Average Partial Effects in Nonseparable Panel Data Models: Identification and Testing

Dalia A. Ghanem*

University of California, San Diego

November 13, 2012

Abstract

In many empirical settings involving microeconomic panel data, the researcher's objective is to identify the average partial effect of a variable on an outcome of interest. This paper examines nonparametric identification of average partial effects in this setting and proposes tests of identifying assumptions that do not rely on functional-form restrictions. The paper first studies the nonparametric identification problem starting from a data-generating process that exhibits both individual and time heterogeneity. The trade-off between identifying assumptions that restrict individual and/or time heterogeneity is formally characterized. The paper then proposes a menu of identifying assumptions that the empirical researcher may choose from. To test the identifying assumptions, bootstrap-adjusted Kolmogorov-Smirnov and Cramer-von-Mises statistics are proposed and are shown to be asymptotically valid. The tests include a nonparametric test of the fixed effects assumption, which is applied to the human capital earnings function using a subsample of the national longitudinal survey of youth, previously used in Angrist and Newey (1991).

*I am grateful for the guidance and support of Graham Elliott and Andres Santos. I am especially thankful to my advisor Graham Elliott for nurturing me intellectually. I would also like to thank Ivana Komunjer, Prashant Bharadwaj and Julie Cullen for helpful discussions, and Richard Carson, Philip Neary, Leah Nelson, and participants at the UCSD econometrics seminar for helpful comments. Finally, I am especially grateful to Prashant Bharadwaj for sharing the dataset used here.

1 Introduction

In many empirical settings, the researcher's objective is to identify the *ceteris paribus* effect of a particular variable on an outcome of interest; for instance, the effect of a job training program on participants' employment or income, the effect of having children on female labor participation, and the effect of schooling on income. While economic theory gives us some direction about which variables to include in order to identify the *ceteris paribus* effect, the functional form governing the relationship between the variables is often left unspecified. This is where nonparametric identification becomes particularly useful, even when parametric models are used as an approximation to the unknown relationship for estimation purposes.

The paper at hand examines the identification of the average partial effect (APE) of a discrete regressor in a nonseparable panel data model, where the time dimension, T, is fixed. This reflects the situation in many microeconomic panel data settings, where we observe a large number of individuals over a short time horizon. In this setup, the APE is point-identified only for a subpopulation.[1]

This paper starts with an unrestricted data-generating process (DGP), where the outcome variable for individual i in period t, Y_{it}, is given by the following,

$$Y_{it} = \xi_t(X_{it}, \mathcal{A}_i, \mathcal{U}_{it}), \quad \text{for } i = 1, 2, ..., n \text{ and } t = 1, 2, ..., T \tag{1}$$

where X_{it} is a vector of regressors, $\mathcal{A}_i$ includes time-invariant, individual-specific variables that are unobservable,[2] and $\mathcal{U}_{it}$ are idiosyncratic shocks. This paper is concerned with static models. Hence, X_{it} does not include lags of the outcome variable as well as other regressors that may introduce feedback mechanisms from Y_{it} to X_{is} for $s \geq t$.[3] The structural function, $\xi_t(.)$, is assumed to be unknown and is allowed to vary over time.[4]

[1] For the purposes of this paper, we will use use identification and point-identification interchangeably. If *a priori* bounds exist for the outcome variable, then bounds on the APE can be constructed as in Chernozhukov, Fernandez-Val, Hahn and Newey (2010). However, for the purposes of this paper, I am concerned with point-identification only.

[2] In linear models, $\mathcal{A}_i$ is referred to as an individual fixed effect. I refrain from this terminology, since it misrepresents what $\mathcal{A}_i$ stands for.

[3] This rules out dynamic selection.

[4] It is structural in the sense that $Y_{it}^x = \xi_t(x, \mathcal{A}_i, \mathcal{U}_{it})$, where Y_{it}^x is the outcome variable when X_{it} is fixed to x.

Furthermore, observables and unobservables are allowed to interact in arbitrary ways in the structural function. Without further assumptions, the DGP reflects arbitrary individual and time heterogeneity, which I formally define in Section 3.

Restrictions on this general DGP achieve identification of the APE for a subpopulation.[5] There are three goals behind starting with a general DGP: (1) to formally characterize the trade-off between identifying assumptions that restrict the structural function, individual and/or time heterogeneity, (2) to present a menu of identifying assumptions, (3) to propose tests thereof.

From a theoretical viewpoint, the characterization of the trade-off contributes to the understanding of nonparametric identification of APEs in nonseparable panel data models. The menu of identifying assumptions includes existing methods in the literature as a special case. It also suggests new methods that fall under both the fixed-effects and correlated-random-effects categories, which are referred to in this paper as within-group and within-period identification, respectively. Finally, the testing problem addresses new issues for testing the equality of distributions in the two-sample problem, where samples are dependent and the data is possibly demeaned. Bootstrap-adjusted Kolmogorov-Smirnov (KS) and Cramer-von-Mises (CM) statistics are proposed and shown to be asymptotically valid.

For practical purposes, the menu of identifying assumptions provides the empirical researcher with a set of alternative identification strategies to choose from. Since the identification strategies may not be justified *a priori*, the tests are tools to aid the empirical researcher in justifying the choice of a particular identification strategy. The tests proposed here include a nonparametric test of the fixed-effects assumption in the presence of time effects.

Angrist and Newey (1991) propose an over-identification test of the fixed-effects assumption in the linear model and find evidence against it for the human capital earnings function using a subsample of the national longitudinal survey of youth (NLSY). We revisit the same dataset in the empirical illustration. When we apply our nonparametric test, we do

[5]The intuition here is similar to the distinction between average treatment effect (ATE) and Treatment on the Treated (TOT). For fixed T, we only observe a subpopulation with and without the treatment. In the presence of unobservable heterogeneity, we can only point-identify the APE for a subpopulation and cannot point-identify the APE for the entire population.

not find evidence against the fixed-effects assumption.[6] We also estimate the APE. Our results indicate that implications of the linear model, such as constant APEs across time, are violated. We conclude that testing nonparametric identification may be useful to the empirical researcher, even when parametric models are used for estimation purposes to approximate the unknown structural relationship.

Related Literature. Athey and Imbens (2006), Chernozhukov, Fernandez-Val, Hahn, and Newey (2010), hereinafter, CFHN2010, and Hoderlein and White (2009) impose restrictions on time heterogeneity to achieve identification. We will refer to this category of restrictions as time homogeneity, but it is also referred to as time invariance and time stationarity in the literature.[7] The papers mentioned above are categorized as 'fixed effects', since they do not restrict individual heterogeneity. As CFHN2010 point out, time homogeneity is a very strong assumption, and its empirical content may be thought of as time being "randomly assigned." Bester and Hansen (2009) propose a correlated random effects approach where they impose restrictions on individual heterogeneity, but allow for time heterogeneity. Altonji and Matzkin (2005) discuss the issue of identifying average effects for cross-sectional and panel data models. Although they do not explicitly allow for time heterogeneity, they propose exchangeability restrictions that are similar in spirit to Bester and Hansen (2009)'s approach.

Similar to the paper at hand, the approach of the aforementioned papers is focused on the identification of a particular object of interest, the APE of a discrete or continuous regressor. The approach in this strand of the literature extends the intuition of differencing out individual heterogeneity in linear models to nonseparable models, which was termed by Magnac (2004) and Hoderlein and White (2009) as quasi-differencing. Hence, this literature is quite relevant for situations where the linear model is used to approximate an unknown, possibly nonseparable model. In the absence of separability, quasi-differencing here is simply averaging over unobservables in an "appropriate way."

The quasi-differencing approach originated in parametric binary choice models, such as

[6] The test in Angrist and Newey (1991) uses restrictions implied by the linear model in addition to restrictions implied by the fixed-effects assumptions. Hence, the rejection of the test may be either due to violations of the linear model or the fixed-effects assumption.

[7] CFHN2010 show that in the presence of location and scale time effects, one can still identify the effect of a discrete regressor using time homogeneity.

conditional logit (Chamberlain 1984, 2010), where the presence of a sufficient statistic for the individual effect allowed for the identification of the common parameter while treating individual heterogeneity nonparametrically. Magnac (2004) introduces the concept of quasi-differencing as the presence of a sufficient statistic for the individual effect. The term quasi-differencing has been more generally used to refer to identification strategies that extend the intuition of differencing in the linear model to more general settings, where the distribution of individual heterogeneity is left unrestricted. For semiparamteric binary choice models, Honore and Kyriazydou (2000) use the intuition of quasi-differencing to nonparametrically identify the common parameter. Hoderlein and White (2009) refer to their approach as quasi-differencing. For random coefficient models, Graham and Powell (2012) also use a differencing approach to identify the APE.

This paper further generalizes the concept of quasi-differencing to refer more generally to any approach where the APE is identified using average changes of the outcome variable across time or subpopulations that coincide with a change in the variable of interest.[8] This definition allows us to include works such as Altonji and Matzkin (2005) and Bester and Hansen (2009) under this category.

The key intuition behind the quasi-differencing approach is that if we seek to identify the APE of a regressor from average changes of the outcome variable across time, then we have to assume that the distribution of unobservables does not change over time. However, if we would like to use average changes of the outcome variable across subpopulations, then we have to assume that the distribution of unobservables is the same across subpopulations. Hence, some form of homogeneity assumption is required either across time periods or subpopulations. Time homogeneity assumptions are widely used in the literature as noted above. Homogeneity assumptions across subpopulations are less prevalent in the nonparametric identification literature. For continuous regressors, Bester and Hansen (2009) proposes assumptions that are closest to this in spirit.[9]

Another strand in the literature follows the classical identification approach, which seeks

[8]It is important to note here that the generalization of quasi-differencing here is different from the generalization proposed in Evdokimov (2011), which is used to identify all structural objects. However, both generalizations allow us to include situations where the structural function is not assumed to be stationary across time.

[9]For cross-sectional setups, Angrist (2004) uses this type of homogeneity assumptions to relate local average treatment effect (LATE) to the average treatment effect (ATE).

to identify all structural objects, i.e. the structural function and the conditional distribution of unobservables, which include Altonji and Matzkin (2005), Evdokimov (2010), and Evdokimov (2011). The key differences between this category and the quasi-differencing approach is that the former identifies all objects, whereas the latter focuses on a specific object and hence requires weaker conditions. In other words, the quasi-differencing literature answers the question: What object can one identify under minimal assumptions? The classical identification literature on the other hand answers a different question: What assumptions are sufficient to identify all structural objects? Both contribute immensely to our understanding of the possibilities and limits of identification of various objects of interest from panel data without parametric assumptions. The paper at hand falls under the quasi-differencing category. Its findings are related to the classical identification literature and attempts to point to the relative merits of both approaches.

