersen, E. B. (1970): "Asymptotic Properties of Conditional Maximum-Likelihood Estimators," *Journal of the Royal Statistical Society. Series B (Methodological)*, 32(2), 283–301.

Arellano, M., and S. Bonhomme (2017): "Nonlinear Panel Data Methods for Dynamic Heterogeneous Agent Models," *Annual Review of Economics*, 9, 471–496.

Arkhangelsky, D., and G. Imbens (2019): "The Role of the Propensity Score in Fixed Effect Models," *arXiv preprint arXiv:1807.02099v6*.

Bester, C., and C. Hansen (2009): "Identification of Marginal Effects in a Nonparametric Correlated Random Effects Model," *Journal of Business & Economic Statistics*, 27(2), 235–250.

Blundell, R., and J. L. Powell (2003): *Endogeneity in Nonparametric and Semiparametric Regression Models*vol. 2 of *Econometric Society Monographs*, p. 312–357. Cambridge University Press.

Blundell, R. W., and J. L. Powell (2004): "Endogeneity in semiparametric binary response models," *The Review of Economic Studies*, 71(3), 655–679.

Chamberlain, G. (1980): "Analysis of Covariance with Qualitative Data," *Review of Economic Studies*, 47, 225–238.

Chamberlain, G. (1985): "Heterogeneity, omitted variable bias, and duration dependence," in *Longitudinal Analysis of Labor Market Data*, ed. by J. J. Heckman, and B. S. Singer, Econometric Society Monographs, p. 3–38. Cambridge University Press.

Chamberlain, G. (2010): "Binary Response Models for Panel Data: Identification and Information," *Econometrica*, 78(1), 159–168.

Charlier, E., B. Melenberg, and A. H. van Soest (1995): "A smoothed maximum score estimator for the binary choice panel data model with an application to labour force participation," *Statistica Neerlandica*, 49(3), 324–342.

Chen, S., J. Si, H. Zhang, and Y. Zhou (2017): "Root-N Consistent Estimation of a Panel Data Binary Response Model With Unknown Correlated Random Effects," *Journal of Business & Economic Statistics*, 35(4), 559–571.

Chernozhukov, V., I. Fernández-Val, J. Hahn, and W. Newey (2013): "Average and Quantile Effects in Nonseparable Panel Models," *Econometrica*, 81(2), 535–580.

Chernozhukov, V., I. Fernández-Val, and Y. Luo (2018): "The Sorted Effects Method: Discovering Heterogeneous Effects Beyond Their Averages," *Econometrica*, 86(6), 1911–1938.

Cox, D. R. (1958): "The Regression Analysis of Binary Sequences," *Journal of the Royal Statistical Society: Series B (Methodological)*, 20(2), 215–232.

Davezies, L., X. D'Haultfoeuille, and L. Laage (2021): "Identification and Estimation of Average Marginal Effects in Fixed Effect Logit Models," *arXiv preprint arXiv:2105.00879*.

Dobronyi, C., J. Gu, and K. I. Kim (2021): "Identification of Dynamic Panel Logit Models with Fixed Effects," *arXiv preprint arXiv:2104.04590*.

Evdokimov, K. (2009): "Identification and Estimation of a Nonparametric Panel Data Model with Unobserved Heterogeneity," Mimeo.

Evdokimov, K., and H. White (2012): "Some Extensions of a Lemma of Kotlarski," *Econometric Theory*, 28(4), 925–932.

Fan, J., M. Farmen, and I. Gijbels (1998): "Local maximum likelihood estimation and inference," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(3), 591–608.

Fernández-Val, I. (2009): "Fixed effects estimation of structural parameters and marginal effects in panel probit models," *Journal of Econometrics*, 150(1), 71–85.

Fisher, M., M. J. Jensen, and P. A. Tkac (2019): "Bayesian nonparametric learning of how skill is distributed across the mutual fund industry," *FRB Atlanta Working Paper No. 2019-3*.

Frölich, M. (2006): "Non-parametric regression for binary dependent variables," *The Econometrics Journal*, 9(3), 511–540.

Honoré, B. E., and E. Kyriazidou (2000): "Panel Data Discrete Choice Models with Lagged Dependent Variables," *Econometrica*, 68(4), 839–874.

Honoré, B. E., and A. Lewbel (2002): "Semiparametric Binary Choice Panel Data Models without Strictly Exogeneous Regressors," *Econometrica*, 70(5), 2053–2063.

Honoré, B. E., and M. Weidner (2020): "Dynamic Panel Logit Models with Fixed Effects," *arXiv preprint arXiv:2005.05942*.

Horowitz, J. L. (1992): "A smoothed maximum score estimator for the binary response model," *Econometrica*, 60(3), 505–531.

Ichimura, H., and L.-F. Lee (1991): "Semiparametric least squares estimation of multiple index models: Single equation estimation," in *Nonparametric and Semiparametric Methods in Econometrics and Statistics. Proceedings of the Fifth International Symposium in Economic Theory and Econometrics*, ed. by J. P. William A. Barnett, and G. E. Tauchen, pp. 3–49. Cambridge University Press.

Imbens, G. W., and W. K. Newey (2009): "Identification and estimation of triangular simultaneous equations models without additivity," *Econometrica*, 77(5), 1481–1512.

Killingsworth, M. R., and J. J. Heckman (1986): "Female labor supply: A survey," *Handbook of labor economics*, 1, 103–204.

Kim, J., and D. Pollard (1990): "Cube root asymptotics," *The Annals of Statistics*, 18(1), 191–219.

Kitazawa, Y. (2021): "Transformations and moment conditions for dynamic fixed effects logit models," *Journal of Econometrics*.

Kong, E., O. Linton, and Y. Xia (2010): "Uniform Bahadur representation for local polynomial estimates of M-regression and its application to the additive model," *Econometric Theory*, pp. 1529–1564.

Kotlarski, I. (1967): "On characterizing the gamma and the normal distribution," *Pacific Journal of Mathematics*, 20(1), 69–76.

Kyriazidou, E. (1995): "Essays in estimation and testing of econometric models," Ph.D. thesis, Northwestern University.

Laage, L. (2020): "A Correlated Random Coefficient Panel Model with Time-Varying Endogeneity," *arXiv preprint arXiv:2003.09367*.

Lee, M.-J. (1999): "A Root-N Consistent Semiparametric Estimator for Related-Effect Binary Response Panel Data," *Econometrica*, 67(2), 427–433.

Lee, Y.-Y. (2018): "Partial Mean Processes with Generated Regressors: Continuous Treatment Effects and Nonseparable Models," *arXiv preprint arXiv:1811.00157*.

Liu, L. (2020): "Density Forecasts in Panel Data Models: A Semiparametric Bayesian Perspective," *arXiv preprint arXiv:1805.04178*.

Magnac, T. (2000): "Subsidised Training and Youth Employment: Distinguishing Unobserved Heterogeneity from State Dependence in Labour Market Histories," *The Economic Journal*, 110(466), 805–837.

Magnac, T. (2004): "Panel Binary Variables and Sufficiency: Generalizing Conditional Logit," *Econometrica*, 72(6), 1859–1876.

Mammen, E., C. Rothe, and M. Schienle (2012): "Nonparametric regression with nonparametrically generated covariates," *The Annals of Statistics*, 40(2), 1132–1170.

——— (2016): "Semiparametric Estimation With Generated Covariates," *Econometric Theory*, 32(5), 1140–1177.

Manski, C. F. (1987): "Semiparametric analysis of random effects linear models from binary panel data," *Econometrica*, 55(2), 357–362.

Masry, E. (1996): "Multivariate local polynomial regression for time series: uniform strong consistency and rates," *Journal of Time Series Analysis*, 17(6), 571–599.

Maurer, J., R. Klein, and F. Vella (2011): "Subjective Health Assessments and Active Labor Market Participation of Older Men: Evidence From a Semiparametric Binary Choice Model With Nonadditive Correlated Individual-Specific Effects," *The Review of Economics and Statistics*, 93(3), 764–774.

Mundlak, Y. (1978): "On the pooling of time series and cross section data," *Econometrica*, 46(1), 69–85.

Newey, W. K. (1994): "Kernel estimation of partial means and a general variance estimator," *Econometric Theory*, 10(2), 233–253.

Newey, W. K., J. L. Powell, and F. Vella (1999): "Nonparametric estimation of triangular simultaneous equations models," *Econometrica*, 67(3), 565–603.

Nolan, D., and D. Pollard (1987): "U-processes: rates of convergence," *The Annals of Statistics*, 15(2), 780–799.

Pakes, A., and D. Pollard (1989): "Simulation and the asymptotics of optimization estimators," *Econometrica*, 57(5), 1027–1057.

Racine, J., and Q. Li (2004): "Nonparametric estimation of regression functions with both categorical and continuous data," *Journal of Econometrics*, 119(1), 99–130.

Rasch, G. (1960): *Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests.* Nielsen & Lydiche.

——— (1961): "On general laws and the meaning of measurement in psychology," in *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*, vol. 4, pp. 321–333.

Rothe, C., and S. Firpo (2019): "Properties of doubly robust estimators when nuisance functions are estimated nonparametrically," *Econometric Theory*, 35(5), 1048–1087.

Tibshirani, R., and T. Hastie (1987): "Local likelihood estimation," *Journal of the American Statistical Association*, 82(398), 559–567.

Vaart, A. W. v. d. (1998): *Asymptotic Statistics*, Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Wooldridge, J. (2010): *Econometric Analysis of Cross Section and Panel Data*. MIT press.

Appendix

This appendix is organized as follows. Appendices A-B-C contain proofs for results in Sections 2-3-4 respectively. Appendix D contains additional figures and tables that supplement the results in the main text for the Monte Carlo simulation and empirical illustration.

A Proofs for Section 2

Proof of Lemma 2.1. We first consider the logit case. Let $s, s' \in \{1, \dots, T\}$ satisfy A2.(ii). The probability $\mathbb{P}((Y_s, Y_{s'}) = (1,0)|Y_s + Y_{s'} = 1, \mathbf{X} = \mathbf{x})$ is identified from the joint distribution of $(Y, \mathbf{X})$ when $\mathbf{x} \in \mathcal{X}$. By A1 and A2.(i), it equals

$$\begin{aligned}
&\frac{\mathbb{P}((Y_s, Y_{s'}) = (1,0)|\mathbf{X} = \mathbf{x})}{\mathbb{P}(Y_s + Y_{s'} = 1|\mathbf{X} = \mathbf{x})} \\
&= \frac{\mathbb{P}((Y_s, Y_{s'}) = (1,0)|\mathbf{X} = \mathbf{x})}{\mathbb{P}((Y_s, Y_{s'}) = (1,0)|\mathbf{X} = \mathbf{x}) + \mathbb{P}((Y_s, Y_{s'}) = (0,1)|\mathbf{X} = \mathbf{x})} \\
&= \frac{\int_{\mathcal{C}} \Lambda(x_s'\beta_0 + c)(1 - \Lambda(x_{s'}'\beta_0 + c))\, dF_{C|\mathbf{X}}(c|\mathbf{x})}{\int_{\mathcal{C}} \Lambda(x_s'\beta_0 + c)(1 - \Lambda(x_{s'}'\beta_0 + c))\, dF_{C|\mathbf{X}}(c|\mathbf{x}) + \int_{\mathcal{C}} (1 - \Lambda(x_s'\beta_0 + c))\Lambda(x_{s'}'\beta_0 + c)\, dF_{C|\mathbf{X}}(c|\mathbf{x})} \\
&= \frac{e^{x_s'\beta_0}}{e^{x_s'\beta_0} + e^{x_{s'}'\beta_0}} \\
&= \Lambda((x_s - x_{s'})'\beta_0).
\end{aligned}$$

Since Λ is invertible, $(x_s - x_{s'})'\beta_0$ is identified for all $(x_s, x_{s'}) \in \text{supp}(X_s, X_{s'})$. By A2.(ii), β_0 is then point-identified.

The non-logistic case is shown in Lemma 2 of Manski (1987). □

Proof of Theorem 2.1. We first show that the ASF and APE are point identified under A3.(ii).(a), i.e., when the covariate of interest is continuously distributed. By Lemma 2.1, note that β_0 is point identified when U_t is logistic. In the non-logistic case, β_0 is identified up to scale.

Note that the distribution of V, F_V is identified from the data. Then,

$$\begin{aligned}
\text{ASF}_t(\underline{x}) &= \int_{\mathcal{C}} \mathbb{P}(Y_t = 1|X_t = \underline{x}, C = c)\, dF_C(c) \\
&= \int_{\mathcal{C}} \mathbb{P}(U_t \leq \underline{x}'\beta_0 + c)\, dF_C(c) \\
&= \int_{\mathcal{C}} \int_{\mathcal{V}} \mathbb{P}(U_t \leq \underline{x}'\beta_0 + c)\, dF_{C|V}(c|v)\, dF_V(v) \\
&= \int_{\mathcal{V}} \mathbb{P}(U_t \leq \underline{x}'\beta_0 + C|V = v)\, dF_V(v) \\
&= \int_{\mathcal{V}} \mathbb{P}(U_t \leq \underline{x}'\beta_0 + C|X_t'\beta_0 = \underline{x}'\beta_0, V = v)\, dF_V(v) \\
&= \int_{\mathcal{V}} \mathbb{P}(Y_t = 1|X_t'\beta_0 = \underline{x}'\beta_0, V = v)\, dF_V(v).
\end{aligned}$$

The second equality follows from the independence between U_t and $(\mathbf{X}, C)$. The third equality follows from iterated expectations. The fifth equality follows from index assumption A3.(i). The last expression is identified from the data by the identification of $\mathbb{P}(Y_t = 1|X_t'\beta_0 = \underline{x}'\beta_0, V = v)$ for $v \in \mathcal{V}$: see A3.(ii).(a). Hence, $\text{ASF}_t(\underline{x})$ is point identified. Note that this result does not depend on the scale normalization imposed on β_0 since the conditioning set $\{X_t'\beta_0 = \underline{x}'\beta_0\}$ is invariant to the scale of β_0.

By the continuous support of $X_t^{(k)}$ given $X_t^{(-k)} = \underline{x}^{(-k)}$, $\text{ASF}_t(\underline{x} + \tilde{u}e_k)$ is identified for all $\tilde{u}$ in a neighborhood of zero. This is because for small enough $\tilde{u}$, $(\underline{x} + \tilde{u}e_k)'\beta_0 \in \mathcal{N}$ by continuity. By the differentiability of F_{U_t} (see A1.(iii)), $\text{APE}_{k,t}(\underline{x}) = \frac{\partial}{\partial \underline{x}^{(k)}} \text{ASF}_t(\underline{x}) = \lim_{\tilde{u} \to 0} \frac{\text{ASF}_t(\underline{x} + \tilde{u}e_k) - \text{ASF}_t(\underline{x})}{\tilde{u}}$ is also identified.

If $X_t^{(k)}$ is discretely distributed, the result follows from the identification of $\text{ASF}_t(u)$ at $u \in \{\underline{x}, \widetilde{\underline{x}}_k\}$ and by $\text{APE}_{k,t}(\underline{x}, \widetilde{\underline{x}}_k) = \text{ASF}_t(\widetilde{\underline{x}}_k) - \text{ASF}_t(\underline{x})$. □

B Proofs for Section 3

We now present a sequence of lemmas which are used to prove our two main theorems of Section 3: Theorem 3.1 and Theorem 3.2. When applied to matrices, let $\|\cdot\|$ denote the spectral norm.

Lemma B.1 (Convergence of S_N)**.** Suppose B1–B6 hold. Then,

$$\sup_{z\in\mathcal{Z}_t}\left\|S_N(z;\widehat{\beta})-S_N(z;\beta_0)\right\|=o_p\left(\frac{1}{\sqrt{Nb_N}}\right).$$

Proof of Lemma B.1. Select the same generic entry from matrices $S_N(z;\widehat{\beta})$ and $S_N(z;\beta_0)$. These entries can respectively be written as

$$S_N^{\tau,\tau'}(z;\widehat{\beta})\equiv\frac{1}{N}\sum_{j=1}^{N}\left(\frac{Z_{jt}(\widehat{\beta})-z}{b_N}\right)^{\tau}\left(\frac{Z_{jt}(\widehat{\beta})-z}{b_N}\right)^{\tau'}\mathcal{K}_{b_N}\left(\frac{Z_{jt}(\widehat{\beta})-z}{b_N}\right)$$

and

$$S_N^{\tau,\tau'}(z;\beta_0)\equiv\frac{1}{N}\sum_{j=1}^{N}\left(\frac{Z_{jt}(\beta_0)-z}{b_N}\right)^{\tau}\left(\frac{Z_{jt}(\beta_0)-z}{b_N}\right)^{\tau'}\mathcal{K}_{b_N}\left(\frac{Z_{jt}(\beta_0)-z}{b_N}\right),$$

where τ, τ' are vectors of exponents which satisfy $0\le|\tau|,|\tau'|\le\ell$. Let τ_1 and τ_1' denote the first components of τ and τ', and let τ_{-1} and τ_{-1}' denote vectors with all other components of τ and τ'. We can write

$$\begin{aligned}
&S_N^{\tau,\tau'}(z;\widehat{\beta})-S_N^{\tau,\tau'}(z;\beta_0)\\
&=\frac{1}{N}\sum_{j=1}^{N}\left[\left(\frac{X_{jt}'\widehat{\beta}-u}{b_N}\right)^{\tau_1+\tau_1'}\frac{1}{b_N}K\left(\frac{X_{jt}'\widehat{\beta}-u}{b_N}\right)-\left(\frac{X_{jt}'\beta_0-u}{b_N}\right)^{\tau_1+\tau_1'}\frac{1}{b_N}K\left(\frac{X_{jt}'\beta_0-u}{b_N}\right)\right]\\
&\cdot\left(\frac{V_j-v}{b_N}\right)^{\tau_{-1}+\tau_{-1}'}\mathcal{K}_{b_N}^{V}\left(\frac{V_j-v}{b_N}\right)\\
&=\frac{1}{N}\sum_{j=1}^{N}\left[\frac{1}{b_N}\Gamma\left(\frac{X_{jt}'\widehat{\beta}-u}{b_N}\right)-\frac{1}{b_N}\Gamma\left(\frac{X_{jt}'\beta_0-u}{b_N}\right)\right]\left(\frac{V_j-v}{b_N}\right)^{\tau_{-1}+\tau_{-1}'}\mathcal{K}_{b_N}^{V}\left(\frac{V_j-v}{b_N}\right)
\end{aligned}$$

where $\mathcal{K}_{b_N}^{V}(v)=b_N^{-d_V}\cdot\prod_{k=1}^{d_V}K(v_k)$, and $\Gamma(u)\equiv u^{\tau_1+\tau_1'}K(u)$ for generic $u\in\mathbb{R}$.