Outline of the Paper. The rest of the paper is organized as follows: Section 2 illustrates the basic identification problem with an empirical example. Section 3 presents the unrestricted DGP and includes the characterization of the trade-off between identifying assumptions. It also gives a menu of identification strategies. Section 4 includes the results for the bootstrap-adjusted tests. Section 5 includes an empirical illustration using a subsample of the national longitudinal survey of youth (NLSY). Section 6 concludes.

2 Basic Identification Problem: Job Training Example

Now let our variable of interest, X, be a binary variable for the participation in a job training program and our outcome of interest, Y, be earnings. Our object of interest is the effect of participating in a job training program on earnings, and we observe individuals in two time periods, $t = 1, 2$. We assume that there are three subpopulations: (1) individuals that never participate in the job training program, $X_i = (0, 0)$, (2) individuals that participate in the first time period, but not in the second, $X_i = (1, 0)$, (3) individuals that participate in the second time period, but not in the first, $X_i = (0, 1)$.

Now there are several unobservable factors here that may confound our effect of interest

if we examine average changes across time or subpopulations that coincide with changes in job training status: (1) macroeconomic conditions, which may change over time, (2) time-invariant individual characteristics, such as innate ability, which makes different subpopulations incomparable, (3) time-varying unobservable individual characteristics, such as the development of new skills, which may change the probability of participation across time. If all of these confounding factors are present, then the effect of job training on earnings cannot be identified for any subpopulation. For instance, the average change of earnings across time for subpopulation $X_i = (0, 1)$ is partly due to the change in job training status and partly due to changing macroeconomic conditions and time-varying ability.

Now if one assumes that macroeconomic conditions are the same across both time periods and ability was only time-invariant, then one can identify the APE of job training for the subpopulations that participate in the program using average within-group changes across time, specifically, $E[Y_{i2} - Y_{i1}|X_i = (0, 1)]$ and $E[Y_{i1} - Y_{i2}|X_i = (1, 0)]$.

However, if macroeconomic conditions were changing over time and affect individuals' decision to participate in the job training program, within-group changes over time can no longer be interpreted as the APE of the job training program. The change over time will be confounded by the changing macroeconomic climate and are only partly due to changes in job training status. In this setup, one could examine within-period changes across groups. If one assumes that subpopulations $X_i = (0, 1)$ and $X_i = (1, 0)$ have the same unobservable tendencies, e.g. they have the same proportion of high-types to low-types, then comparing individuals with $X_i = (0, 1)$ and $X_i = (1, 0)$ within the same period would not confound our effect of interest. In this case, $\{E[Y_{i1}|X_i = (1, 0)] - E[Y_{i1}|X_i = (0, 1)]\}$ and $\{E[Y_{i2}|X_i = (0, 1)] - E[Y_{i2}|X_i = (1, 0)]\}$ identify the effect of job training for the two subpopulations in the first and second period, respectively. Note that in the presence of time-varying unobservables, the APE is different in each period.

Finally, another possible strategy for within-period identification is assuming that job training status in the second period is random conditional on the first. This implies selection on observables, as introduced in Heckman and Robb (1985), in the second time period conditional on job training status in the first time period. In this case, one can

identify the effect of job training on earnings from $\{E[Y_{i2}|X_i = (0,1)] - E[Y_{i2}|X_i = (0,0)]\}$. One of the key points of the paper at hand is that neither of the above identification strategies is justified *a priori*. However, they all have clear testable implications. Hence, this paper proposes strategies to test for these identifying assumptions. We first proceed to a formal discussion of the identification problem.

3 Nonparametric Identification of APEs

3.1 DGP and Definitions

We begin with a general DGP exhibiting arbitrary individual and time heterogeneity, formally defined below. Let Y_{it} be the outcome variable of interest with support $\mathcal{Y} \subseteq \mathbb{R}$. For $i = 1, 2, ..., n$, $t = 1, 2$,

$$Y_{it} = \xi_t(X_{it}, \mathcal{A}_i, \mathcal{U}_{it}) \tag{2}$$

where X_{it} is a $d \times 1$ vector of discrete regressors, $\mathcal{A}_i$ denotes unobservable individual-specific, time-invariant factors, and $\mathcal{U}_{it}$ denotes individual-specific, time-varying unobservables. $\{Y_{it}, X_{it}\}$ are observables, whereas $\{\mathcal{A}_i, \mathcal{U}_{it}\}$ are unobservables that may confound our effect of interest. For simplicity, we assume that $\mathcal{A}_i$ and $\mathcal{U}_{it}$ are scalar real-valued random variables for $t = 1, 2$. However, the results would hold for any finite-dimensional vectors $\mathcal{A}_i$ and $\mathcal{U}_{it}$.[10] Calligraphic letters are used to distinguish unobservable from observable factors. ξ_t is unknown and is referred to as the structural function in the sense that $Y_{it}(x) = \xi_t(x, \mathcal{A}_i, \mathcal{U}_{it})$.

Notation. Let $\mathcal{X}$ denote the support of X_{it}. x and x' denote elements in $\mathcal{X}$. Now $X_i \equiv (X_{i1}, X_{i2}, ..., X_{iT})$, a $d \times T$ matrix with support $\mathcal{X}^T \equiv \times_{t=1}^{T} \mathcal{X}$. $\underline{x}$ and $\underline{x}'$ denote elements in $\mathcal{X}^T$. Note that $\underline{x} = (\underline{x}_1, \underline{x}_2, ..., \underline{x}_T)$, hence $\underline{x}_t$ denotes the t^{th} column of $\underline{x}$. For random variables, W_{it} and Z_i, let $F_{W_{it}|Z_i}(.|.)$ denote the conditional distribution function of W_{it} given Z_i in period t. For a set S, $|S|$ of a set denotes the cardinality of S.

Assumption 3.1 *(General DGP)*

[10] The support of $\mathcal{A}_i$ and $\mathcal{U}_{it}$ could also be different from $\mathbb{R}$.

(i) $Y_{it} = \xi_t(X_{it}, \mathcal{A}_i, \mathcal{U}_{it})$, *where* $\xi_t : \mathcal{X} \times \mathbb{R}^2 \mapsto \mathcal{Y}$, *where* $\mathcal{Y} \subseteq \mathbb{R}$,

(ii) $\mathcal{X}$ *is finite,* $|\mathcal{X}| = K$,

(iii) $E[Y_{it}] < \infty$ *for all* $t = 1, 2, .., T$,

(iv) $P(X_i = \underline{x}) > 0$ *for all* $\underline{x} \in \mathcal{X}^T$.

The main content of the above assumption is the finite support of X_{it} in (ii).[11] Since $\mathcal{X}$ is finite and T is fixed, $\mathcal{X}^T$ is also a finite set. Assumption 3.1 (i) may be thought of as a 'correct specification' assumption. However, it is important to note that the choice of variables to include in X_{it} is so far not restrictive, since the assumption allows for an arbitrary, time-varying structural relationship and there are no assumptions on the distribution of unobservables.[12] It is also important to note that $\mathcal{A}_i$ and $\mathcal{U}_{it}$ may be any finite-dimensional vector, we assume that they are scalar real-valued random variables for simplicity. Assumption 3.1 (iii) and (iv) are regularity conditions that ensure that the APE exists for all elements in the support of X_i, which simplifies our analysis.

The general DGP does not impose any further restrictions on the structural functions ξ_t or the conditional distributions of unobservables $F_{\mathcal{A}_i,\mathcal{U}_{it}|X_i} = F_{\mathcal{U}_{it}|X_i,\mathcal{A}_i} F_{\mathcal{A}_i|X_i}$.[13] Now we formally define what we mean by a nonstationary structural function, arbitrary individual and time heterogeneity.

Definition 1 *(Nonstationary Structural Function)* $\xi_t(.)$ *may vary for* $t = 1, 2, ..., T$.

Definition 2 *(Arbitrary Individual Heterogeneity) A DGP exhibits arbitrary individual heterogeneity, if* $F_{\mathcal{A}_i|X_i}(.|.)$ *is unrestricted.*

The above is the fundamental definition of a 'fixed effect'. Arellano (2003) shows how the linear fixed effects estimator is equivalent to the maximum likelihood estimator where the distribution of individual heterogeneity is unrestricted. Fixed effects logit and Poisson are examples of nonlinear parametric models where the presence of a sufficient statistic for

[11] It is important to note that the identification component of this paper applies to a discrete change of continuous variables as well.

[12] Once we impose identifying assumptions, the choice of X_{it} will be very important.

[13] Depending on the nature of Y_{it}, *a priori* bounds may be given as assumed in CFHN2010. It is important to note however that, unlike the case of set-identification, *a priori* bounds are irrelevant for the discussion of point-identification.

individual heterogeneity allows one to identify parameters of interest without assumptions on the distribution of individual heterogeneity.

Definition 3 *(Arbitrary Time Heterogeneity) A DGP exhibits arbitrary time heterogeneity, when* $F_{\mathcal{U}_{it}|X_i,\mathcal{A}_i}(.|.,.)$ *may vary for* $t = 1, 2, ..., T$.

The key content in the above definition is that the distribution of time-varying unobservables may change over time. However, the definition does not impose that $F_{\mathcal{U}_{it}|X_i,\mathcal{A}_i}$ is unrestricted. This is a key distinction between the definition of arbitrary individual and time heterogeneity.

We will also use the terminology of movers and stayers introduced in Chamberlain (1982) to refer to subpopulations that change the regressor vector, X_{it}, over time and those who do not, respectively.[14] The following definition uses the realizations of X_i to define a subpopulation.

Definition 4 *(Subpopulation) A subpopulation is defined by its realization of* $X_i = \underline{x}$, *where* $\underline{x} \in \mathcal{X}^T$.

Since $|\mathcal{X}^T|$ is finite, we have finitely many subpopulations. It is important to note that each subpopulation, $\underline{x} \in \mathcal{X}^T$, is characterized by having the same distribution of unobservables for all time periods, i.e. $F_{\mathcal{A}_i,\mathcal{U}_{it}|X_i}(.|\underline{x})$. Together with the structural function, this yields a distribution of the outcome variable for each subpopulation. Hence, we can think of the subpopulation as infinitely many realizations from the same distribution. This is not to be confused with individuals in the same subpopulation having the same 'fixed effect'.[15] Back to our job training example, individuals in a subpopulation have the same distribution of ability, this does not mean that they all have the same ability or that they all respond in the same way to macroeconomic shocks. But rather that they have the same likelihood of being high-achieving or fast-learners.

[14] For the latter subpopulations, all columns of X_i are equal, i.e. $X_{i1} = ... = X_{iT}$ as defined in Chamberlain (1982).

[15] The case here is really one where correlated random effects and fixed effects are equivalent. Correlated random effects are approaches that integrate over the distribution of individual effects, whereas fixed effects treat them as fixed in repeated sampling. In the linear model, fixed effects and correlated random effects are equivalent under certain assumptions, as shown in Mundlak (1978).