By B3, Γ is continuously differentiable. A first-order Taylor expansion yields

$$S_N^{\tau,\tau'}(z;\widehat{\beta})-S_N^{\tau,\tau'}(z;\beta_0)=\frac{1}{N}\sum_{j=1}^{N}\frac{1}{b_N^2}\gamma\left(\frac{X_{jt}'\widetilde{\beta}-u}{b_N}\right)\left(\frac{V_j-v}{b_N}\right)^{\tau_{-1}+\tau_{-1}'}\mathcal{K}_{b_N}^{V}\left(\frac{V_j-v}{b_N}\right)X_{jt}'(\widehat{\beta}-\beta_0)$$

where $\widetilde{\beta}$ is such that $X_{jt}'\widetilde{\beta}$ is between $X_{jt}'\widehat{\beta}$ and $X_{jt}'\beta_0$, and where $\gamma(u)\equiv\Gamma'(u)=(\tau_1+\tau_1')u^{\tau_1+\tau_1'-1}K(u)+u^{\tau_1+\tau_1'}K'(u)$.

Since $\mathbb{P}(\widehat{\beta}\in\mathcal{B}_\varepsilon)\to1$ as $N\to\infty$, with probability arbitrarily close to 1, we have that

$$\begin{aligned}
\sup_{z\in\mathcal{Z}_t}\left|S_N^{\tau,\tau'}(z;\widehat{\beta})-S_N^{\tau,\tau'}(z;\beta_0)\right|&\le\frac{1}{b_N^2}\sup_{z\in\mathcal{Z}_t}\left\|\frac{1}{N}\sum_{j=1}^{N}\gamma\left(\frac{X_{jt}'\widetilde{\beta}-u}{b_N}\right)\left(\frac{V_j-v}{b_N}\right)^{\tau_{-1}+\tau_{-1}'}\mathcal{K}_{b_N}^{V}\left(\frac{V_j-v}{b_N}\right)X_{jt}\right.\\
&\quad\left.-\mathbb{E}\left[\gamma\left(\frac{X_t'\widetilde{\beta}-u}{b_N}\right)\left(\frac{V-v}{b_N}\right)^{\tau_{-1}+\tau_{-1}'}\mathcal{K}_{b_N}^{V}\left(\frac{V-v}{b_N}\right)X_t\right]\right\|\|\widehat{\beta}-\beta_0\|\\
&\quad+\sup_{z\in\mathcal{Z}_t}\frac{1}{b_N^2}\left\|\mathbb{E}\left[\mathcal{K}_{b_N}^{V}\left(\frac{V-v}{b_N}\right)\left(\frac{V-v}{b_N}\right)^{\tau_{-1}+\tau_{-1}'}\gamma\left(\frac{X_t'\widetilde{\beta}-u}{b_N}\right)X_t\right]\right\|\|\widehat{\beta}-\beta_0\|\\
&\le\frac{1}{b_N^2}\sup_{z\in\mathcal{Z}_t,\beta\in\mathcal{B}_\varepsilon,b\in(0,\bar{b}]}\left\|\frac{1}{N}\sum_{j=1}^{N}\gamma\left(\frac{X_{jt}'\beta-u}{b}\right)\left(\frac{V_j-v}{b}\right)^{\tau_{-1}+\tau_{-1}'}\mathcal{K}_{b}^{V}\left(\frac{V_j-v}{b}\right)X_{jt}\right.\\
&\quad\left.-\mathbb{E}\left[\gamma\left(\frac{X_t'\beta-u}{b}\right)\left(\frac{V-v}{b}\right)^{\tau_{-1}+\tau_{-1}'}\mathcal{K}_{b}^{V}\left(\frac{V-v}{b}\right)X_t\right]\right\|\|\widehat{\beta}-\beta_0\| \qquad \text{(B.1)}\\
&\quad+\sup_{z\in\mathcal{Z}_t,\beta\in\mathcal{B}_\varepsilon}\frac{1}{b_N^2}\left\|\mathbb{E}\left[\gamma\left(\frac{X_t'\beta-u}{b_N}\right)\left(\frac{V-v}{b_N}\right)^{\tau_{-1}+\tau_{-1}'}\mathcal{K}_{b_N}^{V}\left(\frac{V-v}{b_N}\right)X_t\right]\right\|\|\widehat{\beta}-\beta_0\|, \qquad \text{(B.2)}
\end{aligned}$$

where $\bar{b}>0$. To obtain the stochastic order of term (B.1), define the class of functions

$$\widetilde{\mathcal{F}}=\left\{\gamma\left(\frac{X_t'\beta-u}{b}\right):u\in\mathbb{R},\beta\in\mathcal{B}_\varepsilon,b\in(0,\bar{b}]\right\}.$$

These functions are of the form $\gamma(X_t'c+d)$ where $c=\beta/b$ and $d=-u/b$. Since K has bounded domain and is twice continuously differentiable with bounded derivatives (Assumption B3), the function $\gamma(u)$ is of bounded variation on $\mathbb{R}$. By Nolan and Pollard

(1987) Lemma 22.(ii), the above class of functions is Euclidean. It is also bounded since K is bounded. Similarly, the classes

$$\mathcal{F}_{V_k} = \left\{ \left(\frac{V_k - v_k}{b} \right)^{\tau_{k+1}+\tau'_{k+1}} K\left(\frac{V_k - v_k}{b} \right) : v_k \in \mathbb{R}, b \in (0, \bar{b}] \right\}$$

are Euclidean and bounded for $k = 1, \ldots, d_V$ by the same argument as above. Here τ_{k+1} and τ'_{k+1} denote the $(k+1)$th components of τ and τ'. The product of bounded Euclidean classes is also bounded and Euclidean, hence

$$\mathcal{F}_V = \left\{ \gamma\left(\frac{X_t'\beta - u}{b} \right) \left(\frac{V - v}{b} \right)^{\tau_{-1}+\tau'_{-1}} \mathcal{K}^V\left(\frac{V - v}{b} \right) : z \in \mathcal{Z}_t, \beta \in \mathcal{B}_\varepsilon, b \in (0, \bar{b}] \right\}$$

is bounded and Euclidean. By B5, $\mathbb{E}[\|X_t\|^2] < \infty$. Hence, by Lemma 2.14 (ii) in Pakes and Pollard (1989), the class

$$\mathcal{F} = \left\{ \gamma\left(\frac{X_t'\beta - u}{b} \right) \left(\frac{V - v}{b} \right)^{\tau_{-1}+\tau'_{-1}} \mathcal{K}^V\left(\frac{V - v}{b} \right) X_t : z \in \mathcal{Z}_t, \beta \in \mathcal{B}_\varepsilon, b \in (0, \bar{b}] \right\}$$

is also Euclidean, and hence Donsker. Therefore, by the continuous mapping theorem,

$$\begin{aligned}
&\frac{1}{\sqrt{N} b_N^{2+d_V}} \sup_{z \in \mathcal{Z}_t, \beta \in \mathcal{B}_\varepsilon, b \in (0,\bar{b}]} \left\| \frac{1}{\sqrt{N}} \sum_{j=1}^N \left\{ \gamma\left(\frac{X_{jt}'\beta - u}{b} \right) \left(\frac{V_j - v}{b} \right)^{\tau_{-1}+\tau'_{-1}} \mathcal{K}^V\left(\frac{V_j - v}{b} \right) X_{jt} \right.\right. \\
&\left.\left. - \mathbb{E}\left[\gamma\left(\frac{X_t'\beta - u}{b} \right) \left(\frac{V - v}{b} \right)^{\tau_{-1}+\tau'_{-1}} \mathcal{K}^V\left(\frac{V - v}{b} \right) X_t \right] \right\} \right\| \\
&= \frac{1}{\sqrt{N b_N^{4+2d_V}}} \cdot O_p(1) \\
&= O_p\left((N b_N^{4+2d_V})^{-1/2} \right).
\end{aligned}$$

Thus, term (B.1) can be written as

$$\begin{aligned}
&\frac{1}{b_N^2} \sup_{z \in \mathcal{Z}_t, \beta \in \mathcal{B}_\varepsilon, b \in (0,\bar{b}]} \left\| \frac{1}{N} \sum_{j=1}^N \gamma\left(\frac{X_{jt}'\beta - u}{b} \right) \left(\frac{V_j - v}{b} \right)^{\tau_{-1}+\tau'_{-1}} \mathcal{K}_b^V\left(\frac{V_j - v}{b} \right) X_{jt} \right. \\
&\left. - \mathbb{E}\left[\gamma\left(\frac{X_t'\beta - u}{b} \right) \left(\frac{V - v}{b} \right)^{\tau_{-1}+\tau'_{-1}} \mathcal{K}_b^V\left(\frac{V - v}{b} \right) X_t \right] \right\| \|\widehat{\beta} - \beta_0\| \\
&= O_p\left((N b_N^{4+2d_V})^{-1/2} \right) \cdot O_p(a_N^{-1}) \\
&= o_p\left((N b_N)^{-1/2} \right),
\end{aligned}$$

where the last line follows from $a_N^2 b_N^{3+2d_V} \to \infty$ as $N \to \infty$ (Assumption B6).

To bound term (B.2), we first note that

$$\begin{aligned}
\frac{1}{b_N^2} \left\| \mathbb{E}\left[\gamma\left(\frac{X_t'\beta - u}{b_N} \right) \left(\frac{V - v}{b_N} \right)^{\tau_{-1}+\tau'_{-1}} \mathcal{K}_{b_N}^V\left(\frac{V - v}{b_N} \right) X_t \right] \right\| &= \left\| \mathbb{E}\left[\frac{\partial}{\partial \beta} \left(\frac{Z_t(\beta) - z}{b_N} \right)^{\tau+\tau'} \mathcal{K}_{b_N}\left(\frac{Z_t(\beta) - z}{b_N} \right) \right] \right\| \\
&= \left\| \int \frac{\partial}{\partial \beta} \left(\frac{\widetilde{z} - z}{b_N} \right)^{\tau+\tau'} \mathcal{K}_{b_N}\left(\frac{\widetilde{z} - z}{b_N} \right) f_{Z_t(\beta)}(\widetilde{z})\, d\widetilde{z} \right\| \\
&= \left\| \int \frac{\partial}{\partial \beta} a^{\tau+\tau'} \mathcal{K}(a)\, f_{Z_t(\beta)}(z + a b_N)\, da \right\|.
\end{aligned}$$

The last equality follows from the change of variables $\widetilde{z} = z + a b_N$. We then have that

$$\begin{aligned}
\sup_{z \in \mathcal{Z}_t, \beta \in \mathcal{B}_\varepsilon} \left\| \int \frac{\partial}{\partial \beta} a^{\tau+\tau'} \mathcal{K}(a)\, f_{Z_t(\beta)}(z + a b_N) da \right\| &\leq \sup_{z \in \mathcal{Z}_t, \beta \in \mathcal{B}_\varepsilon} \left\| \frac{\partial}{\partial \beta} f_{Z_t(\beta)}(z) \right\| \left| \int a^{\tau+\tau'} \mathcal{K}(a)\, da \right| \\
&< \infty.
\end{aligned}$$

To see that the last inequality holds, recall Assumption B4.(ii), and that $\mathcal{K}$ is a bounded function with compact support, hence $a^{\tau+\tau'}\mathcal{K}(a)$ is bounded with compact support. Therefore, term (B.2) is of order $O(1) \cdot \|\widehat{\beta} - \beta_0\| = O_p(a_N^{-1}) = o_p\left((N b_N)^{-1/2} \right)$ since, by B6, $N b_N a_N^{-2} \to 0$ as $N \to \infty$.

Combining the rates of convergence of terms (B.1) and (B.2), we obtain

$$\sup_{z\in\mathcal{Z}_t}\left|S_N^{\tau,\tau'}(z;\widehat{\beta})-S_N^{\tau,\tau'}(z;\beta_0)\right|=o_p\left(\frac{1}{\sqrt{Nb_N}}\right)$$

Since this rate of convergence applies uniformly in $z\in\mathcal{Z}_t$ to a generic element of $S_N^{\tau,\tau'}(z;\widehat{\beta})-S_N^{\tau,\tau'}(z;\beta_0)$, it also applies uniformly in $z\in\mathcal{Z}_t$ to the matrix norm of $S_N(z;\widehat{\beta})-S_N(z;\beta_0)$, which concludes the proof. □

Define

$$S(z;\beta_0)=\int\xi(a)\xi(a)'\mathcal{K}(a)\,da\cdot f_{Z_t(\beta_0)}(z).$$

Lemma B.2 (Convergence of S_N to S)**.** Suppose B1–B6 hold. Then,

$$\sup_{z\in\mathcal{Z}_t}\|S_N(z;\beta_0)-S(z;\beta_0)\|=O_p\left(\left(\frac{\log(N)}{Nb_N^{1+d_V}}\right)^{1/2}\right)+O(b_N).$$

Proof of Lemma B.2. This is Corollary 1.(ii) in Masry (1996) with $\theta=1$ (in his notation), therefore we verify its assumptions. His condition 1(b) holds by B4.(iv). His conditions 2 and 3 hold by B3 and B4.(iii). Finally, the rate conditions of Theorem 2 in Masry (1996) hold by B6. Therefore, all assumptions of his corollary holds and the above result holds. □

Lemma B.3 (Convergence of T_N)**.** Suppose B1–B6 hold. Then,

$$\sup_{z\in\mathcal{Z}_t}\left\|T_N(z;\widehat{\beta})-T_N(z;\beta_0)\right\|=o_p\left(\frac{1}{\sqrt{Nb_N}}\right).$$

Proof of Lemma B.3. Select the same generic component from $T_N(z;\widehat{\beta})$ and $T_N(z;\beta_0)$. These components can respectively be written as

$$\begin{aligned}T_N^{\tau}(z;\widehat{\beta})&\equiv\frac{1}{N}\sum_{j=1}^{N}\left(\frac{Z_{jt}(\widehat{\beta})-z}{b_N}\right)^{\tau}Y_{jt}\mathcal{K}_{b_N}\left(\frac{Z_{jt}(\widehat{\beta})-z}{b_N}\right)\\T_N^{\tau}(z;\beta_0)&\equiv\frac{1}{N}\sum_{j=1}^{N}\left(\frac{Z_{jt}(\beta_0)-z}{b_N}\right)^{\tau}Y_{jt}\mathcal{K}_{b_N}\left(\frac{Z_{jt}(\beta_0)-z}{b_N}\right),\end{aligned}$$

where τ is a vector of exponents which satisfies $0\le|\tau|\le\ell$. Again let τ_1 denote the first component of τ and let τ_{-1} denote all other components of τ. Let $\Gamma(u)\equiv u^{\tau_1}K(u)$ and $\gamma(u)\equiv\Gamma'(u)=\tau_1u^{\tau_1-1}K(u)+u^{\tau_1}K'(u)$. As in the proof of Lemma B.1, we write

$$\begin{aligned}&T_N^{\tau}(z;\widehat{\beta})-T_N^{\tau}(z;\beta_0)\\&=\frac{1}{N}\sum_{j=1}^{N}Y_{jt}\left[\frac{1}{b_N}\Gamma\left(\frac{X_{jt}'\widehat{\beta}-u}{b_N}\right)-\frac{1}{b_N}\Gamma\left(\frac{X_{jt}'\beta_0-u}{b_N}\right)\right]\left(\frac{V_j-v}{b_N}\right)^{\tau_{-1}}\mathcal{K}_{b_N}^{V}\left(\frac{V_j-v}{b_N}\right)\\&=\frac{1}{N}\sum_{j=1}^{N}Y_{jt}\frac{1}{b_N^2}\gamma\left(\frac{X_{jt}'\widetilde{\beta}-u}{b_N}\right)\left(\frac{V_j-v}{b_N}\right)^{\tau_{-1}}\mathcal{K}_{b_N}^{V}\left(\frac{V_j-v}{b_N}\right)X_{jt}'(\widehat{\beta}-\beta_0)\end{aligned}$$

By the same arguments as in the proof of Lemma B.1, and by Y_{jt} being bounded, we can show that

$$\begin{aligned}\sup_{z\in\mathcal{Z}_t}\left|T_N^{\tau}(z;\widehat{\beta})-T_N^{\tau}(z;\beta_0)\right|&\le\frac{1}{b_N^2}\sup_{z\in\mathcal{Z}_t,\beta\in\mathcal{B}_\varepsilon,b\in(0,\bar{b}]}\left\|\frac{1}{N}\sum_{j=1}^{N}Y_{jt}X_{jt}\gamma\left(\frac{X_{jt}'\beta-u}{b}\right)\left(\frac{V_j-v}{b}\right)^{\tau_{-1}}\mathcal{K}_b^V\left(\frac{V_j-v}{b}\right)\right.\\&\quad\left.-\mathbb{E}\left[Y_tX_t\gamma\left(\frac{X_t'\beta-u}{b}\right)\left(\frac{V-v}{b}\right)^{\tau_{-1}}\mathcal{K}_b^V\left(\frac{V-v}{b}\right)\right]\right\|\|\widehat{\beta}-\beta_0\|\\&\quad+\sup_{z\in\mathcal{Z}_t,\beta\in\mathcal{B}_\varepsilon}\frac{1}{b_N^2}\left\|\mathbb{E}\left[Y_tX_t\gamma\left(\frac{X_t'\beta-u}{b_N}\right)\left(\frac{V-v}{b_N}\right)^{\tau_{-1}}\mathcal{K}_{b_N}^V\left(\frac{V-v}{b_N}\right)\right]\right\|\|\widehat{\beta}-\beta_0\|\\&=O_p\left(\frac{1}{\sqrt{Nb_N^{4+2d_V}}}\right)\cdot O_p(a_N^{-1})+O(1)\cdot O_p(a_N^{-1})\\&=o_p\left(\frac{1}{\sqrt{Nb_N}}\right)\end{aligned}$$

holds with probability arbitrarily close to 1 as $N\to\infty$ since $\mathbb{P}(\widehat{\beta}\in\mathcal{B}_\varepsilon)\to1$. The last equality follows from B6.