3.2 Object of Interest and the Quasi-differencing Approach

Now we would like to learn about the average effect of a discrete regressor X on Y for a particular subpopulation $\underline{x}$. The problem is that there is both individual and time heterogeneity that may confound our effect if we simply look at the change in Y as X changes over time. Formally, our object of interest is the APE of changing X_{it} from x to x', $x \neq x'$, for subpopulation $X_i = \underline{x}$, given as follows

$$\beta_t(x \to x'|X_i = \underline{x}) \quad = \quad E[Y_{it}^{x'}|X_i = \underline{x}] - E[Y_{it}^{x}|X_i = \underline{x}],$$

where $Y_{it}^{x} = \xi_t(x, a, u)$ is the counterfactual notation, i.e. Y_{it}^{x} is the outcome variable if the regressor vector was fixed at x. We distinguish between x and x' and the columns of $\underline{x}$, since they may not be equal. For instance, if we are interested in the effect of obtaining a master degree (x') on top of a bachelor degree (x), then for individuals that only have a high school degree, i.e. $\underline{x} = (12, ..., 12)$, $\underline{x}_t \neq x$ and $\underline{x}_t \neq x'$ for all $t = 1, 2, ..., T$.[16]
Note that the APE is indexed by the time period t and the subpopulation $\underline{x}$. This reflects the presence of time heterogeneity (*the effect is different across time periods*) as well as individual heterogeneity (*the effect is different across subpopulations*).
Given our DGP in Assumption 3.1, we can write the APE as follows

$$\begin{aligned} & E[Y_{it}^{x'}|X_i = \underline{x}] - E[Y_{it}^{x}|X_i = \underline{x}] \\ = & \int \xi_t(x', a, u) dF_{\mathcal{A}_i, \mathcal{U}_{it}|X_i}(a, u|\underline{x}) - \int \xi_t(x, a, u) dF_{\mathcal{A}_i, \mathcal{U}_{it}|X_i}(a, u|\underline{x}) \end{aligned} \tag{3}$$

It is important to note that in (3) the only difference between the two terms is that one is evaluated at x and the other at x'. Otherwise, the structural function ξ_t and the unobservable distribution $F_{\mathcal{A}_i, \mathcal{U}_{it}|X_i}(.,.|\underline{x})$ are the same. This is the key difficulty in identifying a *ceteris paribus* effect in panel data. In an experiment, one could randomize individuals from the same subpopulation in the same period, thereby holding individual and time heterogeneity fixed. In observational settings, assumptions on $\{\xi_t, F_{\mathcal{A}_i, \mathcal{U}_{it}|X_i}\}_{t=1}^{T}$ are required to ensure that average within-group or within-period changes are taken with respect to the same structural function and unobservable distribution. This is the key

[16]Note that in this situation, identification of the APE for these subpopulations requires assumptions on individual heterogeneity in the spirit of Angrist (2004).

intuition behind quasi-differencing in general nonseparable panel data models that will be examined in the following.

3.3 Characterization of the Trade-off between Identifying Assumptions

To simplify notation, we assume $T = 2$. More specifically, we would like to identify the APE of a subpopulation (x, x')

$$\beta_t(x \to x'|X_i = (x, x')) = \int (\xi_t(x', a, u) - \xi_t(x, a, u))dF_{\mathcal{A}_i,\mathcal{U}_{it}|X_i}(a, u|(x, x')) \quad (4)$$

In observational settings, an obvious candidate for the identification of $\beta_t(x \to x'|X_i = (x, x'))$ is $E[Y_{i2} - Y_{i1}|X_i = (x, x')]$, which we can decompose as follows

$$\begin{aligned} E[Y_{i2} - Y_{i1}|X_i = (x, x')] &= E[Y_{i2} - Y_{i2}^x|X_i = (x, x')] + E[Y_{i2}^x - Y_{i1}|X_i = (x, x')] \\ &= \beta_2(x \to x'|X_i = (x, x')) + E[Y_{i2}^x - Y_{i1}|X_i = (x, x')]. \end{aligned}$$

The identification of $E[Y_{i2}^x - Y_{i1}|X_i = (x, x')]$ is necessary and sufficient for the identification of $\beta_2(x \to x'|X_i = (x, x'))$. We will refer to the former as the counterfactual trend, i.e. it is the change that would have occurred to the subpopulation (x, x') had the change from x to x' not occurred.

In the following, we will give a sufficient condition under which we can identify the counterfactual trend, $E[Y_{i2}^x - Y_{i1}|X_i = (x, x')]$, from a stayer subpopulation, (x, x), i.e.

$$E[Y_{i2}^x - Y_{i1}|X_i = (x, x')] = E[Y_{i2} - Y_{i1}|X_i = (x, x)]. \quad (5)$$

This condition will help us characterize the trade-off between various identifying assumptions. We first introduce some standard regularity conditions on the distribution of unobservables.

Assumption 3.2 *(Distribution of Unobservables)*

(i) $F_{\mathcal{A}_i,\mathcal{U}_{i1},\mathcal{U}_{i2}|X_i}(.,.,.|.)$ *admits a density,* $f_{\mathcal{A}_i,\mathcal{U}_{i1},\mathcal{U}_{i2}|X_i}(.,.,.|.)$,

(ii) $f_{\mathcal{A}_i,\mathcal{U}_{i1},\mathcal{U}_{i2}|X_i}(.,.,.|.) > 0$, $f_{\mathcal{A}_i,\mathcal{U}_{it}|X_i}(.,.|.) > 0$ *for* $t = 1, 2$, $f_{\mathcal{A}_i|X_i}(.|.) > 0$.

The following theorem gives a sufficient condition for (5).

Theorem 3.1 *Let Assumptions 3.1 and 3.2 hold,*

$$E[Y_{i2}^x - Y_{i1}|X_i = (x, x')] \quad = \quad E[Y_{i2} - Y_{i1}|X_i = (x, x)]$$

if

$$\begin{aligned} & (\xi_2(x, a, u_2) - \xi_1(x, a, u_1)) \\ \times \quad & (f_{\mathcal{A}_i, \mathcal{U}_{i1}, \mathcal{U}_{i2}|X_i}(a, u_1, u_2|(x, x')) - f_{\mathcal{A}_i, \mathcal{U}_{i1}, \mathcal{U}_{i2}|X_i}(a, u_1, u_2|(x, x))) = 0. \\ & \forall (a, u_1, u_2) \in \mathbb{R}^3 \end{aligned}$$

All proofs of Section 3 are given in Appendix A. The above condition depends on the change in the structural function and the difference between unobservable distributions, both across time and subpopulations.

We illustrate how the condition characterizes the trade-off between different identifying assumptions in the following variations on the above theorem. The first two variations reflect a situation where the researcher would like to leave individual heterogeneity unrestricted, which reflects the standard fixed-effects approach.

Variation 3.1 *(Stationary Structural Function up to Generalized Time Effect) Let Assumptions 3.1 and 3.2 hold.*

Under arbitrary individual heterogeneity and $\mathcal{U}_{i1}|X_i, \mathcal{A}_i \stackrel{d}{=} \mathcal{U}_{i2}|X_i, \mathcal{A}_i$*,*

$$E[Y_{i2}^x - Y_{i1}|X_i = (x, x')] \quad = \quad E[Y_{i2} - Y_{i1}|X_i = (x, x)]$$

if

$$\xi_t(x, a, u) \quad = \quad \xi(x, a, u) + \lambda_t(x) \; \forall (a, u) \in \mathbb{R}^2$$

In the above variation on Theorem 3.1, time homogeneity, i.e. $\mathcal{U}_{i1}|X_i, \mathcal{A}_i \stackrel{d}{=} \mathcal{U}_{i2}|X_i, \mathcal{A}_i$, is maintained. If individual heterogeneity is left unrestricted, then the structural function has to be decomposed into a stationary and nonstationary component, $\xi(x, a, u)$ and $\lambda_t(x)$, respectively. The nonstationary component may only depend on observable regressors. We

will refer to this as stationary in unobservables. Hence, in the quasi-differencing approach, time homogeneity alone is not sufficient. Hoderlein and White (2009) and CFHN2010 assume both time homogeneity and a stationary structural function without time effects to achieve identification of the APE for discrete and continuous regressors, respectively.[17] Note that for continuous regressors, conditional independence restrictions are required to ensure that the marginal change in the regressor vector does not change the distribution of the unobservables, as in Altonji and Matzkin (2005) and Hoderlein and White (2009).

Variation 3.2 *(Time Homogeneity) Let Assumptions 3.1 and 3.2 hold.*
Under arbitrary individual heterogeneity and $\xi_t(x,a,u) = \xi(x,a,u) + \lambda_t(x)$,

$$E[Y_{i2}^x - Y_{i1}|X_i = (x,x')] \quad = \quad E[Y_{i2} - Y_{i1}|X_i = (x,x)]$$

if $\forall \underline{x} \in \{(x,x),(x,x')\}$

$$f_{\mathcal{U}_{i2}|X_i,\mathcal{A}_i}(u_2|\underline{x},a) \quad = \quad f_{\mathcal{U}_{i1}|X_i,\mathcal{A}_i}(u_1|\underline{x},a) \;\; \forall (a,u_1,u_2) \in \mathbb{R}^2$$

In Variation 3.2, the structural function is assumed to be stationary up to a generalized time effect. In this case, time homogeneity has to be assumed. Hence, the last two variations show that the stationarity of the structural function and time homogeneity are both required to identify the APE using within-group changes across time. This is not at all surprising for the quasi-differencing approach, since we simply take average within-group changes and remain agnostic about the functional form of the structural function and the distribution of unobservables. Thus, without information about these objects, we cannot hope to use time homogeneity without the stationarity of the structural function in unobservables and vice versa.

Variations 3.1 and 3.2 characterize a situation where identification would be achieved within-group, since individual heterogeneity is unrestricted. Section 3.4.1 generalizes the setup here where subpopulations other than (x,x) can be used to identify the generalized time effect.

So far we have seen that restrictions on time heterogeneity and the stationarity of the

[17] With an additional restriction, CFHN2010 can allow for nonstochastic location and scale time effects, i.e. $Y_{it} = \xi(X_{it},\mathcal{A}_i,\mathcal{U}_{it})\sigma_t + \lambda_t$.

structural function are required if we would like individual heterogeneity to be unrestricted. In this setup, we use average within-group changes across time to identify the APE for a subpopulation. Now we move to the case where we would like to have unrestricted time heterogeneity. Here, we will see that we have to impose restrictions on individual heterogeneity.

Variation 3.3 *(Individual Homogeneity) Given Assumptions 3.1, 3.2.*
Under arbitrary time heterogeneity

$$E[Y_{i2}^{x} - Y_{i1}|X_i = (x, x')] \quad = \quad E[Y_{i2} - Y_{i1}|X_i = (x, x)]$$

if

$$f_{\mathcal{A}_i,\mathcal{U}_{i1},\mathcal{U}_{i2}|X_i}(a, u_1, u_2|(x, x')) \quad = \quad f_{\mathcal{A}_i,\mathcal{U}_{i1},\mathcal{U}_{i2}|X_i}(a, u_1, u_2|(x, x)) \; \forall (a, u_1, u_2) \in \mathbb{R}^3$$

The above condition implies that the subpopulations (x, x') and (x, x) are homogeneous, i.e. they have the same distribution of unobservables.[18] This allows us to identify the APE using average within-period changes across subpopulations. Specifically,

$$\begin{aligned}
\beta_2(x \to x'|X_i = (x, x')) &= \int (\xi_2(x', a, u) - \xi_2(x, a, u)) dF_{\mathcal{A}_i,\mathcal{U}_{i2}|X_i}(a, u|(x, x')) \\
&= \int \xi_2(x', a, u) dF_{\mathcal{A}_i,\mathcal{U}_{i2}|X_i}(a, u|(x, x')) - \int \xi_2(x, a, u) dF_{\mathcal{A}_i,\mathcal{U}_{i2}|X_i}(a, u|(x, x)) \\
&= E[Y_{i2}|X_i = (x, x')] - E[Y_{i2}|X_i = (x, x)],
\end{aligned}$$

where the penultimate equality follows from the individual homogeneity assumption. Note that in this setup, we are using the information obtained from the first time period, i.e. X_{i1}, to identify the APE in the second time period, which is a cross-section of all individuals.[19] Section 3.4.2 generalizes this type of homogeneity assumptions to allow subpopulations other than (x, x) as a control group for (x, x'). Angrist (2004) discusses the use of this type of homogeneity assumptions to relate the local average treatment effect (LATE) to the average treatment effect (ATE) for instrumental variables methods in the

[18] It is important to note here that in the above case, the result holds because $E[Y_{i2}^{x}|X_i = (x, x')] = E[Y_{i2}|X_i = (x, x)]$ and $E[Y_{i1}|X_i = (x, x')] = E[Y_{i1}|X_i = (x, x)]$.

[19] Recall that microeconomic panel data models are also called cross-sectional time series. For each time period, we observe a cross-section.

cross-sectional setup.

So far we have characterized the obvious trade-off between individual and time heterogeneity. Since we only have two dimensions that we are differencing over, we cannot leave both unrestricted. In the next variation on Theorem 3.1, the setup assumes that we would like to leave both individual and time heterogeneity unrestricted. The result shows that separability and conditional independence assumptions are required to aid identification. We first introduce some more regularity conditions.

Assumption 3.3 *(Smoothness and Dominating Function)*

(i) $\partial\xi(x,a,u)/\partial a$, $\partial^2\xi(x,a,u)/\partial a\partial u_t$, $f_{\mathcal{U}_{it}|X_i,\mathcal{A}_i}(u|\underline{x},a)$, $\partial f_{\mathcal{U}_{it}|X_i,\mathcal{A}_i}(u|\underline{x},a)/\partial a$ *exist and are continuous for all* a,u, $x\in\mathcal{X}$, $\underline{x}\in\mathcal{X}^T$ *and* $t=1,2,...,T$,

(ii) $|\xi(x,a,u_t)f_{\mathcal{U}_{it}|X_i,\mathcal{A}_i}(u|\underline{x},a)|\leq g(x,u,\underline{x})$, *where* $E[|g(x,\mathcal{U}_{it},X_i)||X_i=\underline{x}]<\infty\ \forall\underline{x}\in\mathcal{X}^T$.

The characterization of separability of a function in two variables is characterized by the cross-partial derivative being zero.[20] Hence, Assumption 3.3 (i) ensures that sufficient smoothness conditions hold. Assumption 3.3 (ii) ensures that we can apply the dominating convergence theorem.

Variation 3.4 *(Separability and Independence) Let Assumptions 3.1, 3.2 and 3.3 hold. Under unrestricted individual and time heterogeneity, and* $\xi_t(x,a,u)=\xi(x,a,u)$, $t=1,2$,

$$E[Y_{i2}^x-Y_{i1}|X_i=(x,x')] \quad = \quad E[Y_{i2}-Y_{i1}|X_i=(x,x)]$$

if

$$\begin{aligned}\partial^2\xi(x,a,u)/\partial a\partial u &= 0\ \forall(a,u)\in\mathbb{R}^2, && \textit{(Separability)}\\ (\mathcal{U}_{i1},\mathcal{U}_{i2}) &\perp (X_i,\mathcal{A}_i).\ \textit{for } t=1,2 && \textit{(Independence)}\end{aligned}$$

There are two interesting findings in the above result. The first is that separability aids identification in this setup if it is coupled with independence. The separability of $\mathcal{A}_i$ and $\mathcal{U}_{it}$ alone is not sufficient if they are dependent. We can think of independence as

[20] For instance, $m(a,u)=a^2+u^2$, then $\partial m(a,u)/\partial a=2a$ and $\partial m(a,u)/\partial u=2u$. Thus, $\partial^2 m(a,u)/\partial a\partial u=0$.

"stochastic separability".[21] Hence, both functional-form and "stochastic separability" are required. The other interesting issue to note here is that even though we have arbitrary individual and time heterogeneity, i.e. $F_{\mathcal{A}_i|X_i}$ is unrestricted, and $F_{\mathcal{U}_{it}|X_i,\mathcal{A}_i}$ is allowed to change over time, the idiosyncratic shocks, $\mathcal{U}_{it}$ are independent of individual heterogeneity and observables, since $F_{\mathcal{U}_{it}|X_i,\mathcal{A}_i} = F_{\mathcal{U}_{it}}$.

Relation to the Classical Identification Literature. In the classical identification literature, stronger assumptions are imposed to identify all structural objects, whereas the quasi-differencing is much more parsimonious. In the following, we will discuss the relative merits of both approaches with special attention given to the model in Evdokimov (2010). It is given by the following

$$Y_{it} \quad = \quad m(X_{it}, \mathcal{A}_i) + \mathcal{U}_{it}. \tag{6}$$

Under the separability of $\mathcal{U}_{it}$ in the structural function, the conditional independence $F_{\mathcal{U}_{it}|X_i,\mathcal{A}_i,\mathcal{U}_{i(t)}} = F_{\mathcal{U}_{it}|X_{it}}$,[22] monotonicity of $m(x,a)$ in a and regularity conditions, Evdokimov (2010) shows the identification of $F_{\mathcal{U}_{it}|X_{it}}$ for $t = 1, 2$, $F_{\mathcal{A}_i|X_i}$ and $m(x,a)$, $\forall x, a$. Under the above model, we can decompose the average within-group change for subpopulation (x, x') as follows,

$$\begin{aligned} E[Y_{i2} - Y_{i1}|X_i = (x,x')] &= E[m(x', \mathcal{A}_i) - m(x, \mathcal{A}_i)|X_i = (x,x')] \\ &+ E[\mathcal{U}_{i2}|X_{i2} = x'] - E[\mathcal{U}_{i1}|X_{i1} = x] \\ &= \beta(x \to x'|X_i = (x,x')) + \Delta. \end{aligned} \tag{7}$$

Since $F_{\mathcal{U}_{it}|X_{it}}$ is identified for $t = 1, 2$, we can identify Δ. Hence, the above identification strategy allows us to identify our object of interest in the presence of time heterogeneity, while allowing individual heterogeneity to be unrestricted.

The quasi-differencing approach gives two strategies to identify the same objects. First, Variation 3.1 and 3.2 show that neither time homogeneity nor stationarity of the structural

[21] This term is used here to convey the intuition and is not related to the definition of separability of stochastic processes.

[22] $\mathcal{U}_{i(t)}$ is the idiosyncratic shock for the other period $\tau \neq t$.

function are sufficient on their own. Hence, the following model

$$\begin{aligned} Y_{it} &= \xi(X_{it}, \mathcal{A}_i, \mathcal{U}_{it}) + \lambda_t(X_{it}), \;\; t = 1, 2 \\ \mathcal{U}_{i1}|X_i, \mathcal{A}_i &\overset{d}{=} \mathcal{U}_{i2}|X_i, \mathcal{A}_i, \end{aligned} \tag{8}$$

where $\lambda_1(x) = 0$ for all $x \in \mathcal{X}$. Hence, $\xi_1(x, a, u) = \xi(x, a, u)$ and $\xi_2(x, a, u) = \xi(x, a, u) + \lambda_2(x)$. Now the APE of moving from x to x' for subpopulation (x, x') in the period t is given by

$$\begin{aligned} \beta_t(x \to x'|X_i = (x, x')) &= \int (\xi_t(x', a, u) - \xi_t(x, a, u)) dF_{\mathcal{A}_i, \mathcal{U}_{it}|X_i}(a, u|(x, x')) \\ &\overset{(8)}{=} \int (\xi_t(x', a, u) - \xi_t(x, a, u)) dF_{\mathcal{A}_i, \mathcal{U}_{i1}|X_i}(a, u|(x, x')) \end{aligned} \tag{9}$$

Note that using average within-group changes

$$\begin{aligned} E[Y_{i2} - Y_{i1}|X_i = (x, x')] &= E[Y_{i2} - Y_{i2}^x|X_i = (x, x')] + E[Y_{i2}^x - Y_{i1}|X_i = (x, x')] \\ &= \int (\xi_2(x', a, u) - \xi_2(x, a, u)) dF_{\mathcal{A}_i, \mathcal{U}_{i1}|X_i}(a, u|(x, x')) \\ &+ \int (\xi_2(x, a, u) - \xi_1(x, a, u)) dF_{\mathcal{A}_i, \mathcal{U}_{i1}|X_i}(a, u|(x, x')) \\ &= \beta_2(x \to x'|X_i = (x, x')) + \lambda_2(x) \end{aligned} \tag{10}$$

The APE is indexed by the time period, since the structural function is not stationary in the observables, $\xi_2(x, a, u) = \xi_1(x, a, u) + \lambda_2(x)$.[23] Now to identify the APE in (10), we have to identify $\lambda_2(x)$, the counterfactual trend. Now noting that

$$\begin{aligned} E[Y_{i2} - Y_{i1}|X_i = (x, x)] &= \int (\xi_2(x, a, u) - \xi_1(x, a, u)) dF_{\mathcal{A}_i, \mathcal{U}_{i1}|X_i}(a, u|(x, x)) \\ &= \int \lambda_2(x) dF_{\mathcal{A}_i, \mathcal{U}_{i1}|X_i}(a, u|(x, x)) = \lambda_2(x). \end{aligned} \tag{11}$$

The last equality illustrates why identification of the counterfactual trend would not be achieved if the nonstationary component of the structural function depended on the unobservables as pointed out in Variation 3.1. Plugging (11) into (10) identifies the APE.