Since this rate of convergence applies uniformly in $z \in \mathcal{Z}_t$ to generic components of the vector $T_N(z;\widehat{\beta}) - T_N(z;\beta_0)$, it applies to its vector norm uniformly in $z \in \mathcal{Z}_t$ as well, which concludes the proof. □

Let

$$T(z;\beta_0) = \int \xi(a)\mathcal{K}(a)\,da \cdot \mathbb{E}[Y_t|Z_t(\beta_0) = z] f_{Z_t(\beta_0)}(z).$$

Also, recall that $Z_t \equiv Z_t(\beta_0)$.

Lemma B.4 (Convergence of T_N to T)**.** Suppose B1–B6 hold. Then,

$$\sup_{z\in\mathcal{Z}_t} \|T_N(z;\beta_0) - T(z;\beta_0)\| = O_p\left(\left(\frac{\log(N)}{Nb_N^{1+d_V}}\right)^{1/2}\right) + O(b_N).$$

Proof of Lemma B.4. By the triangle inequality,

$$\sup_{z\in\mathcal{Z}_t} \|T_N(z;\beta_0) - T(z;\beta_0)\| \leq \sup_{z\in\mathcal{Z}_t} \|T_N(z;\beta_0) - \mathbb{E}[T_N(z;\beta_0)]\| + \sup_{z\in\mathcal{Z}_t} \|\mathbb{E}[T_N(z;\beta_0)] - T(z;\beta_0)\|.$$

Generic components of $T_N(z;\beta_0) - \mathbb{E}[T_N(z;\beta_0)]$ can be written as

$$\sup_{z\in\mathcal{Z}_t} \left| \frac{1}{N}\sum_{j=1}^N \left(\frac{Z_{jt}-z}{b_N}\right)^\tau Y_{jt}\mathcal{K}_{b_N}\left(\frac{Z_{jt}-z}{b_N}\right) - \mathbb{E}\left[\left(\frac{Z_t-z}{b_N}\right)^\tau Y_t\mathcal{K}_{b_N}\left(\frac{Z_t-z}{b_N}\right)\right]\right|.$$

By an argument similar to that used in Corollary 1.(ii) in Masry (1996) or in Lemma B.ii.(2) in Rothe and Firpo (2019), this term is of order $O_p\left(\left(\frac{\log(N)}{Nb_N^{1+d_V}}\right)^{1/2}\right)$.

Next, note that generic elements of $\mathbb{E}[T_N(z;\beta_0)]$ are of the form

$$\begin{aligned}
\mathbb{E}\left[\left(\frac{Z_t-z}{b_N}\right)^\tau Y_t\mathcal{K}_{b_N}\left(\frac{Z_t-z}{b_N}\right)\right] &= \int \left(\frac{\tilde{z}-z}{b_N}\right)^\tau \mathbb{E}[Y_t|Z_t=\tilde{z}]\mathcal{K}_{b_N}\left(\frac{\tilde{z}-z}{b_N}\right) f_{Z_t}(\tilde{z})\,d\tilde{z} \\
&= \int a^\tau \mathcal{K}(a)\mathbb{E}[Y_t|Z_t = z+ab_N] f_{Z_t}(z+ab_N)\,da \\
&\leq \mathbb{E}[Y_t|Z_t=z]f_{Z_t}(z)\int a^\tau\mathcal{K}(a)\,da + b_N \sup_{z\in\mathcal{Z}_t}\left\|\frac{\partial}{\partial z}\left(\mathbb{E}[Y_t|Z_t=z]f_{Z_t}(z)\right)\right\| \cdot \left\|\int a^\tau\mathcal{K}(a)\cdot a\,da\right\|.
\end{aligned}$$

The second equality follows from a change in variables. Note that $\mathbb{E}[Y_t|Z_t=z]f_{Z_t}(z)\int a^\tau\mathcal{K}(a)\,da$ is the corresponding element of $T(z;\beta_0)$. Therefore,

$$\begin{aligned}
&\sup_{z\in\mathcal{Z}_t}\left|\int a^\tau\mathcal{K}(a)\mathbb{E}[Y_t|Z_t=z+ab_N]f_{Z_t}(z+ab_N)\,da - \mathbb{E}[Y_t|Z_t=z]f_{Z_t}(z)\int a^\tau\mathcal{K}(a)\,da\right| \\
&\leq b_N \sup_{z\in\mathcal{Z}_t}\left\|\frac{\partial}{\partial z}\left(\mathbb{E}[Y_t|Z_t=z]f_{Z_t}(z)\right)\right\| \cdot \left\|\int a^\tau\mathcal{K}(a)\cdot a\,da\right\|.
\end{aligned}$$

By B3, $\left\|\int a^\tau\mathcal{K}(a)\cdot a\,da\right\| < \infty$. By B4.(iii), we have that $\sup_{z\in\mathcal{Z}_t}\left\|\frac{\partial}{\partial z}\left(\mathbb{E}[Y_t|Z_t=z]f_{Z_t}(z)\right)\right\| < \infty$. Therefore,

$$\sup_{z\in\mathcal{Z}_t}\|\mathbb{E}[T_N(z;\beta_0)] - T(z;\beta_0)\| = O(b_N)$$

and

$$\sup_{z\in\mathcal{Z}_t}\|T_N(z;\beta_0) - T(z;\beta_0)\| = O_p\left(\left(\frac{\log(N)}{Nb_N^{1+d_V}}\right)^{1/2}\right) + O(b_N).$$

□

Lemma B.5 (Convergence of S_N part 2)**.** Suppose B1–B6 hold. Then,

$$\sup_{z\in\mathcal{Z}_t}\left\|\frac{\partial}{\partial u}S_N(z;\beta_0)\right\| = o_p\left(\frac{a_N}{\sqrt{Nb_N}}\right).$$

Proof of Lemma B.5. As in the proof of Lemma B.1, consider a generic entry of $S_N(z;\beta_0)$, which we write as

$$S_N^{\tau,\tau'}(z;\beta_0) = \frac{1}{N}\sum_{j=1}^N \left(\frac{Z_{jt}-z}{b_N}\right)^{\tau+\tau'}\mathcal{K}_{b_N}\left(\frac{Z_{jt}-z}{b_N}\right).$$

Its derivative with respect to u, the first element of z, is

$$\frac{\partial}{\partial u} S_N^{\tau,\tau'}(z;\beta_0) = \frac{-1}{b_N^{2+d_V}} \frac{1}{N} \sum_{j=1}^{N} \gamma\left(\frac{X_{jt}'\beta_0 - u}{b_N}\right) \left(\frac{V_j - v}{b_N}\right)^{\tau_{-1}+\tau_{-1}'} \mathcal{K}^V\left(\frac{V_j - v}{b_N}\right)$$

where $\gamma(u) = (\tau_1 + \tau_1')u^{\tau_1+\tau_1'-1}K(u) + u^{\tau_1+\tau_1'}K'(u)$.

Therefore, we have that

$$\begin{aligned}
\sup_{z\in\mathcal{Z}_t} \left|\frac{\partial}{\partial u} S_N^{\tau,\tau'}(z;\beta_0)\right| &= \sup_{z\in\mathcal{Z}_t} \left|\frac{-1}{b_N^{2+d_V}} \frac{1}{N} \sum_{j=1}^{N} \gamma\left(\frac{X_{jt}'\beta_0 - u}{b_N}\right) \left(\frac{V_j - v}{b_N}\right)^{\tau_{-1}+\tau_{-1}'} \mathcal{K}^V\left(\frac{V_j - v}{b_N}\right)\right| \\
&\leq \sup_{z\in\mathcal{Z}_t, b\in(0,\bar{b}]} \frac{1}{\sqrt{N} b_N^{2+d_V}} \left| \frac{1}{\sqrt{N}} \sum_{j=1}^{N} \left\{ \gamma\left(\frac{X_{jt}'\beta_0 - u}{b}\right) \left(\frac{V_j - v}{b}\right)^{\tau_{-1}+\tau_{-1}'} \mathcal{K}^V\left(\frac{V_j - v}{b}\right) \right.\right. \\
&\quad \left.\left. - \mathbb{E}\left[\gamma\left(\frac{X_t'\beta_0 - u}{b}\right) \left(\frac{V - v}{b}\right)^{\tau_{-1}+\tau_{-1}'} \mathcal{K}^V\left(\frac{V - v}{b}\right)\right]\right\}\right| \qquad \text{(B.3)} \\
&\quad + \sup_{z\in\mathcal{Z}_t} \frac{1}{b_N^2} \left|\mathbb{E}\left[\gamma\left(\frac{X_t'\beta_0 - u}{b}\right) \left(\frac{V - v}{b}\right)^{\tau_{-1}+\tau_{-1}'} \mathcal{K}_{b_N}^V\left(\frac{V - v}{b}\right)\right]\right|. \qquad \text{(B.4)}
\end{aligned}$$

The class

$$\left\{\gamma\left(\frac{X_t'\beta_0 - u}{b}\right) \left(\frac{V - v}{b}\right)^{\tau_{-1}+\tau_{-1}'} \mathcal{K}^V\left(\frac{V - v}{b}\right) : z \in \mathcal{Z}_t, b \in (0,\bar{b}]\right\}$$

is a subset of $\mathcal{F}_V$ which is Euclidean, therefore it is also Euclidean and hence Donsker. We therefore have that term (B.3) is of order $O_p\left(\frac{1}{\sqrt{Nb_N^{4+2d_V}}}\right)$.

We can bound term (B.4) as follows,

$$\begin{aligned}
&\sup_{z\in\mathcal{Z}_t} \frac{1}{b_N^2} \left|\mathbb{E}\left[\gamma\left(\frac{X_t'\beta_0 - u}{b_N}\right) \left(\frac{V - v}{b_N}\right)^{\tau_{-1}+\tau_{-1}'} \mathcal{K}_{b_N}^V\left(\frac{V - v}{b_N}\right)\right]\right| \\
&= \sup_{z\in\mathcal{Z}_t} \left|\mathbb{E}\left[\frac{\partial}{\partial u}\left(\frac{Z_t - z}{b_N}\right)^{\tau+\tau'} \mathcal{K}_{b_N}\left(\frac{Z_t - z}{b_N}\right)\right]\right| \\
&= \sup_{z\in\mathcal{Z}_t} \left|\int \frac{\partial}{\partial u}\left(\frac{\tilde{z} - z}{b_N}\right)^{\tau+\tau'} \mathcal{K}_{b_N}\left(\frac{\tilde{z} - z}{b_N}\right) f_{Z_t(\beta_0)}(\tilde{z})\, d\tilde{z}\right| \\
&= \sup_{z\in\mathcal{Z}_t} \left|\int \frac{\partial}{\partial u} a^{\tau+\tau'} \mathcal{K}(a)\, f_{Z_t}(z + ab_N)\, da\right| \\
&\leq \sup_{z\in\mathcal{Z}_t} \left|\frac{\partial}{\partial u} f_{Z_t}(z)\right| \left|\int a^{\tau+\tau'} \mathcal{K}(a)\, da\right| \\
&= O(1).
\end{aligned}$$

The third equality follows from the change of variables $\tilde{z} = z + ab_N$. The final line follow from B3 and B4.(iii).

Therefore,

$$\begin{aligned}
\sup_{z\in\mathcal{Z}_t} \left|\frac{\partial}{\partial z_1} S_N^{\tau,\tau'}(z;\beta_0)\right| &= O_p\left(\frac{1}{\sqrt{Nb_N^{4+2d_V}}}\right) + O(1) \\
&= o_p\left(\frac{a_N}{\sqrt{Nb_N}}\right)
\end{aligned}$$

since, as $N \to \infty$, $\frac{1}{\sqrt{Nb_N^{4+2d_V}}} \cdot \frac{\sqrt{Nb_N}}{a_N} = O(N^{\epsilon-\delta(3/2+d_V)}) = o(1)$ by B6, and since $\frac{\sqrt{Nb_N}}{a_N} \cdot O(1) = O(N^{1/2-\epsilon-\delta/2}) = o(1)$, also by B6. Since this holds for a generic entry of the matrix $\frac{\partial}{\partial u} S_N(z;\beta_0)$, it holds for its matrix norm as well, which concludes this lemma. □

Lemma B.6 (Convergence of T_N part 2)**.** Suppose B1–B6 hold. Then,

$$\sup_{z\in\mathcal{Z}_t}\left\|\frac{\partial}{\partial u}T_N(z;\beta_0)\right\| = o_p\left(\frac{a_N}{\sqrt{Nb_N}}\right).$$

Proof of Lemma B.6. As in the proof of Lemma B.3, consider a generic component the vector of $T_N(z;\beta_0)$. Write this element as

$$T_N^\tau(z;\beta_0) = \frac{1}{N}\sum_{j=1}^N\left(\frac{Z_{jt}-z}{b_N}\right)^\tau Y_{jt}\mathcal{K}_{b_N}\left(\frac{Z_{jt}-z}{b_N}\right).$$

Its derivative with respect to u is

$$\frac{\partial}{\partial u}T_N^\tau(z;\beta_0) = \frac{-1}{b_N^{2+d_V}}\frac{1}{N}\sum_{j=1}^N Y_{jt}\gamma\left(\frac{X_{jt}'\beta-u}{b}\right)\left(\frac{V_j-v}{b}\right)^{\tau-1}\mathcal{K}^V\left(\frac{V_j-v}{b}\right).$$

where $\gamma(u)=\tau_1 u^{\tau_1-1}K(u)+u^{\tau_1}K'(u)$. The rest of the proof follows directly from the arguments used in the proofs of Lemma B.3 and B.5. □

Lemma B.7 (Convergence of indicators)**.** Suppose B1–B6 hold. Suppose $\widetilde{\beta}\xrightarrow{p}\beta_0$. Let $\pi_{it}(\beta)\equiv\mathbb{1}((\underline{x}'\beta,V_i)\in\mathcal{Z}_t)$. Then,

$$\mathbb{P}\left(\sup_{i=1,\ldots,N}\left|\pi_{it}(\widetilde{\beta})-\pi_{it}(\beta_0)\right|=0\right)\to 1$$

as $N\to\infty$.

Proof of Lemma B.7. We note that

$$\begin{aligned}\sup_{i=1,\ldots,N}|\pi_{it}(\widetilde{\beta})-\pi_{it}(\beta_0)| &= \sup_{i=1,\ldots,N}\left(\mathbb{1}((\underline{x}'\widetilde{\beta},V_i)\in\mathcal{Z}_t,(\underline{x}'\beta_0,V_i)\notin\mathcal{Z}_t)+\mathbb{1}((\underline{x}'\widetilde{\beta},V_i)\notin\mathcal{Z}_t,(\underline{x}'\beta_0,V_i)\in\mathcal{Z}_t)\right)\\ &\le \sup_{i=1,\ldots,N}\left(\mathbb{1}(\underline{x}'\widetilde{\beta}\in\mathcal{Z}_{1t},\underline{x}'\beta_0\notin\mathcal{Z}_{1t})+\mathbb{1}(\underline{x}'\widetilde{\beta}\notin\mathcal{Z}_{1t},\underline{x}'\beta_0\in\mathcal{Z}_{1t})\right)\\ &= \mathbb{1}(\underline{x}'\widetilde{\beta}\in\mathcal{Z}_{1t},\underline{x}'\beta_0\notin\mathcal{Z}_{1t})+\mathbb{1}(\underline{x}'\widetilde{\beta}\notin\mathcal{Z}_{1t},\underline{x}'\beta_0\in\mathcal{Z}_{1t}),\end{aligned}$$

where $\mathcal{Z}_{1t}=\{z_1=e_1'z : z\in\mathcal{Z}_t\}$. By B4.(v), $\underline{x}'\beta_0\in\mathcal{Z}_{1t}$, and therefore $\mathbb{1}(\underline{x}'\widetilde{\beta}\in\mathcal{Z}_{1t},\underline{x}'\beta_0\notin\mathcal{Z}_{1t})=0$, and $\mathbb{1}(\underline{x}'\widetilde{\beta}\notin\mathcal{Z}_{1t},\underline{x}'\beta_0\in\mathcal{Z}_{1t})=\mathbb{1}(\underline{x}'\widetilde{\beta}\notin\mathcal{Z}_{1t})$.

By assumption, $\widetilde{\beta}$ converges in probability to β_0. By Theorem 18.9.(v) in Vaart (1998), $\mathbb{P}(\underline{x}'\widetilde{\beta}\in\mathcal{Z}_{1t})\to\mathbb{1}(\underline{x}'\beta_0\in\mathcal{Z}_{1t})=1$ since $\underline{x}'\beta_0$ is not in the boundary of $\mathcal{Z}_{1t}$ by B4.(v).

Therefore,

$$\mathbb{P}\left(\sup_{i=1,\ldots,N}|\pi_{it}(\widetilde{\beta})-\pi_{it}(\beta_0)|=0\right)\ge\mathbb{P}(\mathbb{1}(\underline{x}'\widetilde{\beta}\notin\mathcal{Z}_{1t})=0)=\mathbb{P}(\underline{x}'\widetilde{\beta}\in\mathcal{Z}_{1t})\to\mathbb{P}(\underline{x}'\beta_0\in\mathcal{Z}_{1t})=1$$

as $N\to\infty$. □

Lemma B.8 (ASF convergence in distribution)**.** Suppose B1–B6 hold. Then,

$$\sqrt{Nb_N}\left(\frac{1}{N}\sum_{i=1}^N\widehat{h}_1(\underline{x}'\beta_0,V_i;\beta_0)\pi_{it}-\mathbb{E}[h_1(\underline{x}'\beta_0,V;\beta_0)\pi_t]\right)\xrightarrow{d}\mathcal{N}(0,\sigma^2_{\mathrm{ASF}_t}(\underline{x}'\beta_0)).$$

Proof of Lemma B.8. This proof builds on the proof of Corollary 2 in Kong, Linton, and Xia (2010) (KLX hereafter). First we verify that Assumptions A1–A7 of KLX hold under ours. Their A1 holds with our squared-loss function, and we note that $\psi(\varepsilon_i)\equiv -2(Y_{it}-\mathbb{E}[Y_t|Z_{it}])$ in their notation. Since $Y_{it}\in\{0,1\}$, $\mathbb{E}[|\psi(\varepsilon_i)|^{\nu_1}]<\infty$ holds for arbitrary large ν_1. Their A2 holds immediately. A3 holds by Assumption B3. A4 and A5 holds by B4.(iii). A6 holds if

$$\begin{aligned}Nb_N^{1+d_V}/\log(N)&\to\infty\\ Nb_N^{1+d_V+2(\ell+1)}/\log(N)&=O(1)\\ N^{\nu_2/8-\lambda_1-1/4}b_N^{(1+d_V)(\nu_2/8-\lambda_1+3/4)}\log(N)^{-\nu_2/8+5/4+\lambda_1}&\to\infty,\end{aligned}$$

for some $2 < \nu_2 \leq \nu_1$. Since $b_N = \kappa \cdot N^{-\delta}$, these conditions are equivalent to

$$\begin{aligned} 1 - \delta(1 + d_V) &> 0 \\ 1 - \delta(3 + 2\ell + d_V) &\leq 0 \\ \nu_2/8 - \lambda_1 - 1/4 - \delta(1 + d_V)(\nu_2/8 - \lambda_1 + 3/4) &> 0. \end{aligned}$$

Since ν_1 can be made arbitrarily large, ν_2 can also taken to be arbitrarily large, and the last inequality is equivalent to

$$\delta < \frac{1}{1 + d_V}.$$

By our B6, these rate conditions all hold. Finally, their A7 holds by B4.(v). Since these assumptions hold for $\lambda_1 = 1$, we can use equation (13) in KLX and their Corollary 1 to write

$$\widehat{h}_1(z;\beta_0) = h_1(z;\beta_0) + B_{1,N}(z) + \frac{1}{N}\sum_{j=1}^{N} \phi_{1,jN}(z) + R_{1,N}(z)$$

where $B_{1,N}(z)$ is a bias term satisfying $\sup_{z \in \mathcal{Z}_t} |B_{1,N}(z)| = O(b_N^{\ell+1})$ if ℓ is odd or $O(b_N^{\ell+2})$ if ℓ is even, where $\phi_{1,jN}(z)$ are mean-zero random variables, and where $R_{1,N}(z)$ is a higher-order term satisfying $\sup_{z \in \mathcal{Z}_t} |R_{1,N}(z)| = O_p\left(\frac{\log(N)}{N b_N^{1+d_V}}\right)$.

Second, we note that

$$\sqrt{Nb_N}\left(\frac{1}{N}\sum_{i=1}^{N} \widehat{h}_1(\underline{x}'\beta_0, V_i;\beta_0)\pi_{it} - \mathbb{E}[h_1(\underline{x}'\beta_0, V;\beta_0)\pi_t]\right) = \sqrt{Nb_N}\frac{1}{N}\sum_{i=1}^{N}\left(\widehat{h}_1(\underline{x}'\beta_0, V_i;\beta_0) - h_1(\underline{x}'\beta_0, V_i;\beta_0)\right)\pi_{it} \tag{B.5}$$

$$+ \sqrt{b_N} \cdot \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left(h_1(\underline{x}'\beta_0, V_i;\beta_0)\pi_{it} - \mathbb{E}[h_1(\underline{x}'\beta_0, V;\beta_0)\pi_t]\right). \tag{B.6}$$

To analyze term (B.5), we use the fact that

$$\begin{aligned} &\sqrt{Nb_N}\frac{1}{N}\sum_{i=1}^{N}\left(\widehat{h}_1(\underline{x}'\beta_0, V_i;\beta_0) - h_1(\underline{x}'\beta_0, V_i;\beta_0)\right)\pi_{it} \\ &= \sqrt{Nb_N}\frac{1}{N}\sum_{i=1}^{N} B_{1,N}(\underline{x}'\beta_0, V_i)\pi_{it} + \sqrt{Nb_N}\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\phi_{1,jN}(\underline{x}'\beta_0, V_i)\pi_{it} + \sqrt{Nb_N}\frac{1}{N}\sum_{i=1}^{N} R_{1,N}(\underline{x}'\beta_0, V_i)\pi_{it}. \end{aligned}$$

When ℓ is odd, $\sqrt{Nb_N}\frac{1}{N}\sum_{i=1}^{N} B_{1,N}(\underline{x}'\beta_0, V_i)\pi_{it}$ is $o(1)$ because

$$\begin{aligned} \left|\sqrt{Nb_N}\frac{1}{N}\sum_{i=1}^{N} B_{1,N}(\underline{x}'\beta_0, V_i)\pi_{it}\right| &\leq \sqrt{Nb_N} \cdot \sup_{z \in \mathcal{Z}_t}\left|B_{1,N}(z)\right| \\ &= \sqrt{Nb_N} \cdot O(b_N^{\ell+1}) \\ &= O(\sqrt{Nb_N^{2\ell+3}}) \end{aligned}$$

and by B6. A similar derivation applies when ℓ is even.

We now show that term $\sqrt{Nb_N}\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\phi_{1,jN}(\underline{x}'\beta_0, V_i)\pi_{it}$ converges in distribution to a normal distribution. By standard arguments from Masry (1996), which are also referred to in the proof of Corollary 2 in KLX, we have that

$$\begin{aligned} &\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\phi_{1,jN}(\underline{x}'\beta_0, V_i)\pi_{it} \\ &= \frac{-1}{Nb_N}\sum_{i=1}^{N}(Y_{it} - \mathbb{E}[Y_t|Z_{it}])f_V(V_i)\mathbb{1}((\underline{x}'\beta_0, V_i) \in \mathcal{Z}_t) \\ &\cdot e_1' S_N(\underline{x}'\beta_0, V_i;\beta_0)^{-1}\int \mathcal{K}\left(\frac{X_{it}'\beta_0 - \underline{x}'\beta_0}{b_N}, v\right)\xi\left(\frac{X_{it}'\beta_0 - \underline{x}'\beta_0}{b_N}, v\right) dv\left(1 + O_p\left(\left(\frac{\log(N)}{Nb_N^{d_V}}\right)^{1/2}\right)\right) \\ &= \frac{-1}{Nb_N}\sum_{i=1}^{N}(Y_{it} - \mathbb{E}[Y_t|Z_{it}])f_V(V_i)\mathbb{1}((\underline{x}'\beta_0, V_i) \in \mathcal{Z}_t) \\ &\cdot e_1' S_N(\underline{x}'\beta_0, V_i;\beta_0)^{-1}\int \mathcal{K}\left(\frac{X_{it}'\beta_0 - \underline{x}'\beta_0}{b_N}, v\right)\xi\left(\frac{X_{it}'\beta_0 - \underline{x}'\beta_0}{b_N}, v\right) dv + o_p(1). \end{aligned}$$

We now calculate the asymptotic variance of

$$\frac{-1}{Nb_N}\sum_{i=1}^{N}(Y_{it}-\mathbb{E}[Y_t|Z_{it}])f_V(V_i)\mathbb{1}((\underline{x}'\beta_0,V_i)\in\mathcal{Z}_t)\cdot e_1'S_N(\underline{x}'\beta_0,V_i;\beta_0)^{-1}\int\mathcal{K}\left(\frac{X_{it}'\beta_0-\underline{x}'\beta_0}{b_N},v\right)\xi\left(\frac{X_{it}'\beta_0-\underline{x}'\beta_0}{b_N},v\right)dv.$$

We have that

$$\begin{aligned}
&\mathrm{Var}\left(\frac{-1}{Nb_N}\sum_{i=1}^{N}(Y_t-\mathbb{E}[Y_t|Z_t])f_V(V)\mathbb{1}((\underline{x}'\beta_0,V)\in\mathcal{Z}_t)\cdot e_1'S_N(\underline{x}'\beta_0,V;\beta_0)^{-1}\int\mathcal{K}\left(\frac{X_t'\beta_0-\underline{x}'\beta_0}{b_N},v\right)\xi\left(\frac{X_t'\beta_0-\underline{x}'\beta_0}{b_N},v\right)dv\right)\\
&=\frac{1}{Nb_N^2}\mathbb{E}\left[(Y_t-\mathbb{E}[Y_t|Z_t])^2f_V(V)^2\mathbb{1}((\underline{x}'\beta_0,V)\in\mathcal{Z}_t)e_1'S_N(\underline{x}'\beta_0,V;\beta_0)^{-1}\left(\int\mathcal{K}\left(\frac{X_t'\beta_0-\underline{x}'\beta_0}{b_N},v\right)\xi\left(\frac{X_t'\beta_0-\underline{x}'\beta_0}{b_N},v\right)dv\right)\right.\\
&\left.\left(\int\mathcal{K}\left(\frac{X_t'\beta_0-\underline{x}'\beta_0}{b_N},v\right)\xi\left(\frac{X_t'\beta_0-\underline{x}'\beta_0}{b_N},v\right)dv\right)'S_N(\underline{x}'\beta_0,V;\beta_0)^{-1}e_1\right]
\end{aligned}$$

Recall that $S_N(z;\beta_0)=S(z;\beta_0)+o_p(1)=\int\xi(a)\xi(a)'\mathcal{K}(a)\,da\cdot f_{Z_t(\beta_0)}(z)+o_p(1)$ uniformly in $z\in\mathcal{Z}_t$ by Lemma B.2. Therefore,

$$\begin{aligned}
&=\frac{1}{Nb_N^2}\mathbb{E}\left[\mathrm{Var}(Y_t|Z_t(\beta_0))\frac{f_V(V)^2}{f_{Z_t(\beta_0)}(\underline{x}'\beta_0,V)^2}\mathbb{1}((\underline{x}'\beta_0,V)\in\mathcal{Z}_t)e_1'\left(\int\xi(a)\xi(a)'\mathcal{K}(a)\,da\right)^{-1}\left(\int\mathcal{K}\left(\frac{X_t'\beta_0-\underline{x}'\beta_0}{b_N},v\right)\xi\left(\frac{X_t'\beta_0-\underline{x}'\beta_0}{b_N},v\right)dv\right)\right.\\
&\left.\cdot\left(\int\mathcal{K}\left(\frac{X_t'\beta_0-\underline{x}'\beta_0}{b_N},v\right)\xi\left(\frac{X_t'\beta_0-\underline{x}'\beta_0}{b_N},v\right)dv\right)'\left(\int\xi(a)\xi(a)'\mathcal{K}(a)\,da\right)^{-1}e_1\right]+o((Nb_N)^{-1})\\
&=\frac{1}{Nb_N^2}\mathbb{E}\left[\int\mathrm{Var}(Y_t|X_t'\beta_0=\tilde{u},V)\frac{f_V(V)^2}{f_{Z_t(\beta_0)}(\underline{x}'\beta_0,V)^2}\mathbb{1}((\underline{x}'\beta_0,V)\in\mathcal{Z}_t)e_1'\left(\int\xi(a)\xi(a)'\mathcal{K}(a)\,da\right)^{-1}\left(\int\mathcal{K}\left(\frac{\tilde{u}-\underline{x}'\beta_0}{b_N},v\right)\xi\left(\frac{\tilde{u}-\underline{x}'\beta_0}{b_N},v\right)dv\right)\right.\\
&\left.\cdot\left(\int\mathcal{K}\left(\frac{\tilde{u}-\underline{x}'\beta_0}{b_N},v\right)\xi\left(\frac{\tilde{u}-\underline{x}'\beta_0}{b_N},v\right)dv\right)'\left(\int\xi(a)\xi(a)'\mathcal{K}(a)\,da\right)^{-1}e_1\,f_{X_t'\beta_0|V}(\tilde{u}|V)\,d\tilde{u}\right]+o((Nb_N)^{-1})\\
&=\frac{1}{Nb_N}\mathbb{E}\left[\int\mathrm{Var}(Y_t|X_t'\beta_0=\underline{x}'\beta_0+b_Nu,V)\frac{f_V(V)^2}{f_{Z_t(\beta_0)}(\underline{x}'\beta_0,V)^2}\mathbb{1}((\underline{x}'\beta_0,V)\in\mathcal{Z}_t)e_1'\left(\int\xi(a)\xi(a)'\mathcal{K}(a)\,da\right)^{-1}\left(\int\mathcal{K}(z)\,\xi(z)\,dv\right)\right.\\
&\left.\cdot\left(\int\mathcal{K}(z)\,\xi(z)\,dv\right)'\left(\int\xi(a)\xi(a)'\mathcal{K}(a)\,da\right)^{-1}e_1f_{X_t'\beta_0|V}(\underline{x}'\beta_0+b_Nu|V)\,du\right]+o((Nb_N)^{-1})\\
&=\frac{1}{Nb_N}\mathbb{E}\left[\mathrm{Var}(Y_t|X_t'\beta_0=\underline{x}'\beta_0,V)\frac{f_V(V)}{f_{Z_t(\beta_0)}(\underline{x}'\beta_0,V)}\mathbb{1}((\underline{x}'\beta_0,V)\in\mathcal{Z}_t)\right]\\
&\cdot e_1'\left(\int\xi(a)\xi(a)'\mathcal{K}(a)\,da\right)^{-1}\int\left(\int\mathcal{K}(z)\,\xi(z)\,dv\right)\left(\int\mathcal{K}(z)\,\xi(z)\,dv\right)'du\left(\int\xi(a)\xi(a)'\mathcal{K}(a)\,da\right)^{-1}e_1+o((Nb_N)^{-1})\\
&=\frac{1}{Nb_N}\sigma^2_{\mathrm{ASF}_t}(\underline{x}'\beta_0)+o((Nb_N)^{-1}).
\end{aligned}$$

The third equality follows from the change of variables $\tilde{u}=\underline{x}'\beta_0+b_Nu$. The above equations re-derive and fix a minor typo in equation (A.42) in KLX. By the proof of Corollary 2 in KLX, we have that

$$\sqrt{Nb_N}\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\phi_{1,jN}(\underline{x}'\beta_0,V_i)\pi_{it}\xrightarrow{d}\mathcal{N}(0,\sigma^2_{\mathrm{ASF}_t}(\underline{x}'\beta_0)).$$

Also, the term $\sqrt{Nb_N}\frac{1}{N}\sum_{i=1}^{N}R_{1,N}(\underline{x}'\beta_0,V_i)\pi_{it}$ is $o_p(1)$ because

$$\begin{aligned}
\left|\sqrt{Nb_N}\frac{1}{N}\sum_{i=1}^{N}R_{1,N}(\underline{x}'\beta_0,V_i)\pi_{it}\right|&\le\sqrt{Nb_N}\cdot\sup_{z\in\mathcal{Z}_t}\left|R_{1,N}(z)\right|\\
&=\sqrt{Nb_N}\cdot O_p\left(\frac{\log(N)}{Nb_N^{1+d_V}}\right)\\
&=O_p\left(\frac{\log(N)}{\sqrt{Nb_N^{1+2d_V}}}\right)\\
&=o_p(1)
\end{aligned}$$

by B6.