Note that the setup of Evdokimov (2010) may also accommodate generalized time effects

[23] This implies that $\xi_2(x', a, u) - \xi_1(x, a, u) = \xi_1(x', a, u) + \lambda(x') - \xi_1(x, a, u) - \lambda(x)$, which is equal to $\xi_1(x', a, u) - \xi_1(x, a, u)$ iff $\lambda_t(x') = \lambda_t(x)$. For instance, if $\lambda_t(x) = \lambda$, then this condition would be fulfilled and the APE would be constant across time.

as in $\lambda_t(X_{it})$. The key difference between the two approaches is that with the identification of the entire distribution of $\mathcal{U}_{it}|X_{it}$, Evdokimov (2010) can allow for time heterogeneity, whereas the quasi-differencing approach used here does not allow for time heterogeneity. Since the structural functions in (6) and (8) are different, it may be hard to compare the two models. On the other hand, (6) is very similar to the setup in Variation 3.4, which is given by

$$\begin{aligned} Y_{it} &= \mu(X_{it}, \mathcal{A}_i) + \mathcal{U}_{it} \\ \mathcal{U}_{it}|X_i, \mathcal{A}_i &\overset{d}{=} \mathcal{U}_{it}, \quad t = 1, 2 \end{aligned} \tag{12}$$

$$\begin{aligned} E[Y_{i2} - Y_{i1}|X_i = (x, x')] &= \int (\mu(x', a) + u_2 - (\mu(x, a) + u_1)) dF_{\mathcal{A}_i, \mathcal{U}_{i1}, \mathcal{U}_{i2}|X_i}(a, u_1, u_2|(x, x')) \\ &= \int (\mu(x', a) - \mu(x, a)) dF_{\mathcal{A}_i|X_i}(a|(x, x')) \\ &+ E[\mathcal{U}_{i2}|X_i = (x, x')] - E[\mathcal{U}_{i1}|X_i = (x, x')] \\ &\overset{(12)}{=} \beta(x \to x'|X_i = (x, x')) + E[\mathcal{U}_{i2}] - E[\mathcal{U}_{i1}] \end{aligned} \tag{13}$$

Here, the counterfactual trend is given by $\{E[\mathcal{U}_{i2}] - E[\mathcal{U}_{i1}]\}$. Since the distribution of $\mathcal{U}_{it}$ is independent of X_i and $\mathcal{A}_i$, we can use any stayer subpopulation to identify the counterfactual trend.

$$E[Y_{i2} - Y_{i1}|X_i = (x, x)] = E[\mathcal{U}_{i2} - \mathcal{U}_{i1}|X_i = (x, x)] \overset{(12)}{=} E[\mathcal{U}_{i2}] - E[\mathcal{U}_{i1}]. \tag{14}$$

Plugging (14) into (13) identifies the APE. In Evdokimov (2010), $E[\mathcal{U}_{it}|X_{it}] \neq E[\mathcal{U}_{it}]$. If this were true, then we could not identify the APE using the quasi-differencing approach. It is important to point out that the classical identification results in Evdokimov (2010) require stronger assumptions than the quasi-differencing approach. Hence, it is hard to compare the two approaches. The above examples are simply meant to illustrate the advantages of imposing stronger assumptions and identifying all structural objects.

Relation to the Classical Difference-in-Difference Model. Let $X_{it} \in \{0, 1\}$ and $T = 2$. Now consider the following two linear models. Model 1 follows the setup in

Variations 3.1 and 3.2,

$$\begin{aligned} Y_{it} &= \beta X_{it} + \mathcal{A}_i + \lambda_t + \mathcal{U}_{it}, \quad t = 1, 2 \\ \mathcal{U}_{i2}|X_i, \mathcal{A}_i &\overset{d}{=} \mathcal{U}_{i1}|X_i, \mathcal{A}_i. \end{aligned} \tag{15}$$

Model 2 follows the setup in Variation 3.4,

$$\begin{aligned} Y_{it} &= \beta X_{it} + \mathcal{A}_i + \mathcal{U}_{it} \\ \mathcal{U}_{it}|X_i, \mathcal{A}_i &\overset{d}{=} \mathcal{U}_{it}, \quad t = 1, 2. \end{aligned} \tag{16}$$

Under both models, β is identified by the probability limit of the difference-in-difference estimator $\hat{\beta} \equiv \sum_{i=1}^n \Delta Y_i 1\{X_i = (0,1)\} / \sum_{i=1}^n 1\{X_i = (0,1)\} - \sum_{i=1}^n \Delta Y_i 1\{X_i = (0,0)\} / \sum_{i=1}^n 1\{X_i = (0,0)\}$.

For Model 1,

$$\begin{aligned} \hat{\beta} &\xrightarrow{p} E[Y_{i2} - Y_{i1}|X_i = (0,1)] - E[Y_{i2} - Y_{i1}|X_i = (0,0)] \\ &= \beta + E[\mathcal{A}_i + \lambda_2 + \mathcal{U}_{i2} - (\mathcal{A}_i + \lambda_1 + \mathcal{U}_{i1})|X_i = (0,1)] \\ &- E[\mathcal{A}_i + \lambda_2 + \mathcal{U}_{i2} - (\mathcal{A}_i + \lambda_1 + \mathcal{U}_{i1})|X_i = (0,0)] \\ &\overset{(15)}{=} \beta + E[\mathcal{U}_{i1} - \mathcal{U}_{i1}|X_i = (0,1)] - E[\mathcal{U}_{i1} - \mathcal{U}_{i1}|X_i = (0,0)] \\ &= \beta \end{aligned} \tag{17}$$

Note that in the above, idiosyncratic shocks may be correlated with the regressors as well as unobservable individual heterogeneity in arbitrary ways. However, time homogeneity has to hold, i.e. the conditional distribution of idiosyncratic shocks has to be the same across time. Furthermore, the structural relationship has to be stationary up to a non-stochastic time effect. Thus, the generalization of Model 1 to a nonseparable model yields Model 1', given by

$$\begin{aligned} Y_{it} &= \xi(X_{it}, \mathcal{A}_i, \mathcal{U}_{it}) + \lambda_t, \ \ t = 1, 2 \\ \mathcal{U}_{i2}|X_i, \mathcal{A}_i &\overset{d}{=} \mathcal{U}_{i1}|X_i, \mathcal{A}_i. \end{aligned} \tag{18}$$

The above is a generalization of Athey and Imbens (2006) for the panel context, where a nonstochastic time effect is allowed for.

For Model 2, on the other hand,

$$\begin{aligned}
\hat{\beta} &\xrightarrow{p} E[Y_{i2} - Y_{i1}|X_i = (0,1)] - E[Y_{i2} - Y_{i1}|X_i = (0,0)] \\
&= \beta + E[\mathcal{U}_{i2} - \mathcal{U}_{i1}|X_i = (0,1)] - E[\mathcal{U}_{i2} - \mathcal{U}_{i1}|X_i = (0,0)] \\
&\overset{(16)}{=} \beta + E[\mathcal{U}_{i2} - \mathcal{U}_{i1}] - E[\mathcal{U}_{i2} - \mathcal{U}_{i1}] \\
&= \beta. \qquad (19)
\end{aligned}$$

The intuition here is that, even though idiosyncratic shocks are heterogeneously distributed across time, they are independent of regressors and individual heterogeneity. Hence, counterfactual trends for mover subpopulations, (x, x') can be identified from stayer subpopulations, (x, x). Generalizing this model to the nonseparable setup yields Model 2', given by

$$\begin{aligned}
Y_{it} &= \xi(X_{it}, \mathcal{A}_i, \mathcal{U}_{it}), \\
\mathcal{U}_{it}|X_i, \mathcal{A}_i &\overset{d}{=} \mathcal{U}_{it}.\ t = 1, 2 \qquad (20)
\end{aligned}$$

Recall that the parallel-trends assumption in the linear model implies that $Cov(\mathcal{U}_{it}, \mathcal{A}_i) = 0$ and $Cov(\mathcal{U}_{it}, X_{it}) = 0$ for $t = 1, 2$. Hence, the above may be viewed as the nonseparable version of the parallel-trends assumption.[24]

3.4 Menu of Identifying Assumptions

The previous section gives directions to some new identification strategies. For within-group identification, we use time homogeneity and the stationarity of the structural function in unobservables. For instance, the structural function may be given as follows

$$Y_{it} = \xi(X_{it}, \mathcal{A}_i, \mathcal{U}_{it}) + \lambda_t(X_{it}). \qquad (21)$$

[24] In the empirical literature, the parallel-trends assumption is tested by look at pre-trial years and ensuring that both control and treatment groups follow the same trends in average within-group changes across time. It is important to note that if only changes are considered, both Model 1' and Model 2' would yield the same 'parallel-trends' prediction. However, they have different testable implications if we look at the distribution of unobservables as a whole. This issue will be left for future work.

For within-period identification, restrictions on individual heterogeneity, such as $F_{\mathcal{A}_i,\mathcal{U}_{it}|X_i}(.,.|(x,x)) = F_{\mathcal{A}_i,\mathcal{U}_{it}|X_i}(.,.|(x,x'))$, allow us to identify the effect of interest in period t in the presence of time heterogeneity.

In the following section, we will give generalizations of (21) as well as other options for within-period identification.

3.4.1 Within-Group Identification: Restrictions on Structural Function and Time Heterogeneity

Time homogeneity was proposed in CFHN2010, where the structural function is assumed to be stationary in both observables and unobservables. In this paper, we allow the structural function to be nonstationary in observables, however we maintain its stationarnity in unobservables. The following theorem gives conditions under which one can identify the APE of a subpopulation using average within-group changes across time, when we allow for generalized time effects in the structural function.

$$Y_{it} \quad = \quad \xi(X_{it}, \mathcal{A}_i, \mathcal{U}_{it}) + \lambda_t(h(X_i)) \tag{22}$$

where $h : \mathcal{X}^T \mapsto \mathbb{R}$.[25] The following theorem gives conditions under which one can identify the APE given the above structural relationship. Let $\underline{x}_t$ denote the t^{th} column of $\underline{x}$.

Theorem 3.2 *(Within-Group Identification)*

Given Assumption 3.1, if the following conditions are satisfied for some $t, \tau \in \{1, 2, ...T\}$, $\tau \neq t$,

(i) $F_{\mathcal{U}_{it}|X_i,\mathcal{A}_i}(.|.,.) = F_{\mathcal{U}_{i\tau}|X_i,\mathcal{A}_i}(.|.,.)$,

(ii) $\xi_\tau(x,a,u) = \xi_t(x,a,u) + \lambda_\tau(h(\underline{x}))$, $\forall (x,a,u) \in \mathcal{X} \times \mathbb{R}^2$

(iii) $h(\underline{x}) = h(\underline{x}^c)$ *and* $\underline{x}^c_t = \underline{x}^c_\tau$,

then

$$\beta_t(\underline{x}_t \to \underline{x}_\tau | X_i = \underline{x}) \quad = \quad E[Y_{i\tau} - Y_{it} | X_i = \underline{x}] - E[Y_{i\tau} - Y_{it} | X_i = \underline{x}^c]$$

[25] Note that this mapping is not onto. $\mathcal{X}^T$ is a finite set, whereas $\mathbb{R}$ is infinite.