Third, term (B.6) above is of order $O_p(\sqrt{b_N})=o_p(1)$ by an application of the central limit theorem.

Finally, we obtain that

$$\sqrt{Nb_N}\left(\frac{1}{N}\sum_{i=1}^{N}\widehat{h}_1(\underline{x}'\beta_0,V_i;\beta_0)\pi_{it}-\mathbb{E}[h_1(\underline{x}'\beta_0,V;\beta_0)\pi_t]\right)=\sqrt{Nb_N}\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\phi_{jN}(\underline{x}'\beta_0,V_i)\pi_{it}+o_p(1)$$
$$\xrightarrow{d}\mathcal{N}(0,\sigma^2_{\mathrm{ASF}_t}(\underline{x}'\beta_0)).$$

□

We use the following technical lemma in the proof of Theorem 3.1.

Lemma B.9. Let A and B be positive-definite, symmetric matrices. Let $\lambda_{\min}(A)$ denote the minimum eigenvalue of A. Then,

$$|\lambda_{\min}(A)-\lambda_{\min}(B)|\leq\|A-B\|.$$

Proof of Lemma B.9. Since A and B are positive-definite and symmetric, they are invertible and $\lambda_{\min}(A)=\|A^{-1}\|^{-1}>0$ and $\lambda_{\min}(B)=\|B^{-1}\|^{-1}>0$. We then have

$$\begin{aligned}
|\lambda_{\min}(A)-\lambda_{\min}(B)| &= |\|A^{-1}\|^{-1}-\|B^{-1}\|^{-1}| \\
&= |\|A^{-1}\|-\|B^{-1}\||\cdot\frac{1}{\|A^{-1}\|\|B^{-1}\|} \\
&\leq \|A^{-1}-B^{-1}\|\cdot\frac{1}{\|A^{-1}\|\|B^{-1}\|} \\
&= \|B^{-1}(B-A)A^{-1}\|\cdot\frac{1}{\|A^{-1}\|\|B^{-1}\|} \\
&\leq \|B^{-1}\|\|A-B\|\|A^{-1}\|\cdot\frac{1}{\|A^{-1}\|\|B^{-1}\|} \\
&= \|A-B\|.
\end{aligned}$$

The first inequality follows from the triangle inequality, and the second inequality from $\|CD\|\leq\|C\|\|D\|$ for the spectral norm and square matrices C and D. □

Proof of Theorem 3.1. We have the following decomposition:

$$\sqrt{Nb_N}\left(\widehat{\mathrm{ASF}}_t(\underline{x})-\mathrm{ASF}^{\pi}_t(\underline{x})\right)=\sqrt{Nb_N}\left(\frac{1}{N}\sum_{i=1}^{N}\left(\widehat{h}_1(\underline{x}'\widehat{\beta},V_i;\widehat{\beta})-\widehat{h}_1(\underline{x}'\widehat{\beta},V_i;\beta_0)\right)\hat{\pi}_{it}\right) \tag{B.7}$$
$$+\sqrt{Nb_N}\left(\frac{1}{N}\sum_{i=1}^{N}\left(\widehat{h}_1(\underline{x}'\widehat{\beta},V_i;\beta_0)-\widehat{h}_1(\underline{x}'\beta_0,V_i;\beta_0)\right)\hat{\pi}_{it}\right) \tag{B.8}$$
$$+\sqrt{Nb_N}\left(\frac{1}{N}\sum_{i=1}^{N}\widehat{h}_1(\underline{x}'\beta_0,V_i;\beta_0)(\hat{\pi}_{it}-\pi_{it})\right) \tag{B.9}$$
$$+\sqrt{Nb_N}\left(\frac{1}{N}\sum_{i=1}^{N}\widehat{h}_1(\underline{x}'\beta_0,V_i;\beta_0)\pi_{it}-\mathbb{E}[h_1(\underline{x}'\beta_0,V;\beta_0)\pi_t]\right). \tag{B.10}$$

We break down the proof in four parts. In the first three parts, we show that terms (B.7)–(B.9) are $o_p(1)$. In the fourth and last part, we show that term (B.10) converges in distribution.

Part 1: Convergence of Term (B.7)

We have that

$$\begin{aligned}
&\sqrt{Nb_N}\cdot\left|\frac{1}{N}\sum_{i=1}^{N}\left(\widehat{h}_1(\underline{x}'\widehat{\beta},V_i;\widehat{\beta})-\widehat{h}_1(\underline{x}'\widehat{\beta},V_i;\beta_0)\right)\hat{\pi}_{it}\right| \\
&=\sqrt{Nb_N}\cdot\left|\frac{1}{N}\sum_{j=1}^{N}e_1'\left(S_N(\underline{x}'\widehat{\beta},V_i;\widehat{\beta})^{-1}T_N(\underline{x}'\widehat{\beta},V_i;\widehat{\beta})-S_N(\underline{x}'\widehat{\beta},V_i;\beta_0)^{-1}T_N(\underline{x}'\widehat{\beta},V_i;\beta_0)\right)\mathbb{1}((\underline{x}'\widehat{\beta},V_i)\in\mathcal{Z}_t)\right| \\
&=\sqrt{Nb_N}\cdot\left|\frac{1}{N}\sum_{j=1}^{N}e_1'\left(S_N(\underline{x}'\widehat{\beta},V_i;\widehat{\beta})^{-1}(T_N(\underline{x}'\widehat{\beta},V_i;\widehat{\beta})-T_N(\underline{x}'\widehat{\beta},V_i;\beta_0))\right.\right. \\
&\left.\left.+S_N(\underline{x}'\widehat{\beta},V_i;\widehat{\beta})^{-1}\left(S_N(\underline{x}'\widehat{\beta},V_i;\beta_0)-S_N(\underline{x}'\widehat{\beta},V_i;\widehat{\beta})\right)S_N(\underline{x}'\widehat{\beta},V_i;\beta_0)^{-1}T_N(\underline{x}'\widehat{\beta},V_i;\beta_0)\right)\mathbb{1}((\underline{x}'\widehat{\beta},V_i)\in\mathcal{Z}_t)\right| \\
&\leq\sqrt{Nb_N}\cdot\|e_1\|\sup_{z\in\mathcal{Z}_t}\left\|S_N(z;\widehat{\beta})^{-1}\right\|\sup_{z\in\mathcal{Z}_t}\left\|T_N(z;\widehat{\beta})-T_N(z;\beta_0)\right\| \\
&+\sqrt{Nb_N}\cdot\|e_1\|\sup_{z\in\mathcal{Z}_t}\left\|S_N(z;\widehat{\beta})^{-1}\right\|\sup_{z\in\mathcal{Z}_t}\left\|S_N(z;\widehat{\beta})-S_N(z;\beta_0)\right\|\sup_{z\in\mathcal{Z}_t}\left\|S_N(z;\beta_0)^{-1}\right\|\sup_{z\in\mathcal{Z}_t}\|T_N(z;\beta_0)\|.
\end{aligned}$$

The terms in the previous expressions are of these asymptotic orders:

- $\|e_1\|=1$.
- $\left\|S_N(z;\widehat{\beta})^{-1}\right\|=\lambda_{\min}\left(S_N(z;\widehat{\beta})\right)^{-1}$, where $\lambda_{\min}(\cdot)$ denotes the minimum eigenvalue of a symmetric matrix. We have that

$$\begin{aligned}
\sup_{z\in\mathcal{Z}_t}\left|\lambda_{\min}\left(S_N(z;\widehat{\beta})\right)-\lambda_{\min}\left(S(z;\beta_0)\right)\right|&\leq\sup_{z\in\mathcal{Z}_t}\left\|S_N(z;\widehat{\beta})-S(z;\beta_0)\right\| \\
&\leq\sup_{z\in\mathcal{Z}_t}\left\|S_N(z;\widehat{\beta})-S_N(z;\beta_0)\right\|+\sup_{z\in\mathcal{Z}_t}\|S_N(z;\beta_0)-S(z;\beta_0)\| \\
&=o_p\left(\frac{1}{\sqrt{Nb_N}}\right)+O_p\left(\left(\frac{\log(N)}{Nb_N^{1+d_V}}\right)^{1/2}\right)+O(b_N) \\
&=o_p(1).
\end{aligned}$$

  The first line follows from Lemma B.9. The second line follows from the triangle inequality. The third line follows from Lemma B.1 and B.2. The last line follows from B6. Also note that

$$\inf_{z\in\mathcal{Z}_t}\lambda_{\min}\left(S(z;\beta_0)\right)=\inf_{z\in\mathcal{Z}_t}f_{Z_t}(z)\cdot\lambda_{\min}\left(\int\xi(a)\xi(a)'\mathcal{K}(a)\,da\right)>0.$$

  This follows from the definition of the set $\mathcal{Z}_t$, which is such that $\inf_{z\in\mathcal{Z}_t}f_{Z_t}(z)>0$: see B4.(ii). $\int\xi(a)\xi(a)'\mathcal{K}(a)\,da$ is positive definite since, for $c\in\mathbb{R}^{\bar{N}}$ such that $c\neq\mathbf{0}$,

$$c'\left(\int\xi(a)\xi(a)'\mathcal{K}(a)\,da\right)c=\int(c'\xi(a))^2\mathcal{K}(a)\,da=0$$

  implies that $c'\xi(a)=0$ for all a in the support of $\mathcal{K}(a)$. Since $\xi(a)$ is comprised of products of powers of components of a, $c'\xi(a)=0$ over this entire support implies $c=\mathbf{0}$, a contradiction. Therefore $\lambda_{\min}\left(\int\xi(a)\xi(a)'\mathcal{K}(a)\,da\right)>0$ and $\inf_{z\in\mathcal{Z}_t}\lambda_{\min}\left(S(z;\beta_0)\right)>0$.
  This implies that,

$$\begin{aligned}
\sup_{z\in\mathcal{Z}_t}\left\|S_N(z;\widehat{\beta})^{-1}\right\|&=\frac{1}{\inf_{z\in\mathcal{Z}_t}\lambda_{\min}\left(S_N(z;\widehat{\beta})\right)} \\
&\leq\frac{1}{\inf_{z\in\mathcal{Z}_t}\lambda_{\min}\left(S(z;\beta_0)\right)-\sup_{z\in\mathcal{Z}_t}\left|\lambda_{\min}\left(S_N(z;\widehat{\beta})\right)-\lambda_{\min}\left(S(z;\beta_0)\right)\right|} \\
&=\frac{1}{\inf_{z\in\mathcal{Z}_t}\lambda_{\min}\left(S(z;\beta_0)\right)-o_p(1)} \\
&=O_p(1).
\end{aligned}$$

- By Lemma B.1, we have that $\sup_{z\in\mathcal{Z}_t}\left\|S_N(z;\widehat{\beta})-S_N(z;\beta_0)\right\|=o_p\left(\frac{1}{\sqrt{Nb_N}}\right)$.
- By Lemma B.3, we have that $\sup_{z\in\mathcal{Z}_t}\left\|T_N(z;\widehat{\beta})-T_N(z;\beta_0)\right\|=o_p\left(\frac{1}{\sqrt{Nb_N}}\right)$
- As above, we have that $\sup_{z\in\mathcal{Z}_t}\left\|S_N(z;\beta_0)^{-1}\right\|=O_p(1)$.

- We have that

$$\sup_{z\in\mathcal{Z}_t}\|T_N(z;\beta_0)\| \leq \sup_{z\in\mathcal{Z}_t}\|T_N(z;\beta_0)-T(z;\beta_0)\| + \sup_{z\in\mathcal{Z}_t}\|T(z;\beta_0)\|$$

where

$$\sup_{z\in\mathcal{Z}_t}\|T_N(z;\beta_0)-T(z;\beta_0)\| = O_p\left(\left(\frac{\log(N)}{Nb_N^{1+d_V}}\right)^{1/2}\right) + O(b_N)$$

by Lemma B.4. We also have that

$$\begin{aligned}\sup_{z\in\mathcal{Z}_t}\|T(z;\beta_0)\| &= \sup_{z\in\mathcal{Z}_t}|\mathbb{E}[Y_t|Z_t=z]f_{Z_t}(z)|\cdot\left\|\int\xi(a)\mathcal{K}(a)\,da\right\| \\ &\leq 1\cdot\sup_{z\in\mathcal{Z}_t}f_{Z_t}(z)\cdot O(1) \\ &= O(1)\end{aligned}$$

by $\sup_{z\in\mathcal{Z}_t}f_{Z_t}(z)<\infty$ (Assumption B4.(iii)), and by $\left\|\int\xi(a)\mathcal{K}(a)\,da\right\|<\infty$ (Assumption B3). Therefore,

$$\begin{aligned}\sup_{z\in\mathcal{Z}_t}\|T_N(z;\beta_0)\| &= O_p\left(\left(\frac{\log(N)}{Nb_N^{1+d_V}}\right)^{1/2}\right) + O(b_N) + O(1) \\ &= O_p(1),\end{aligned}$$

by B6.

Combining the asymptotic orders of the above six terms, we have

$$\begin{aligned}\sqrt{Nb_N}\cdot\left|\frac{1}{N}\sum_{i=1}^N\left(\widehat{h}_1(\underline{x}'\widehat{\beta},V_i;\widehat{\beta})-\widehat{h}_1(\underline{x}'\widehat{\beta},V_i;\beta_0)\right)\hat{\pi}_{it}\right| &\leq \sqrt{Nb_N}\cdot O_p(1)\cdot o_p\left(\frac{1}{\sqrt{Nb_N}}\right) \\ &\quad+\sqrt{Nb_N}\cdot O_p(1)\cdot o_p\left(\frac{1}{\sqrt{Nb_N}}\right)\cdot O_p(1)\cdot O_p(1) \\ &= o_p(1).\end{aligned}$$

Part 2: Convergence of Term (B.8)

We have that

$$\begin{aligned}&\left|\frac{1}{N}\sum_{i=1}^N\left(\widehat{h}_1(\underline{x}'\widehat{\beta},V_i;\beta_0)-\widehat{h}_1(\underline{x}'\beta_0,V_i;\beta_0)\right)\hat{\pi}_{it}\right| \\ &= \left|\frac{1}{N}\sum_{j=1}^N e_1'\left(S_N(\underline{x}'\widehat{\beta},V_i;\beta_0)^{-1}T_N(\underline{x}'\widehat{\beta},V_i;\beta_0)-S_N(\underline{x}'\beta_0,V_i;\beta_0)^{-1}T_N(\underline{x}'\beta_0,V_i;\beta_0)\right)\mathbb{1}((\underline{x}'\widehat{\beta},V_i)\in\mathcal{Z}_t)\right| \\ &= \Bigg|\frac{1}{N}\sum_{j=1}^N e_1'\Big(S_N(\underline{x}'\widehat{\beta},V_i;\beta_0)^{-1}(T_N(\underline{x}'\widehat{\beta},V_i;\beta_0)-T_N(\underline{x}'\beta_0,V_i;\beta_0)) \\ &\quad+S_N(\underline{x}'\widehat{\beta},V_i;\beta_0)^{-1}\left(S_N(\underline{x}'\beta_0,V_i;\beta_0)-S_N(\underline{x}'\widehat{\beta},V_i;\beta_0)\right)S_N(\underline{x}'\beta_0,V_i;\beta_0)^{-1}T_N(\underline{x}'\beta_0,V_i;\beta_0)\Big)\mathbb{1}((\underline{x}'\widehat{\beta},V_i)\in\mathcal{Z}_t)\Bigg| \\ &\leq \|e_1\|\sup_{z\in\mathcal{Z}_t}\left\|S_N(z;\beta_0)^{-1}\right\|\sup_{z\in\mathcal{Z}_t}\left\|\frac{\partial}{\partial u}T_N(z;\beta_0)\right\|\left\|\underline{x}'\widehat{\beta}-\underline{x}'\beta_0\right\| \\ &\quad+\|e_1\|\sup_{z\in\mathcal{Z}_t}\left\|S_N(z;\beta_0)^{-1}\right\|\sup_{z\in\mathcal{Z}_t}\left\|\frac{\partial}{\partial u}S_N(z;\beta_0)\right\|\left\|\underline{x}'\widehat{\beta}-\underline{x}'\beta_0\right\|\sup_{z\in\mathcal{Z}_t}\left\|S_N(z;\beta_0)^{-1}\right\|\sup_{z\in\mathcal{Z}_t}\|T_N(z;\beta_0)\|. \qquad \text{(B.11)}\end{aligned}$$

The inequality follows from applications of the mean-value theorem and the Cauchy-Schwarz inequality. By Lemma B.3 and B.5,

$$\begin{aligned}\sup_{z\in\mathcal{Z}_t}\left\|\frac{\partial}{\partial u}S_N(z;\beta_0)\right\| &= o_p\left(\frac{a_N}{\sqrt{Nb_N}}\right) \\ \sup_{z\in\mathcal{Z}_t}\left\|\frac{\partial}{\partial u}T_N(z;\beta_0)\right\| &= o_p\left(\frac{a_N}{\sqrt{Nb_N}}\right).\end{aligned}$$

By B2, $\|\underline{x}'\widehat{\beta}-\underline{x}'\beta_0\|\leq\|\underline{x}\|\|\widehat{\beta}-\beta_0\| = O_p(a_N^{-1})$. The asymptotic order of all other terms in equation (B.11) were

characterized in the analysis of the convergence of term (B.7). Therefore

$$\begin{aligned}
&\sqrt{Nb_N} \cdot \left| \frac{1}{N} \sum_{i=1}^{N} \left(\widehat{h}_1(\underline{x}'\widehat{\beta}, V_i; \beta_0) - \widehat{h}_1(\underline{x}'\beta_0, V_i; \beta_0) \right) \hat{\pi}_{it} \right| \\
&= \sqrt{Nb_N} \cdot O_p(1) \cdot o_p\left(\frac{a_N}{\sqrt{Nb_N}} \right) \cdot O_p(a_N^{-1}) + \sqrt{Nb_N} \cdot O_p(1) \cdot o_p\left(\frac{a_N}{\sqrt{Nb_N}} \right) \cdot O_p(a_N^{-1}) \cdot O_p(1) \cdot O_p(1) \\
&= o_p(1).
\end{aligned}$$

Part 3: Convergence of Term (B.9)

First note that

$$\left| \frac{1}{N} \sum_{i=1}^{N} \widehat{h}_1(\underline{x}'\beta_0, V_i; \beta_0)(\hat{\pi}_{it} - \pi_{it}) \right| \leq \frac{1}{N} \sum_{i=1}^{N} \left| \widehat{h}_1(\underline{x}'\beta_0, V_i; \beta_0) \right| \cdot \sup_{i=1,\ldots,N} |\hat{\pi}_{it} - \pi_{it}| .$$

Therefore

$$\begin{aligned}
\mathbb{P}\left(\sqrt{Nb_N} \left| \frac{1}{N} \sum_{i=1}^{N} \widehat{h}_1(\underline{x}'\beta_0, V_i; \beta_0)(\hat{\pi}_{it} - \pi_{it}) \right| = 0 \right) &\geq \mathbb{P}\left(\frac{1}{N} \sum_{i=1}^{N} \left| \widehat{h}_1(\underline{x}'\beta_0, V_i; \beta_0) \right| \cdot \sup_{i=1,\ldots,N} |\hat{\pi}_{it} - \pi_{it}| = 0 \right) \\
&\geq \mathbb{P}\left(\sup_{i=1,\ldots,N} |\hat{\pi}_{it} - \pi_{it}| = 0 \right) \\
&\to 1
\end{aligned}$$

as $N \to \infty$ by Lemma B.7. Therefore

$$\sqrt{Nb_N} \left| \frac{1}{N} \sum_{i=1}^{N} \widehat{h}_1(\underline{x}'\beta_0, V_i; \beta_0)(\hat{\pi}_{it} - \pi_{it}) \right| = o_p(1)$$

Part 4: Convergence of Term (B.10)

By Lemma B.8, this term converges in distribution:

$$\sqrt{Nb_N} \left(\frac{1}{N} \sum_{i=1}^{N} \widehat{h}_1(\underline{x}'\beta_0, V_i; \beta_0)\pi_{it} - E[h_1(\underline{x}'\beta_0, V; \beta_0)\pi_t] \right) \xrightarrow{d} \mathcal{N}(0, \sigma^2_{\mathrm{ASF}_t}(\underline{x}'\beta_0)).$$

The conclusion follows from an application of Slutsky's Theorem. □

Lemma B.10 (APE convergence in distribution)**.** Suppose B1–B6 hold. Then,

$$\sqrt{Nb_N^3} \left(\frac{1}{N} \sum_{i=1}^{N} \widehat{h}_2(\underline{x}'\beta_0, V_i; \beta_0)\pi_{it} - \mathbb{E}[h_2(\underline{x}'\beta_0, V; \beta_0)\pi_t] \right) \xrightarrow{d} \mathcal{N}(0, \sigma^2_{\mathrm{APE}_t}(\underline{x}'\beta_0)).$$

Proof of Lemma B.10. This proof builds on that of Corollary 2 in KLX and our Lemma B.8. Recall that Assumptions A1–A7 of KLX hold under ours. We can then use equation (13) in KLX and their Corollary 1 to write

$$\begin{aligned}
b_N \widehat{h}_2(z; \beta_0) &= b_N h_2(z; \beta_0) + B_{2,N}(z) + \frac{1}{N} \sum_{j=1}^{N} \phi_{2,jN}(z) + R_{2,N}(z) \\
&= e'_{2+d_V} h(z; \beta_0) + B_{2,N}(z) + \frac{1}{N} \sum_{j=1}^{N} \phi_{2,jN}(z) + R_{2,N}(z),
\end{aligned}$$

where $B_{2,N}(z)$ is a bias term satisfying $\sup_{z \in \mathcal{Z}_t} |B_{2,N}(z)| = O(b_N^{\ell+1})$ if ℓ is odd or $O(b_N^{\ell+2})$ if ℓ is even, where $\phi_{2,jN}(z)$ are mean-zero random variables, and where $R_{2,N}(z)$ is a higher-order term satisfying $\sup_{z \in \mathcal{Z}_t} |R_{2,N}(z)| = O_p\left(\frac{\log(N)}{Nb_N^{1+d_V}} \right)$.

Second, note that

$$\begin{aligned}
\sqrt{Nb_N^3} \left(\frac{1}{N} \sum_{i=1}^{N} \widehat{h}_2(\underline{x}'\beta_0, V_i; \beta_0)\pi_{it} - \mathbb{E}[h_2(\underline{x}'\beta_0, V; \beta_0)\pi_t] \right) &= \sqrt{Nb_N^3} \frac{1}{N} \sum_{i=1}^{N} \left(\widehat{h}_2(\underline{x}'\beta_0, V_i; \beta_0) - h_2(\underline{x}'\beta_0, V_i; \beta_0) \right) \pi_{it} \quad \text{(B.12)} \\
&\quad + \sqrt{b_N^3} \cdot \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \left(h_2(\underline{x}'\beta_0, V_i; \beta_0)\pi_{it} - \mathbb{E}[h_2(\underline{x}'\beta_0, V; \beta_0)\pi_t] \right) \quad \text{(B.13)}
\end{aligned}$$

To analyze term (B.12), we use the fact that

$$\begin{aligned}
&\sqrt{Nb_N^3}\frac{1}{N}\sum_{i=1}^{N}\left(\widehat{h}_2(\underline{x}'\beta_0,V_i;\beta_0)-h_2(\underline{x}'\beta_0,V_i;\beta_0)\right)\pi_{it}\\
&=\sqrt{Nb_N}\frac{1}{N}\sum_{i=1}^{N}e'_{2+d_V}\left(\widehat{h}(\underline{x}'\beta_0,V_i;\beta_0)-h(\underline{x}'\beta_0,V_i;\beta_0)\right)\pi_{it}\\
&=\sqrt{Nb_N}\frac{1}{N}\sum_{i=1}^{N}B_{2,N}(\underline{x}'\beta_0,V_i)\pi_{it}+\sqrt{Nb_N}\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\phi_{2,jN}(\underline{x}'\beta_0,V_i)\pi_{it}+\sqrt{Nb_N}\frac{1}{N}\sum_{i=1}^{N}R_{2,N}(\underline{x}'\beta_0,V_i)\pi_{it}.
\end{aligned}$$

The terms $\sqrt{Nb_N}\frac{1}{N}\sum_{i=1}^{N}B_{2,N}(\underline{x}'\beta_0,V_i)\pi_{it}$ and $\sqrt{Nb_N}\frac{1}{N}\sum_{i=1}^{N}R_{2,N}(\underline{x}'\beta_0,V_i)\pi_{it}$ are $o_p(1)$ from the same arguments used in the proof of Lemma B.8.

The term $\sqrt{Nb_N}\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\phi_{2,jN}(\underline{x}'\beta_0,V_i)\pi_{it}$ converges in distribution to

$$\sqrt{Nb_N}\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\phi_{2,jN}(\underline{x}'\beta_0,V_i)\pi_{it}\xrightarrow{d}\mathcal{N}(0,\sigma^2_{\mathrm{APE}_t}(\underline{x}'\beta_0))$$

by standard arguments from Masry (1996) referred to in the proof of Corollary 2 in KLX.

Term (B.13) above is of order $O_p(b_N^{3/2})=o_p(1)$ by an application of the central limit theorem. Therefore,

$$\begin{aligned}
\sqrt{Nb_N^3}\left(\frac{1}{N}\sum_{i=1}^{N}\widehat{h}_2(\underline{x}'\beta_0,V_i;\beta_0)\pi_{it}-\mathbb{E}[h_2(\underline{x}'\beta_0,V;\beta_0)\pi_t]\right)&=\sqrt{Nb_N}\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\phi_{2,jN}(\underline{x}'\beta_0,V_i)\pi_{it}+o_p(1)\\
&\xrightarrow{d}\mathcal{N}(0,\sigma^2_{\mathrm{APE}_t}(\underline{x}'\beta_0)).
\end{aligned}$$

□

Proof of Theorem 3.2. First, we write

$$\begin{aligned}
\sqrt{Nb_N^3}\left(\widehat{\mathrm{APE}}_{k,t}(\underline{x})-\mathrm{APE}^{\pi}_{k,t}(\underline{x})\right)&=\widehat{\beta}^{(k)}\cdot\sqrt{Nb_N^3}\left(\frac{1}{N}\sum_{i=1}^{N}\left(\widehat{h}_2(\underline{x}'\widehat{\beta},V_i;\widehat{\beta})-\widehat{h}_2(\underline{x}'\widehat{\beta},V_i;\beta_0)\right)\hat{\pi}_{it}\right) &\text{(B.14)}\\
&+\widehat{\beta}^{(k)}\cdot\sqrt{Nb_N^3}\left(\frac{1}{N}\sum_{i=1}^{N}\left(\widehat{h}_2(\underline{x}'\widehat{\beta},V_i;\beta_0)-\widehat{h}_2(\underline{x}'\beta_0,V_i;\beta_0)\right)\hat{\pi}_{it}\right) &\text{(B.15)}\\
&+\widehat{\beta}^{(k)}\cdot\sqrt{Nb_N^3}\left(\frac{1}{N}\sum_{i=1}^{N}\widehat{h}_2(\underline{x}'\beta_0,V_i;\beta_0)(\hat{\pi}_{it}-\pi_{it})\right) &\text{(B.16)}\\
&+\widehat{\beta}^{(k)}\cdot\sqrt{Nb_N^3}\left(\frac{1}{N}\sum_{i=1}^{N}\widehat{h}_2(\underline{x}'\beta_0,V_i;\beta_0)\pi_{it}-\mathbb{E}[h_2(\underline{x}'\beta_0,V;\beta_0)\pi_t]\right) &\text{(B.17)}\\
&+\sqrt{Nb_N^3}(\widehat{\beta}^{(k)}-\beta_0^{(k)})\cdot\mathbb{E}[h_2(\underline{x}'\beta_0,V;\beta_0)\pi_t]. &\text{(B.18)}
\end{aligned}$$

We will show that terms (B.14)–(B.16) and (B.18) are $o_p(1)$, and that term (B.17) converges in distribution.

Convergence of Term (B.14)

Note that

$$\begin{aligned}
&\sqrt{Nb_N^3}\cdot\left|\frac{1}{N}\sum_{i=1}^{N}\left(\widehat{h}_2(\underline{x}'\widehat{\beta},V_i;\widehat{\beta})-\widehat{h}_2(\underline{x}'\widehat{\beta},V_i;\beta_0)\right)\hat{\pi}_{it}\right|\\
&=\sqrt{Nb_N}\cdot\left|\frac{1}{N}\sum_{j=1}^{N}e'_{2+d_V}\left(S_N(\underline{x}'\widehat{\beta},V_i;\widehat{\beta})^{-1}T_N(\underline{x}'\widehat{\beta},V_i;\widehat{\beta})-S_N(\underline{x}'\widehat{\beta},V_i;\beta_0)^{-1}T_N(\underline{x}'\widehat{\beta},V_i;\beta_0)\right)\mathbb{1}((\underline{x}'\widehat{\beta},V_i)\in\mathcal{Z}_t)\right|
\end{aligned}$$

by the definition of $\widehat{h}_2$. Also note that $\widehat{\beta}^{(k)}=O_p(1)$. Therefore, we can follow the same steps used in the proof of Theorem 3.1 to show term (B.7) is $o_p(1)$.

Convergence of Term (B.15)

We have that

$$\sqrt{Nb_N^3} \cdot \left| \widehat{\beta}^{(k)} \frac{1}{N} \sum_{i=1}^{N} \left(\widehat{h}_2(\underline{x}'\widehat{\beta}, V_i; \beta_0) - \widehat{h}_2(\underline{x}'\beta_0, V_i; \beta_0) \right) \hat{\pi}_{it} \right|$$

$$= \left| \widehat{\beta}^{(k)} \right| \cdot \sqrt{Nb_N} \cdot \left| \frac{1}{N} \sum_{j=1}^{N} e'_{2+d_V} \left(S_N(\underline{x}'\widehat{\beta}, V_i; \beta_0)^{-1} T_N(\underline{x}'\widehat{\beta}, V_i; \beta_0) - S_N(\underline{x}'\beta_0, V_i; \beta_0)^{-1} T_N(\underline{x}'\beta_0, V_i; \beta_0) \right) \mathbb{1}((\underline{x}'\widehat{\beta}, V_i) \in \mathcal{Z}_t) \right|.$$

Again, we can follow the same steps used in the proof of Theorem 3.1 to show term (B.8) is $o_p(1)$.

Convergence of Term (B.16)

The convergence of this term is shown in an identical manner to that of term (B.9).

Convergence of Term (B.17)

By Lemma B.10, we have that $\sqrt{Nb_N^3} \left(\frac{1}{N} \sum_{i=1}^{N} \widehat{h}_2(\underline{x}'\beta_0, V_i; \beta_0)\pi_{it} - \mathbb{E}[h_2(\underline{x}'\beta_0, V; \beta_0)\pi_t] \right) \xrightarrow{d} \mathcal{N}(0, \sigma^2_{\mathrm{APE}_t}(\underline{x}'\beta_0))$. By B2, $\widehat{\beta}^{(k)} \xrightarrow{p} \beta_0^{(k)}$. Therefore, by Slutsky's Theorem, term (B.17) converges in distribution to a mean-zero Gaussian distribution with variance $(\beta_0^{(k)})^2 \cdot \sigma^2_{\mathrm{APE}_t}(\underline{x}'\beta_0)$.

Convergence of Term (B.18)

Note that $\mathbb{E}[h_2(\underline{x}'\beta_0, V; \beta_0)\pi_t] = O(1)$. Term (B.18) is of order $\sqrt{Nb_N^3}(\widehat{\beta}^{(k)} - \beta_0^{(k)}) \cdot O(1) = O_p\left(\sqrt{Nb_N^3} a_N^{-1}\right)$. By B2, the order of this term is

$$O_p\left(N^{\frac{1}{2}(1-3\delta-2\epsilon)}\right) = o_p(1).$$

This equality follows from $\delta > \frac{1-2\epsilon}{3}$, which can be seen from $\delta > 1 - 2\epsilon$ and $\delta > 0$: see B6.

Combining the convergence of terms (B.14)–(B.18) with Slutsky's Theorem, we obtain our result. □

C Proofs for Section 4

Proof of Proposition 4.1. By A1–A2 and Lemma 2.1, β_0 is point-identified. Under $\mathrm{supp}(X_t'\beta_0, V) = \mathbb{R} \times \mathcal{V}$, the conditional probability $\mathbb{P}(Y_t = 1 | X_t'\beta_0 = u, V = v)$ is identified for all $(u, v) \in \mathbb{R} \times \mathcal{V}$. Assumption A3.(i) and basic manipulations show that this conditional probability equals $F_{U_t - C|V}(u|v)$. Therefore, the conditional distribution of $U_t - C$ conditional on V is point identified. By A1.(ii), U_t and C are independent given V, and

$$\mathbb{E}[\exp(i\zeta(U_t - C))|V = v] = \mathbb{E}[\exp(i\zeta U_t)|V = v] \cdot \mathbb{E}[\exp(-i\zeta C)|V = v]$$

for any $\zeta \in \mathbb{R}$, where $i = \sqrt{-1}$.

By A1, we have that $U_t|V \overset{d}{=} U_t$. By A2.(i), the distribution of U_t is known (standard logistic) and has a characteristic function with no zeros. We can then write

$$\mathbb{E}[\exp(-i\zeta C)|V = v] = \frac{\mathbb{E}[\exp(i\zeta(U_t - C))|V = v]}{\mathbb{E}[\exp(i\zeta U_t)]},$$

where the right-hand side is identified from the data.

From the inversion formula for characteristic functions, this implies the conditional distribution of $C|V = v$ is identified for all $v \in \mathcal{V}$. By the law of iterated expectations, the distribution of C is also identified. □

Proof of Proposition 4.2. By A1, A2', A3, and Lemma 2.1, β_0 is identified up to scale. Under $\mathrm{supp}(X_s'\beta_0, X_t'\beta_0, V) = \mathbb{R}^2 \times \mathcal{V}$, the conditional probability $\mathbb{P}(Y_s = 1, Y_t = 1 | X_s'\beta_0 = u_1, X_t'\beta_0 = u_2, V = v)$ is identified for all $(u_1, u_2, v) \in \mathbb{R}^2 \times \mathcal{V}$. Assumption A3.(i) and basic manipulations show that this conditional probability equals $F_{U_s - C, U_t - C|V}(u_1, u_2|v)$. Therefore, the conditional distribution of $(U_s - C, U_t - C)$ conditional on V is identified. This implies the joint distribution of $(U_s - C, U_t - C)$ is identified by the law of iterated expectations.