The above theorem shows that in order to allow for a time effect that depends on X_i, there has to be a restriction $h(X_i)$ that allows one to identify $\lambda_\tau(h(\underline{x}))$ from another subpopulation, a control group, $\underline{x}^c$, this is implied by condition (iii) in the above theorem. To illustrate this, given (i) and (ii),

$$\begin{aligned} E[Y_{i\tau} - Y_{it}|X_i = \underline{x}] &= \int (\xi_\tau(\underline{x}_\tau, a, u) - \xi_t(\underline{x}_t, a, u))dF_{\mathcal{A}_i,\mathcal{U}_{i1}|X_i}(a, u|\underline{x}) \\ &= \beta_\tau(\underline{x}_t \to \underline{x}_\tau|X_i = \underline{x}) + \lambda_\tau(h(\underline{x})). \end{aligned}$$

Hence, the counterfactual trend here is $\lambda_\tau(h(\underline{x}))$. Now we would like to identify this counterfactual trend from a subpopulation $\underline{x}^c$, given by

$$\begin{aligned} E[Y_{i\tau} - Y_{it}|X_i = \underline{x}^c] &= \int (\xi_\tau(\underline{x}^c_\tau, a, u) - \xi_t(\underline{x}^c_t, a, u))dF_{\mathcal{A}_i,\mathcal{U}_{i1}|X_i}(a, u|\underline{x}^c) \\ &= \int (\xi_\tau(\underline{x}^c_\tau, a, u) - \xi_t(\underline{x}^c_t, a, u))dF_{\mathcal{A}_i,\mathcal{U}_{i1}|X_i}(a, u|\underline{x}^c) \\ &\overset{(\underline{x}^c_\tau = \underline{x}^c_t)}{=} \lambda_\tau(h(\underline{x}^c)) \\ &\overset{(h(\underline{x}) = h(\underline{x}^c))}{=} \lambda_\tau(h(\underline{x})). \end{aligned}$$

The last two equalities follow from the conditions given in (iii).

Within-group identification strategies are very popular in the empirical literature, the above result can be thought of as a generalization of the fixed-effects methods, such as the difference-in-difference model discussed above. Section 4 proposes tests for the models implied by the above theorem, which may be viewed as nonparametric tests of the fixed-effects assumption.

3.4.2 Within-Period Identification: Restrictions on Individual Heterogeneity

Within-period identification solves a cross-sectional identification problem using the panel information. Traditional cross-sectional identification strategies may be applied, such as instrumental variables approaches which require completeness conditions in the nonparametric setup, which are known to be very strong.[26] What we propose here are alternative

[26] Canay, Santos and Shaikh (2011) give conditions under which no non-trivial tests of the completeness condition exist.

strategies to the classical cross-sectional identification strategies. For continuous regressors, Bester and Hansen (2009) show how average effects in a specific time period can be identified through index restrictions on the distribution of individual heterogeneity, while relying on special properties of continuous variables.[27] In many ways, within-period identification strategies presented here may be thought of as the counterpart of Bester and Hansen (2009) for discrete regressors.

Theorem 3.3 *(Within-Period Identification)*

Given Assumption 3.1, if

(i) $F_{\mathcal{A}_i,\mathcal{U}_{it}|X_i}(.,.|.) = F_{\mathcal{A}_i,\mathcal{U}_{it}|h(X_i)}(.,.|.)$,

(ii) $h(\underline{x}) = h(\underline{x}^c)$, *where* $\underline{x} \neq \underline{x}^c$,

then

$$\begin{aligned}\beta_t(\underline{x}_t^c \to \underline{x}_t|X_i = \underline{x}) &= \beta_t(\underline{x}_t^c \to \underline{x}_t|X_i = \underline{x}^c) \\ &= E[Y_{it}|X_i = \underline{x}] - E[Y_{it}|X_i = \underline{x}^c].\end{aligned}$$

Note in the above that $\beta_t(\underline{x}_t^c \to \underline{x}_t|X_i = \underline{x}) = \beta_t(\underline{x}_t^c \to \underline{x}_t|X_i = \underline{x}^c)$ by (i) and (ii). Hence, we identify the effect for the control group as well. The particular effect that we identify is determined by the realizations of X_{it} for both subpopulations in period t, $\underline{x}_t$ and $\underline{x}_t^c$, respectively.

There are two types of restrictions that may be of particular interest here, (1) exchangeability restrictions and (2) exclusion restrictions. Exchangeability restrictions, which are similar in spirit to the restrictions in Bester and Hansen (2009), were proposed in Altonji and Matzkin (2005) to relax the conditional independence assumption in the identification of average derivatives.[28] In that case, exchangeability is not sufficient for identification. However, for discrete regressors, exchangeability is sufficient. We give an empirical example where we use the exchangeability restriction to identify the effect of interest.

Example 3.1 *(Female Labor Participation: Exchangeability Restriction)*

Let Y_{it} *be a binary variable for employment status,* X_{it} *a binary variable for having a child*

[27] Bester and Hansen (2009) impose the restriction that $\mathcal{U}_{it}|X_i, \mathcal{A}_i \overset{d}{=} \mathcal{U}_{it}|X_{it}, \mathcal{A}_i$.

[28] Note that in Hoderlein and White (2009) time homogeneity alone is not sufficient for the identification of the local average structural derivative, but a conditional independence restriction is also imposed.

under 6 months of age.

$$Y_{it} \quad = \quad \xi_t(X_{it}, \mathcal{A}_i, \mathcal{U}_{it}) \tag{23}$$

The effect of child-bearing on female labor participation is a classical example of panel data models using fixed effects approach, i.e. within-group identification, whether in linear or nonlinear models, as in Heckman and MaCurdy (1980) and (1982), Hyslop (1999), and Fernandez-Val (2009). In the presence of time heterogeneity, within-group identification could not identify the APE of child-bearing on female labor participation. However, if we assume $F_{\mathcal{A}_i,\mathcal{U}_{it}|X_i} = F_{\mathcal{A}_i,\mathcal{U}_{it}|\sum_t X_{it}}$*, then within-period identification is possible. The justification behind such an assumption is that, generally speaking, individuals are more likely to select how many children to have in a given time period but they cannot fully control when their children are born. This would imply that the number of times that an individual has children under 6 months of age, i.e.* $\sum_{t=1}^{T} X_{it}$*, characterizes unobservable characteristics, as opposed to* X_i*. For* $T = 2$*, this assumption implies that subpopulations* $(0,1)$ *and* $(1,0)$ *have the same distribution of unobservables, i.e.* $F_{\mathcal{A}_i,\mathcal{U}_{it}|X_i}(.,.|(0,1)) = F_{\mathcal{A}_i,\mathcal{U}_{it}|X_i}(.,.|(1,0))$ *for* $t = 1, 2$*. Hence, they can be used to identify the counterfactual for one another. This yield the following APEs for the first and second time period.*

$$\begin{aligned}
\beta_1(0 \to 1|X_i = (0,1)) \quad &= \quad \beta_1(0 \to 1|X_i = (1,0)) = E[Y_{i1}|X_i = (1,0)] - E[Y_{i1}|X_i = (0,1)] \\
\beta_2(0 \to 1|X_i = (0,1)) \quad &= \quad \beta_2(0 \to 1|X_i = (1,0)) = E[Y_{i2}|X_i = (0,1)] - E[Y_{i2}|X_i = (1,0)].
\end{aligned}$$

Hence, within-period changes across subpopulations identify the effect of interest for individuals that switch their X_{it} *across time. Note that in the presence of time heterogeneity,* $\beta_1(0 \to 1|X_i = (0,1)) \neq \beta_2(0 \to 1|X_i = (0,1))$*. The intuition here is that in the presence of changing macroeconomic conditions, the APE in one year is generally different from another, even for the same subpopulation.*

Another class of restrictions is exclusion restrictions such as $F_{\mathcal{A}_i,\mathcal{U}_{it}|X_i}(.) = F_{\mathcal{A}_i,\mathcal{U}_{it}|X_{i(\tau)}}(.)$, where $X_{i(t)} = \{X_{i1}, .., X_{i,\tau-1}, X_{i,\tau+1}, ..., X_{iT}\}$. The distribution of unobservables is determined by observables, $X_{i(\tau)}$, hence $X_{i\tau}$ does not provide additional information about these unobservables. This assumption implies selection on observables, $X_{i(\tau)}$, as in Heck-

man and Robb (1985).

Example 3.2 *(Hypothetical Example: Family Panel Model with Exclusion Restriction) Now let us think of a panel of families and their children. Our outcome variable is income of the child, and the variable of interest is a binary variable for whether the parents helped their child with college tuition or not.*

$$Y_{ij} = \xi_j(X_{ij}, \mathcal{A}_i, \mathcal{U}_{ij}), \quad i = 1, 2, ..., n, j = 1, 2, ..., J. \tag{24}$$

For simplicity, assume that all families in the panel have exactly 2 children, i.e. J=2. In this setup, it is very important to allow for unobservable heterogeneity at the child level, since every child has his/her own abilities. Thus, a within-period approach is more suitable. In this situation, once we condition on whether parents help their first child with their college tuition, whether they help their second child contains no further information about the unobservable characteristics of the family. This justifies an exclusion restriction on the distribution of unobservables, $F_{\mathcal{A}_i,\mathcal{U}_{ij}|X_i} = F_{\mathcal{A}_i,\mathcal{U}_{ij}|X_{i1}}$, $j = 1, 2$. *As a result, we can identify the effect of interest for the second child. If we observe all subpopulations,* $(0, 0)$, $(0, 1)$, $(1, 0)$ *and* $(1, 1)$, *then we can identify the following objects:*

$$\begin{aligned}
\beta_2(0 \to 1|X_i = (0, 1)) &= \beta_2(0 \to 1|X_i = (0, 1)) = E[Y_{i2}|X_i = (0, 1)] - E[Y_{i2}|X_i = (0, 0)] \\
\beta_2(0 \to 1|X_i = (1, 0)) &= \beta_2(0 \to 0|X_i = (1, 1)) = E[Y_{i2}|X_i = (1, 1)] - E[Y_{i2}|X_i = (1, 0)],
\end{aligned}$$

which are the APE of parents' help with college tuition on the second child's income for subpopulations that did not help their first child with college tuition and the subpopulation that helped their first child with college tuition, respectively.

4 Testing Identifying Assumptions

The purpose of the previous section is to better understand how identification of a specific object is achieved and to characterize the trade-off between individual and time heterogeneity as well as assumptions on the the structural function. It also provides different assumptions that can achieve identification of the same object and may not be justified *a*

priori. Fortunately, the identification strategies proposed here have testable implications. In this section, we propose tests for these implications.