Note that (U_s, U_t, C) are jointly independent by Assumption A1.(ii). Therefore, we can apply Kotlarski's lemma (Kotlarski, 1967) to obtain the distributions of U_s, U_t, and C. □

D Additional Figures and Tables

This section presents additional figures and tables that supplement the main results in the text.

D.1 Monte Carlo Simulation

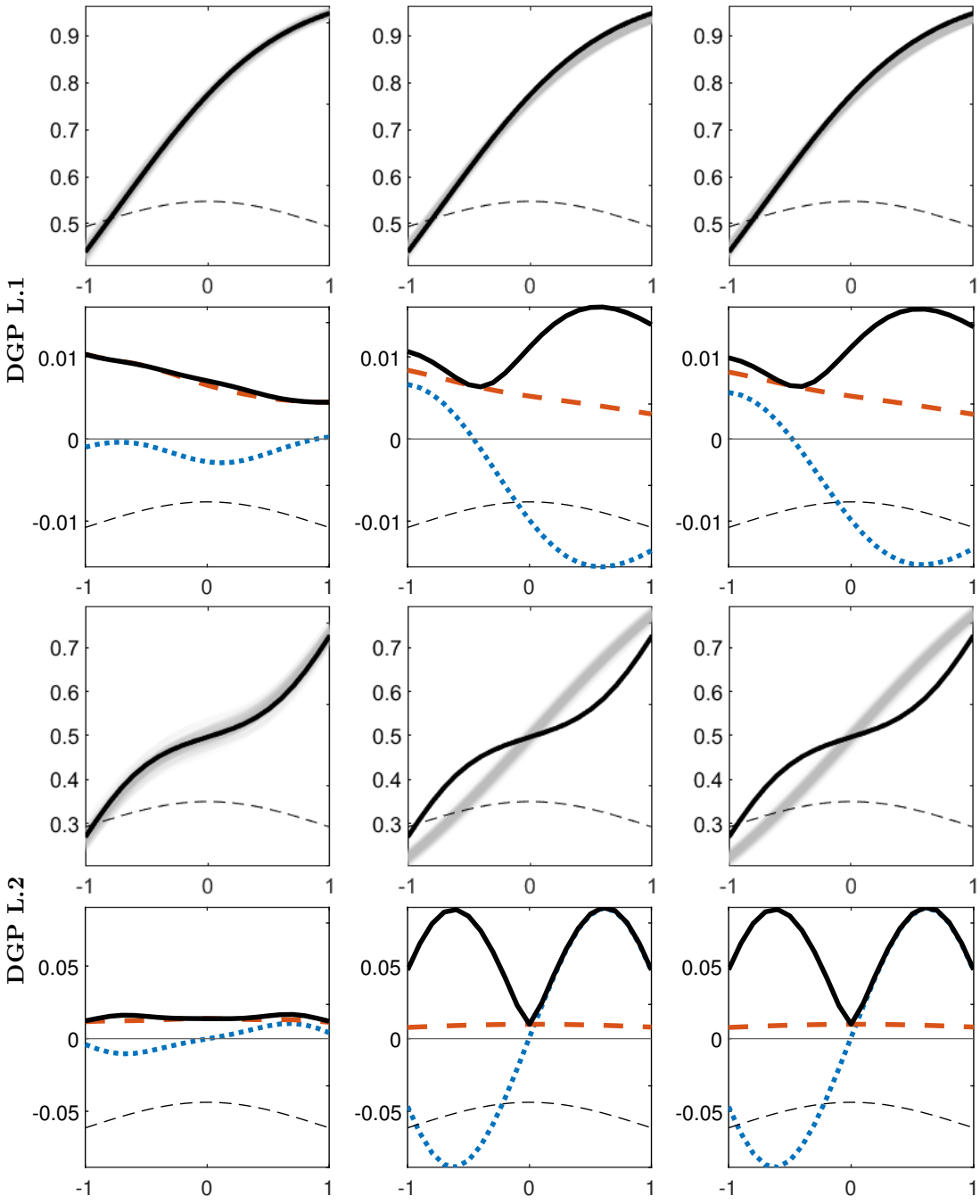
In the logit case, Figure 5 shows the estimated ASFs and their corresponding performance statistics. In the general case, Figure 6 depicts the APE performance statistics, and Figures 7 and 8 present the estimated ASFs and performance statistics, respectively.

Table 6: Estimation of Common Parameter and ASF - Logit Case

		$\widehat{\beta}$			ASF					
		Bias	SD	RMSE	|Bias|	SD	RMSE	Min	Med.	Max
DGP L.1	Semiparam.	-0.001	0.023	0.023	0.006	0.007	**0.007**	0.5%	0.9%	2.3%
	RE	-0.094	0.019	0.096	0.010	0.005	0.012	1.0%	1.6%	2.4%
	CRE	-0.092	0.018	0.094	0.010	0.005	0.011	1.0%	1.6%	2.3%
DGP L.2	Semiparam.	0.002	0.049	0.049	0.012	0.013	**0.015**	1.6%	2.8%	4.6%
	RE	-0.502	0.025	0.502	0.061	0.009	0.062	2.0%	14.0%	22.8%
	CRE	-0.501	0.025	0.501	0.061	0.009	0.062	2.0%	14.0%	22.8%

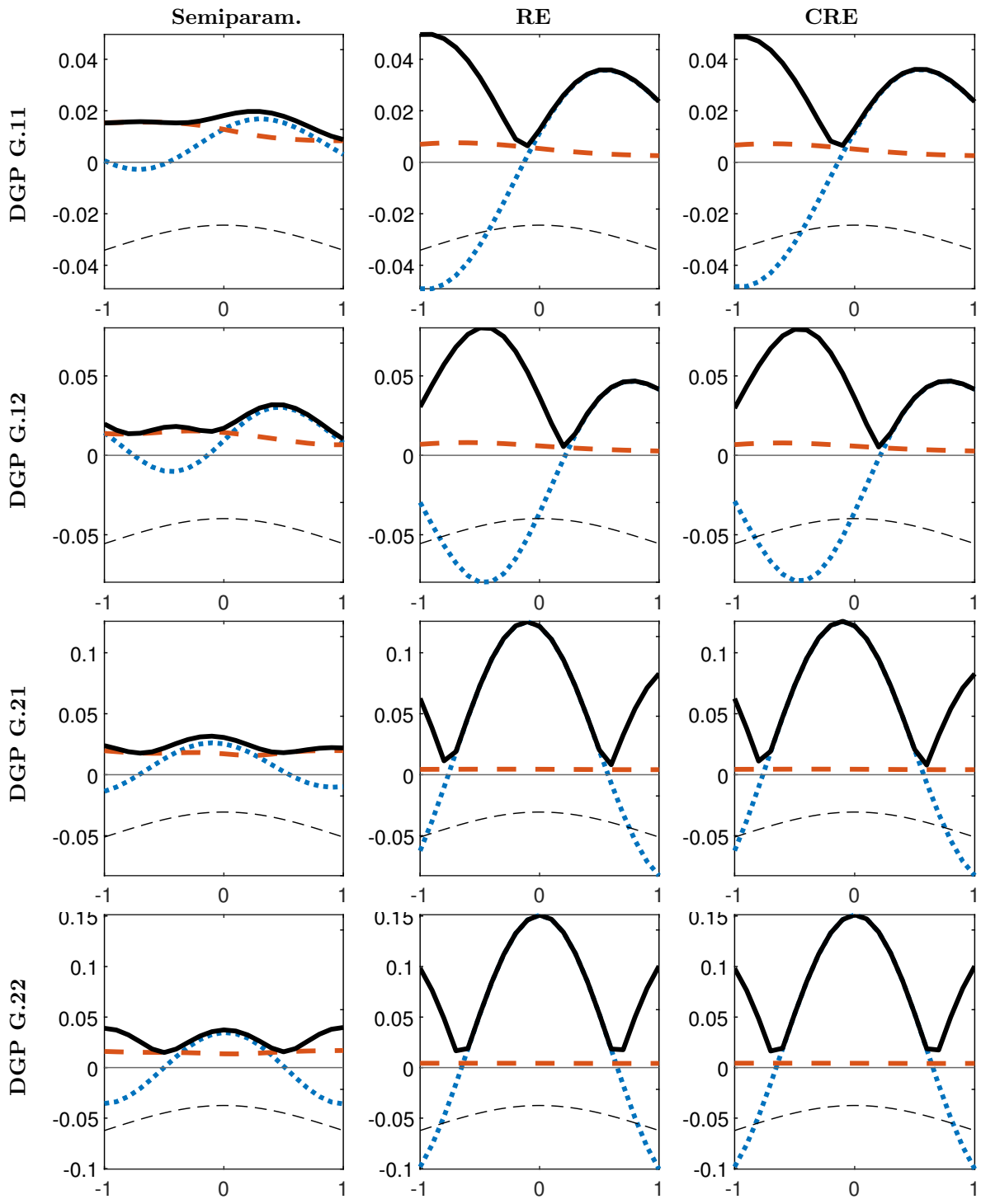
Notes: In the true DGPs, the error term $U_t \sim \sqrt{3}/\pi \cdot$ standard logistic, so that $\text{Var}(U_t) = 1$, whereas the logit estimators are based on standard logistic errors. Here we multiply the estimated $\widehat{\beta}$ by $\sqrt{3}/\pi$ to be comparable to the true β_0. |Bias| indicates the absolute value of the bias. The |Bias|, SD, and RMSE of the ASF are weighted averages across the collection of evaluation points $\underline{x}$, where the weights are proportional to $f_{X_t}(\underline{x})$. Bold entries indicate the best ASF estimator (i.e., with the smallest RMSE) for each DGP. The last three columns are the minimum/median/maximum of RMSE($\underline{x}$)/ASF($\underline{x}$) $\times$ 100% over $\underline{x}$.

Figure 5: ASF Estimation - Logit Case



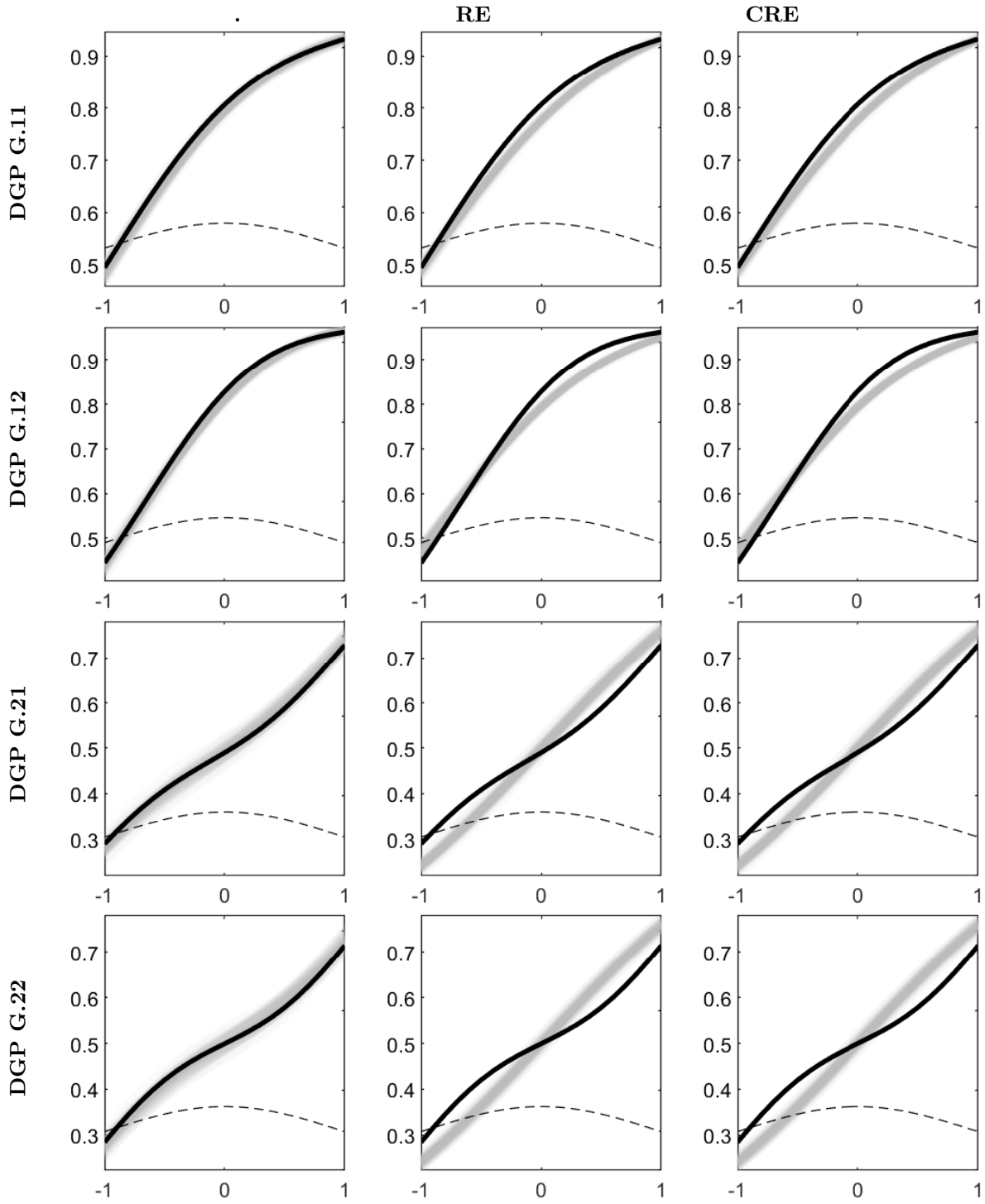
Notes: X-axes are potential values $\underline{x}$. In the first and third rows, black solid lines are the true ASF, gray bands are collections of lines where each line corresponds to the estimated ASF based on one simulation repetition. In the second and fourth rows, black solid / blue dotted / red dashed lines represent the RMSEs / biases / standard deviations of the ASF estimates. Thin dashed lines at the bottom of all panels show $f_{X_t}(\underline{x})$.

Figure 6: Bias, Standard Deviation, and RMSE in APE Estimation - General Case



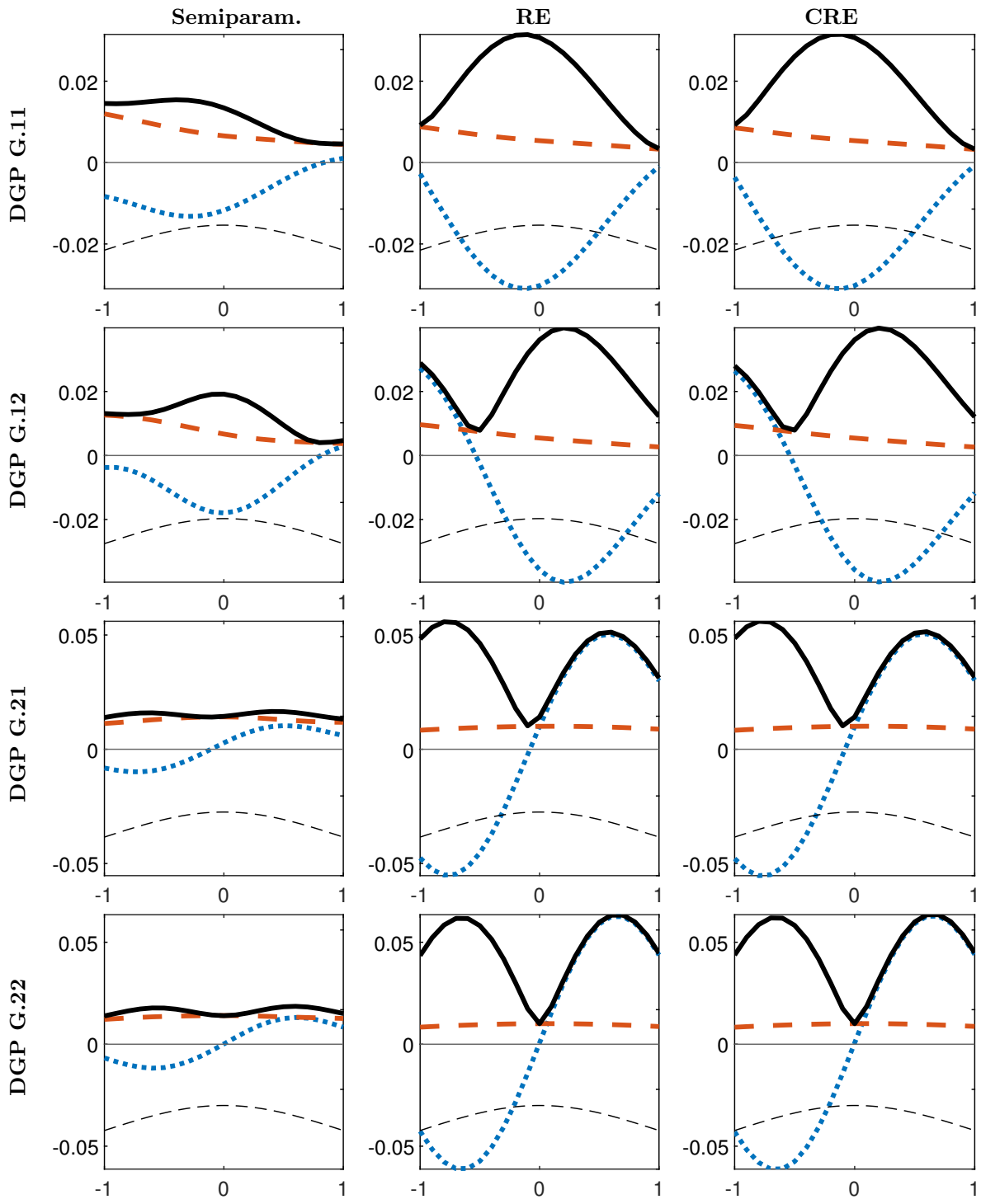
Notes: X-axes are potential values $\underline{x}^{(2)}$. Black solid / blue dotted / red dashed lines represent the RMSEs / biases / standard deviations of the APE estimates. Thin dashed lines at the bottom of all panels show $f_{X_t^{(2)}}\left(\underline{x}^{(2)}\right)$.

Figure 7: Estimated ASF vs True ASF - General Case



Notes: X-axes are potential values $\underline{x}^{(2)}$. Black solid lines are the true ASF. Gray bands are collections of lines where each line corresponds to the estimated ASF based on one simulation repetition. Thin dashed lines at the bottom of all panels show $f_{X_t^{(2)}}\left(\underline{x}^{(2)}\right)$..

Figure 8: Bias, Standard Deviation, and RMSE in ASF Estimation - General Case



Notes: X-axes are potential values $\underline{x}^{(2)}$. Black solid / blue dotted / red dashed lines represent the RMSEs / biases / standard deviations of the ASF estimates. Thin dashed lines at the bottom of all panels show $f_{X_t^{(2)}}\left(\underline{x}^{(2)}\right)$.

Table 7: Estimation of Common Parameter and ASF - General Case

		$\widehat{\beta}^{(2)}$			ASF					
		Bias	SD	RMSE	|Bias|	SD	RMSE	Min	Med.	Max
DGP G.11	Semiparam.	0.011	0.031	0.033	0.010	0.007	**0.011**	0.5%	1.7%	2.9%
	RE	0.004	0.023	0.023	0.020	0.006	0.021	0.4%	2.8%	4.1%
	CRE	0.005	0.022	0.023	0.020	0.006	0.021	0.4%	2.8%	4.2%
DGP G.12	Semiparam.	0.012	0.026	0.028	0.012	0.007	**0.013**	0.4%	2.2%	2.9%
	RE	0.005	0.019	0.020	0.025	0.006	0.026	1.2%	3.4%	6.5%
	CRE	0.006	0.019	0.020	0.025	0.006	0.026	1.2%	3.4%	6.3%
DGP G.21	Semiparam.	0.015	0.064	0.065	0.012	0.013	**0.015**	1.8%	3.0%	4.8%
	RE	0.007	0.041	0.042	0.037	0.010	0.038	2.2%	7.7%	16.8%
	CRE	0.008	0.043	0.043	0.037	0.010	0.039	2.2%	7.7%	16.9%
DGP G.22	Semiparam.	0.011	0.072	0.073	0.013	0.013	**0.016**	2.1%	3.2%	4.9%
	RE	0.004	0.043	0.043	0.044	0.010	0.045	2.0%	9.5%	16.9%
	CRE	0.005	0.044	0.044	0.044	0.010	0.045	2.0%	9.5%	17.0%

Notes: For the RE and CRE, we normalize $\widehat{\beta}$ such that $|\widehat{\beta}^{(1)}| = 1$ to allow comparisons across estimators. |Bias| indicates the absolute value of the bias. The |Bias|, SD, and RMSE of the ASF are weighted averages across the collection of evaluation points $\underline{x}$, where the weights are proportional to $f_{X_t}(\underline{x})$. Bold entries indicate the best ASF estimator (i.e., with the smallest RMSE) for each DGP. The last three columns are the minimum/median/maximum of RMSE($\underline{x}$)/ASF($\underline{x}$) $\times$ 100% over $\underline{x}$.

Table 8: APE Estimation - General Case, Semiparametric Estimator

		|Bias|	SD	RMSE
DGP G.11	Known β_0	0.0125	0.0115	0.0151
	Unknown β_0	0.0133	0.0123	0.0161
DGP G.12	Known β_0	0.0176	0.0117	0.0200
	Unknown β_0	0.0178	0.0123	0.0204
DGP G.21	Known β_0	0.0199	0.0144	0.0232
	Unknown β_0	0.0199	0.0145	0.0233
DGP G.22	Known β_0	0.0217	0.0185	0.0258
	Unknown β_0	0.0218	0.0187	0.0260

Notes: |Bias| indicates the absolute value of the bias. The reported |Bias|, SD, and RMSE are weighted averages across the collection of evaluation points $\underline{x}$, where the weights are proportional to $f_{X_t}(\underline{x})$.

D.2 Empirical Illustration

Alternative specifications. We collapse two discrete index variables into a binary variable and a trinary variable in our benchmark specification to ensure a sufficient number of observations in each cell to implement the semiparametric estimator. To explore the effects of the coarsening scheme on the empirical findings, we examined a range of alternative specifications, and the semiparametric estimator is generally robust across different coarsening schemes. For conciseness, here we focus on two alternatives (see Figure 10 below):

(i) Model "30": The total number of children is collapsed into a trinary variable depending on whether it is below the 33rd quantile, between the 33rd and 67th quantiles, or above the 67th quantile. Initial age is treated as a continuous index variable. Then, the number of continuous index variables equals $d_V = 2$.

(ii) Model "22": Both the total number of children and initial age are collapsed into binary indicators depending on whether the index variable is above or below its median. We have $d_V = 1$ in this case.

Local logit. The local logit estimator is a special case of the local likelihood approach, which locally fits nonlinear parametric models using the MLE (Tibshirani and Hastie, 1987; Fan, Farmen, and Gijbels, 1998; Frölich, 2006). With binary outcomes, the logistic likelihood function becomes an appealing choice, yielding the "local logit" estimator. Let $\lambda(\cdot)$ and $\Lambda(\cdot)$ denote the pdf and cdf of a standard logistic distribution. In the same spirit as the local-polynomial-regression-based estimator in Sections 3.2 and 3.3, we first replace the squared residuals with the logit likelihood in the optimization problem

$$\widehat{h}(z;\widehat{\beta}) = \underset{h\in\mathbb{R}^{\bar{N}}}{\operatorname{argmax}} \sum_{j=1}^{N} \left[Y_{jt} \log \Lambda\left(\xi\left(\frac{Z_{jt}(\widehat{\beta})-z}{b_N}\right)' h\right) + (1-Y_{jt}) \log\left(1-\Lambda\left(\xi\left(\frac{Z_{jt}(\widehat{\beta})-z}{b_N}\right)' h\right)\right)\right] \mathcal{K}_{b_N}\left(\frac{Z_{jt}(\widehat{\beta})-z}{b_N}\right).$$

Then, we obtain the ASF and APE estimates by averaging over the empirical marginal distribution of V_i:

$$\widehat{\mathrm{ASF}}_t(\underline{x}) = \frac{1}{N}\sum_{i=1}^{N} \Lambda\left(\widehat{h}_1(\underline{x}'\widehat{\beta}, V_i;\widehat{\beta})\right)\hat{\pi}_{it},$$

$$\widehat{\mathrm{APE}}_{k,t}(\underline{x}) = \widehat{\beta}^{(k)} \cdot \frac{1}{N}\sum_{i=1}^{N} \lambda\left(\widehat{h}_1(\underline{x}'\widehat{\beta}, V_i;\widehat{\beta})\right)\widehat{h}_2(\underline{x}'\widehat{\beta}, V_i;\widehat{\beta})\hat{\pi}_{it}.$$

Compared with local polynomial regression, one main advantage of local logit is that it is naturally tailored to the binary nature of the outcome, and ensures the estimated ASF is between 0 and 1 (see its estimates based on Model "22" in the bottom row in Figure 10). However, in our numerical experiments, the local logit estimate greatly worsens as the dimension of the index increases given a finite sample. Besides, local logit calls for numerical optimization at each local point $(\underline{x}'\widehat{\beta}, V_i)$ and hence requires more computational efforts than local polynomial regression, which is done using closed-form expressions. For example, it takes around 10 seconds to estimate Model "22" using local polynomial regression but around 5 minutes using local logit with the same polynomial order $\ell = 2$. This difference in computation time is likely larger in models with higher-dimensional indices.

Figures and tables. Table 9 summarizes descriptive statistics of the observables and supplements Figure 3 in the main text.

Figure 9 depicts the estimated coefficients on time dummies which capture time-variation in aggregate participation rates. Point estimates of the time profiles are generally parallel with each other (from top to bottom: the smoothed maximum score, RE, and CRE) and show higher participation rates after 1983, which coincides with the beginning of the Great Moderation. Most of the time-variation within each estimator and difference across estimators are insignificant at the 5% level, and standard errors generally increase with time for all three estimators. The smoothed maximum score yields the widest confidence band, as expected.

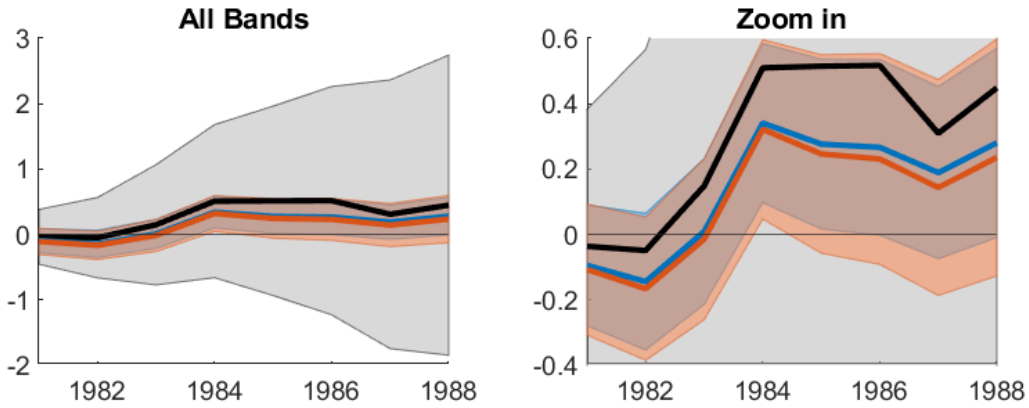
Figure 10 plots the estimated ASF and APE based on alternative specifications. Comparing with the benchmark specification in Figure 4, we see that in general the local-polynomial-regression-based estimates do not change much as we vary the coarsening scheme of the index variables. The local logit estimator helps narrow the bands in Model "22" (the second row versus the third row in Figure 10). Yet, its estimates may perform poorly for multidimensional indices in a finite sample, and it takes longer to compute, so we leave it out of the main analyses.

Table 9: Descriptive Statistics - Female Labor Force Participation

	25%	Med.	75%	Mean	SD	Skew.	Kurt.
(a) Full Sample, #obs = $N \times T$ = 13,149							
Participate	-	-	-	0.72	0.45	-	-
Children 0–2	0	0	0	0.23	0.47	1.99	6.79
Children 3–5	0	0	1	0.29	0.51	1.60	4.85
Children 6–17	0	1	2	1.05	1.10	0.91	3.46
Log Husband's Income	10.09	10.51	10.83	10.43	0.69	-0.89	7.27
Age	30.00	35.00	43.00	37.30	9.22	0.56	2.50
(b) Always Participate, %obs = 46.27%							
Children 0–2	0	0	0	0.18	0.41	2.25	7.56
Children 3–5	0	0	0	0.23	0.46	1.93	6.12
Children 6–17	0	1	2	1.00	1.06	0.91	3.47
Log Husband's Income	10.08	10.47	10.77	10.37	0.65	-1.36	8.89
Age	31.00	36.00	44.00	37.98	9.04	0.51	2.45
(c) Never Participate, %obs = 8.28%							
Children 0–2	0	0	0	0.21	0.47	2.35	8.50
Children 3–5	0	0	0	0.23	0.48	2.05	6.79
Children 6–17	0	1	2	0.99	1.19	1.30	4.54
Log Husband's Income	10.13	10.62	11.04	10.53	0.85	-0.74	6.52
Age	35.00	43.00	52.00	42.98	10.09	-0.06	1.90
(d) Movers, %obs = 45.45%							
Participate	-	-	-	0.57	0.49	-	-
Children 0–2	0	0	1	0.28	0.51	1.70	5.74
Children 3–5	0	0	1	0.36	0.56	1.27	3.82
Children 6–17	0	1	2	1.11	1.11	0.83	3.18
Log Husband's Income	10.11	10.55	10.87	10.47	0.69	-0.59	5.81
Age	29.00	34.00	40.00	35.57	8.71	0.73	2.88

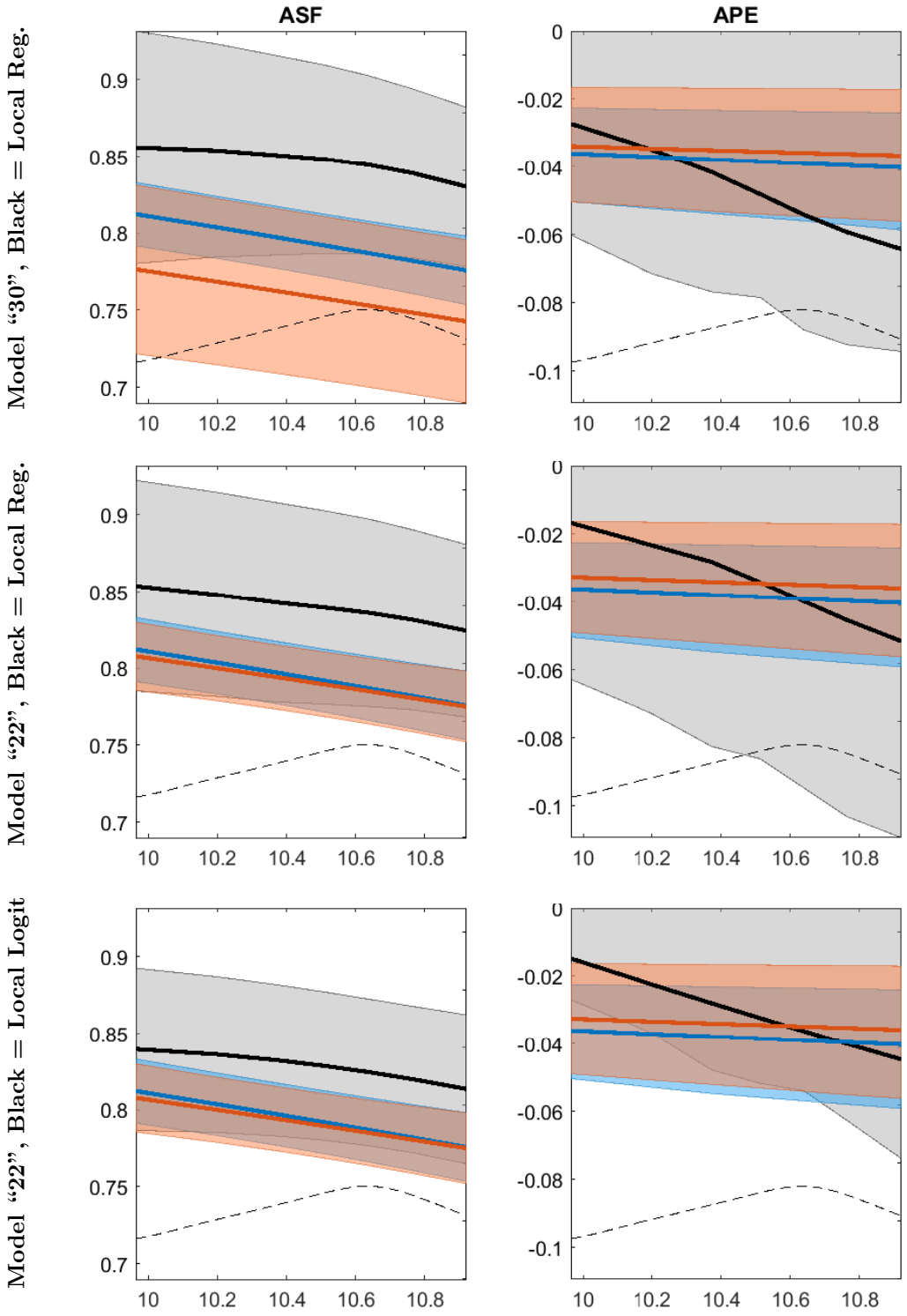
Notes: The sample consists of $N = 1461$ married women observed for $T = 9$ years from the PSID between 1980–1988. "Movers" refers to women who participate in the labor market in some years but not all. See Fernández-Val (2009) for details.

Figure 9: Estimated Coefficients on Time Dummies - Female Labor Force Participation



Notes: Black/blue/orange solid lines represent point estimates of the coefficients on time dummies using the smoothed maximum score/RE/CRE. Bands with corresponding colors indicate the 95% symmetric percentile-t confidence intervals based on bootstrap standard deviations. The right panel further zooms in on y-axis values between -0.4 and 0.6.

Figure 10: Estimated ASF and APE under Alternative Specifications - Female Labor Force Participation



Notes: X-axes are potential values of log husband's income. Blue/orange solid lines represent point estimates of the ASF and APE using the RE/CRE. Bands with corresponding colors indicate the 95% bootstrap confidence intervals. Thin dashed lines at the bottom of all panels show the distribution of log husband's income.