4.1 Basic Testing Problem

The identifying assumptions proposed above impose restrictions on $\{\xi_t, F_{\mathcal{A}_i,\mathcal{U}_{it}|X_i}\}_{t=1}^T$, which imply equality restrictions on the conditional distribution of the outcome variable.[29] The equality of two distributions is a well-known problem in statistics, the so-called two-sample problem, where the two samples are random and independent of each other. When testing within-group identification in the panel setting, the two samples are not independent, since both samples are realizations of the same individuals across time. For within-period identification, our statistic is a linear combination of KS and CM statistics. For both cases, bootstrap methods are proposed to obtain the critical values for the statistics.

It is important to note that the tests proposed here are not over-identification tests. They test implications of the identifying assumptions. Hence, rejecting them is clear evidence against the identifying assumptions. Two examples are given below.

Example 4.1 *(Time Homogeneity for the Scalar Binary Example)*
Let $T = 2$ and $X_{it} \in \{0, 1\}$. Assume a stationary structural function and time homogeneity, $\xi_t = \xi$ and $F_{\mathcal{A}_i,\mathcal{U}_{i1}|X_i} = F_{\mathcal{A}_i,\mathcal{U}_{i2}|X_i}$ for $t = 1, 2$. This implies the following restriction

$$F_{Y_{i1}^{\underline{x}}|X_i}(.|\underline{x}) \quad = \quad F_{Y_{i2}^{\underline{x}}|X_i}(.|\underline{x}), \quad \forall \underline{x}. \tag{25}$$

Back to our job training program, if a subpopulation does not change its job training status across time, then the distribution of its outcome variable should also not change over time. Note that this is exactly what allows us to identify the effect of job training on earnings for subpopulations that switch their job training status across time.

[29] As a result, these restrictions imply restrictions on the set of distributions, $\mathcal{D} \equiv \{F_{Y_{it}^{\underline{x}}|X_i}(.|\underline{x}) : t = 1, 2, ..., T, x \in \mathcal{X}, \underline{x} \in \mathcal{X}^T\}$, where X_{it} is fixed. Note that this set is not observable. Now let us define the set of observable distributions $\mathcal{D}^o \equiv \{F_{Y_{it}|X_i}(.|\underline{x}) : t = 1, 2, ..., T, \underline{x} \in \mathcal{X}^T\}$. If the restrictions imply that the map $\mathcal{D}^o \to \mathcal{D}$ is non-injective, then the identifying assumptions imply equality restrictions on the conditional distribution of the outcome variable.

Thus, we have the following restrictions

$$\begin{aligned} F_{Y_{i1}|X_i}(.|(0,0)) &= F_{Y_{i2}|X_i}(.|(0,0)) \\ F_{Y_{i1}|X_i}(.|(1,1)) &= F_{Y_{i2}|X_i}(.|(1,1)) \end{aligned} \tag{26}$$

However, note that the APE for subpopulations $(0,1)$ *and* $(1,0)$ *is just-identified as follows*

$$\begin{aligned} \beta_1(0 \to 1|X_i = (0,1)) &= E[Y_{i2}|X_i = (0,1)] - E[Y_{i1}|X_i = (0,1)] \\ \beta_1(0 \to 1|X_i = (1,0)) &= E[Y_{i1}|X_i = (1,0)] - E[Y_{i2}|X_i = (1,0)]. \end{aligned}$$

Thus, the tests proposed here may be implemented even if there are no over-identifying restrictions.

Example 4.2 *(Exclusion Restriction Example for the Scalar Binary Example)*
Again, let $T = 2$ *and* $X_{it} \in \{0,1\}$*. We now impose the following exclusion restriction* $F_{\mathcal{A}_i,\mathcal{U}_{it}|X_i}(.,.|\underline{x}) = F_{\mathcal{A}_i,\mathcal{U}_{it}|X_{i1}}(.,.|\underline{x}_1)$*. This assumption implies that*

$$F_{Y_{i1}^x|X_i}(.,.|(x,x')) = F_{Y_{i1}^x|X_i}(.,.|(x,x'')) \quad \forall x, x', x'' \in \mathcal{X}, x' \neq x'' \tag{27}$$

Hence, we have testable implications for the first period, where $F_{Y_{i1}|X_i}(.|(0,0)) = F_{Y_{i1}|X_i}(.|(0,1))$ *and* $F_{Y_{i1}|X_i}(.|(1,0)) = F_{Y_{i1}|X_i}(.|(1,1))$*. Here our object of interest is also just identified.*

$$\begin{aligned} \beta_2(0 \to 1|X_i = (0,1)) &= E[Y_{i2}|X_i = (0,1)] - E[Y_{i2}|X_i = (0,0)] \\ \beta_2(0 \to 1|X_i = (1,0)) &= E[Y_{i2}|X_i = (1,1)] - E[Y_{i2}|X_i = (1,0)] \end{aligned}$$

In the following, the classical two-sample problem is reviewed. Then, the validity of the bootstrap procedures for within-group and within-period identification is shown.

4.2 Review of the Classical Two-Sample Problem

Given two independent random samples of continuous variables, $\{X_i\}_{i=1}^n$ and $\{Y_i\}_{i=1}^m$, i.e. all the random variables are mutually independent, we would like to test the equality of

the two distributions from which the samples are drawn, i.e.

$$H_0 : F_X(.) = F_Y(.) \text{ versus } H_1 : F_X(.) \neq F_Y(.) \tag{28}$$

where $F_W(.)$ denotes the distribution of W. Now let $F_{W,n}(.) = \sum_{i=1}^{n} 1\{W_i \leq .\}/n$, the empirical cumulative distribution function (cdf). Now we can define the Kolmogorov-Smirnov (KS) and Cramer-von-Mises (CM) statistics as follows,

$$KS_{n\mathcal{Z}} = \sup_{z\in\mathcal{Z}} |F_{X,n}(z) - F_{Y,n}(z)| \equiv \|F_{X,n}(.) - F_{Y,m}(.)\|_{\infty,\mathcal{Z}} \tag{29}$$

and

$$CM_{n,F_Z} = \int (F_{X,n}(z) - F_{Y,n}(z))^2 dF_Z(z) \equiv \|F_{X,n}(.) - F_{Y,m}(.)\|_{2,F_Z}, \tag{30}$$

respectively. F_Z is the pooled distribution. Note that $F_Z(.) = F_X(.) = F_W(.)$ under the null. Since $F_Z(.)$ is unknown, CM_{n,F_Z} is infeasible. There is a feasible statistic equivalent to CM_{n,F_Z} that will be discussed in the following.

Under the independence of the two random samples and the continuity of the underlying pooled distribution, both statistics are pivotal, i.e. their distribution does not depend on the distributions of X and Y. The result follows from the probability integral transform theorem and is given for the one-sample KS statistic in Gibbons (1985).[30] For the two-sample CM statistic, Anderson (1962) gives the asymptotic distribution of the two-sample CM statistic. For a detailed outline of the development of the KS statistic, see Andrews (1997).

Let $N = n + m$, and $\{Z_i\}_{i=1}^{N}$ be the pooled sample of $\{X_i\}_{i=1}^{n}$ and $\{Y_i\}_{i=1}^{m}$. Under the above conditions, the aforementioned statistics are equivalent to the following

$$KS_{n,N} = \max_{1\leq i\leq N} |F_{X,n}(Z_i) - F_{Y,n}(Z_i)| \equiv \|F_{X,n}(.) - F_{Y,m}(.)\|_{\infty,N} \tag{31}$$

and

$$CM_{n,N} = \frac{nm}{N^2} \sum_{i=1}^{N} (F_{X,n}(Z_i) - F_{Y,n}(Z_i))^2 \equiv \|F_{X,n}(.) - F_{Y,m}(.)\|_{2,N}, \tag{32}$$

[30]The extension to the two-sample case is straightforward and is given in Section C.

which are based on the empirical measure.

There are two related testing problems that we are interested in here. In the following, the cross-sectional independence assumption is maintained. First, the testing of within-group identifying assumptions, which are generalizations of time homogeneity, is a paired-sample problem, which deviates from the classical two-sample problem due to the dependence between the samples. The dependence is a result of the nature of panel data, where we track the same individuals across time. Hence, the sources of dependence between the observations across time are (1) time-invariant unobservables and (2) possible dependence between idiosyncratic shocks.

Testing the equality of two distributions with dependence has recently attracted some interest in the statistics literature. Quessy and Ethier (2012) propose the use of the multiplier method to adjust CM and characteristic function tests for the k-sample problem, where the samples are dependent. In the presence of a time effect, the data must be appropriately demeaned before testing the equality of distributions. In section 4.3, a bootstrap method is proposed for both the KS and CM tests that is shown to be asymptotically valid to test time homogeneity in the presence of generalized time effects.

Testing within-period identification under cross-sectional independence is a k-sample problem, where the samples are independent. Hence, it is a more straightforward extension of the two-sample problem and is discussed in section 4.4.

4.3 Testing Within-Group Identification

For within-group identification, there are two specific cases, time homogeneity with and without a generalized time effect. In the following, bootstrap-adjusted KS and CM statistics to test both variants of time homogeneity are proposed. To simplify illustration, the case where $T = 2$ is examined. Extensions to $T > 2$ are discussed in section 4.3.3.

4.3.1 Time Homogeneity

Under time homogeneity, we assume that for $i = 1, 2, ..., n$

$$\begin{aligned} Y_{it} &= \xi(X_{it}, \mathcal{A}_i, \mathcal{U}_{it}), \quad t = 1, 2 \\ F_{\mathcal{U}_{i2}|X_i,\mathcal{A}_i}(.|.) &= F_{\mathcal{U}_{i1}|X_i,\mathcal{A}_i}(.|.). \end{aligned} \tag{33}$$

It is important to note that the two time periods need not be adjacent to each other. We refer to them as period 1 and 2 for simplicity, but they could be any two periods in a time series.

Let $X_i = (X_{i1}, X_{i2})$, time homogeneity implies the following

$$F_{Y_{i1}|X_i}(.|(x,x)) = F_{Y_{i2}|X_i}(.|(x,x)), \ \forall x \in \mathcal{X}. \tag{34}$$

Recall that $|\mathcal{X}| = K$ by Assumption 3.1. Integrating the above with respect to $F_{X_i|\Delta X_i}(.|0)$, where $\Delta X_i \equiv X_{i2} - X_{i1}$, we obtain the following hypothesis

$$H_0^1 : F_{Y_{i1}|\Delta X_i}(.|0) = F_{Y_{i2}|\Delta X_i}(.|0) \quad \text{vs.} \quad H_A : F_{Y_{i1}|\Delta X_i}(.|0) \neq F_{Y_{i2}|\Delta X_i}(.|0) \tag{35}$$

Since Y_{i1} and Y_{i2} are observations from the same individual, the above testing problem is a paired-sample problem. To simplify notation, let $F_t(.|\Delta X_i = 0) \equiv F_{Y_{it}|X_i}(.|0)$ and $F_{t,n}(.|\Delta X_i = 0)$ denote the empirical counterpart of $F_t(.|\Delta X_i = 0)$ for $t = 1, 2$, given as follows

$$F_{t,n}(.|\Delta X_i = 0) = \frac{\sum_{i=1}^n 1\{Y_{it} \leq .\} 1\{\Delta X_i = 0\}}{\sum_{i=1}^n 1\{\Delta X_i = 0\}}.$$

Now the paired KS and CM tests are defined by the following

$$\begin{aligned} KS_{n,\mathcal{Y}} &= \sup_{y \in \mathcal{Y}} |\sqrt{n}(F_{1,n}(.|\Delta X_i = 0) - F_{2,n}(.|\Delta X_i = 0))| \\ &\equiv \|F_{1,n}(.|\Delta X_i = 0) - F_{2,n}(.|\Delta X_i = 0)\|_{\infty,\mathcal{Y}} \end{aligned} \tag{36}$$

$$\begin{aligned} CM_{n,\phi} &= \int \{\sqrt{n}(F_{1,n}(y|\Delta X_i = 0) - F_{2,n}(y|\Delta X_i = 0))\}^2 \phi(y) dy \\ &\equiv \|F_{1,n}(.|\Delta X_i = 0) - F_{2,n}(.|\Delta X_i = 0)\|_{2,\phi} \end{aligned} \tag{37}$$

where ϕ is some user-specified density.

Given the dependence between Y_{i1} and Y_{i2}, the probability-integral transform theorem no longer applies and hence the above statistics are not pivotal. The following bootstrap procedure is proposed to adjust the p-value of the above statistic. Let $(Y_i, X_i') = \{(Y_{i1}, Y_{i2})', (X_{i1}, X_{i2})'\}$. Let $\hat{F}^b_{t,n}(.|\Delta X_i = 0)$ be the empirical cdf of the b^{th} bootstrap sample, $(\hat{Y}^b_i, \hat{X}^b_i)$ given as follows

$$\hat{F}^b_{t,n}(.|\Delta X_i = 0) = \frac{\sum_{i=1}^n 1\{\hat{Y}^b_{it} \leq .\}1\{\Delta \hat{X}^b_i = 0\}}{\sum_{i=1}^n 1\{\Delta \hat{X}^b_i = 0\}}.$$

The bootstrap procedure proposed here is given by the following:

1. Compute the statistic $KS_{n,\mathcal{Y}}$ and $CM_{n,\phi}$ for $\{(Y_1, X_1), ..., (Y_n, X_n)\}$, hereinafter the original sample.

2. Resample n observations $\{(\hat{Y}_1, \hat{X}_1), ..., (\hat{Y}_n, \hat{X}_n)\}$ with replacement from the original sample. Compute the centered statistics

$$KS^b_{n,\mathcal{Y}} = \|\sqrt{n}\{\hat{F}^b_{1,n}(.|\Delta X_i = 0) - \hat{F}^b_{2,n}(.|\Delta X_i = 0) - (F_{1,n}(.|\Delta X_i = 0) - F_{2,n}(.|\Delta X_i = 0))\}\|_{\infty,\mathcal{Y}}$$
$$CM^b_{n,\phi} = \|\sqrt{n}\{\hat{F}^b_{1,n}(.|\Delta X_i = 0) - \hat{F}^b_{2,n}(.|\Delta X_i = 0) - (F_{1,n}(.|\Delta X_i = 0) - F_{2,n}(.|\Delta X_i = 0))\}\|_{2,\phi}$$

3. Repeat 1-2 B times.

4. Calculate the p-values of the tests with

$$p_{KS,n} = \sum_{b=1}^B 1\{KS^b_{n,\mathcal{Y}} > KS_{n,\mathcal{Y}}\}$$
$$p_{CM,n} = \sum_{b=1}^B 1\{CM^b_{n,\phi} > CM_{n,\phi}\}.$$

   Reject if p-value is smaller than some significance level α.

Note that as a paired sample problem, the resampling procedure is very similar to a one-sample problem, where we just resample across i. The above procedure approximates the distribution of the KS and CM statistics, under the null $F_1(.|\Delta X_i = 0) = F_2(.|\Delta X_i = 0)$, which allows us to then estimate the p-value. The following theorem shows that the bootstrap-adjusted tests have asymptotic level α and are consistent against fixed alternatives.

Theorem 4.1 *Given that* $\{(Y_i, X_i')\}_{i=1}^n$ *is an iid sequence,* $|\mathcal{X}| = K$*,* $P(\Delta X_i = 0) > 0$*, and* $F_t(.|\Delta X_i)$ *are non-degenerate for* $t = 1, 2$*, the procedure described in 1-4 for* $KS_{n,\mathcal{Y}}$ *and* $CM_{n,\phi}$ *to test* H_0^1 *(i) provides correct asymptotic size* α *and (ii) is consistent against any fixed alternative.*

The proof is in Appendix B. The convergence of the bootstrap empirical process follows by straightforward application of results in Van der Vaart and Wellner (2000). Since the limit process is a tight Brownian bridge and both test statistics are norms thereof, the statistics have positive densities. Hence, the correct asymptotic size and consistency of the test follows.

4.3.2 Testing Time Homogeneity with Time Effects

For time homogeneity with generalized time effects, we assume that

$$\begin{aligned} Y_{it} &= \xi(X_{it}, \mathcal{A}_i, \mathcal{U}_{it}) + \lambda_t(X_{it}), \quad t = 1, 2 \\ F_{\mathcal{U}_{i2}|X_i,\mathcal{A}_i}(.|.) &= F_{\mathcal{U}_{i1}|X_i,\mathcal{A}_i}(.|.), \end{aligned} \tag{38}$$

where we normalize $\lambda_1(.)$ to be zero and hence drop the subscript for $\lambda_2(.) = \lambda(.)$. Hence,

$$\begin{aligned} F_{Y_{i1}^x|X_i}(.|\underline{x}) &= F_{Y_{i2}^x - \lambda_2(x)|X_i}(.|\underline{x}) \\ &= F_{Y_{i2}^x|X_i}(. + \lambda_2(x)|\underline{x}), \quad \forall x, \underline{x} \end{aligned} \tag{39}$$

which implies that

$$F_{Y_{i1}|X_i}(.|(x,x)) = F_{Y_{i2}|X_i}(. + \lambda_2(x)|(x,x)), \quad \forall x \tag{40}$$

Now let $\Lambda = (\lambda(x^1), ..., \lambda(x^K))'$ and recall that $|\mathcal{X}| = K$, we introduce the following notation

$$F_{Y_{i2}|\Delta X_i}(., \Lambda|0) = \sum_{k=1}^{K} P(X_i = (x^k, x^k)) F_{Y_{i2}|X_i}(. + \lambda_2(x^k)|(x^k, x^k)).$$

Hence, we can write our hypothesis as follows

$$H_0^2 : F_{Y_{i1}|\Delta X_i}(.|0) = F_{Y_{i2}|\Delta X_i}(., \Lambda|0) \quad \text{vs.} \quad H_A : F_{Y_{i1}|\Delta X_i}(.|0) \neq F_{Y_{i2}|\Delta X_i}(., \Lambda|0) \tag{41}$$

We will adapt the simplified notation from above $F_2(.,\Lambda|\Delta X_i=0) \equiv F_{Y_{i2}|\Delta X_i}(.,\Lambda|0)$.

$$F_2(.,\Lambda|\Delta X_i=0) = \sum_{k=1}^{K} P(X_{i1}=X_{i2}=x^k|\Delta X_i=0)F_2(.+\lambda(x^k)|X_{i1}=X_{i2}=x^k).$$

$F_{t,n}(.|.)$ is the empirical cdf of $F_t(.|.)$. Λ_n is the sample analogue of Λ, given by

$$\begin{aligned}\Lambda_n &= \begin{pmatrix} \lambda_n(x^1) \\ ... \\ \lambda_n(x^K) \end{pmatrix} \\ &= \begin{pmatrix} \frac{\sum_{i=1}^n \Delta Y_i 1\{X_i=(x^1,x^1)\}}{\sum_{i=1}^n 1\{X_i=(x^1,x^1)} \\ ... \\ \frac{\sum_{i=1}^n \Delta Y_i 1\{X_i=(x^K,x^K)\}}{\sum_{i=1}^n 1\{X_i=(x^K,x^K)} \end{pmatrix} \end{aligned} \tag{42}$$

$$\begin{aligned} KS_{n,\mathcal{Y}}(\Lambda_n) &= \|\sqrt{n}(F_{1,n}(.|\Delta X_i=0) - F_{2,n}(.,\Lambda_n|\Delta X_i=0))\|_{\infty,\mathcal{Y}} \\ CM_{n,\phi}(\Lambda_n) &= \|\sqrt{n}(F_{1,n}(.,\Lambda_n|\Delta X_i=0) - F_{2,n}(.|\Delta X_i=0))\|_{2,\phi}. \end{aligned} \tag{43}$$

The following bootstrap procedure can be used to adjust the p-values of the above statistics. Let $\hat{\Lambda}_n^b$ be the estimator of Λ in the b^{th} bootstrap sample.

1. Compute the statistics $KS_{n,\mathcal{Y}}(\Lambda_n)$ and $CM_{n,\phi}(\Lambda_n)$ for $\{\{Y_1,X_1\},...,\{Y_n,X_n\}\}$, hereinafter the original sample.

2. Resample n observations $\{(\hat{Y}_1,\hat{X}_1),...,(\hat{Y}_n,\hat{X}_n)\}$ with replacement from the original sample. Compute the centered statistics

$$\begin{aligned} & KS_{n,\mathcal{Y}}^b(\hat{\Lambda}_n) \\ = \; & \|\sqrt{n}\{\hat{F}_{1,n}^b(.|\Delta X_i=0) - \hat{F}_{2,n}^b(.,\hat{\Lambda}_n^b|\Delta X_i=0) - (F_{1,n}(.|\Delta X_i=0) - F_{2,n}(.,\Lambda_n|\Delta X_i=0))\}\|_{\infty,\mathcal{Y}} \\ & CM_{n,\phi}^b(\hat{\Lambda}_n) \\ = \; & \|\sqrt{n}\{\hat{F}_{1,n}^b(.|\Delta X_i=0) - \hat{F}_{2,n}^b(.,\hat{\Lambda}_n^b|\Delta X_i=0) - (F_{1,n}(.|\Delta X_i=0) - F_{2,n}(.,\Lambda_n|\Delta X_i=0))\}\|_{2,\phi} \end{aligned}$$

3. Repeat 1-2 B times.

4. Calculate the p-values of the tests with

$$p_{KS,n} = \sum_{b=1}^{B} 1\{KS^b_{n,\mathcal{Y}}(\hat{\Lambda}_n) > KS_{n,\mathcal{Y}}(\Lambda_n)\}$$
$$p_{CM,n} = \sum_{b=1}^{B} 1\{CM^b_{n,\phi}(\hat{\Lambda}_n) > CM_{n,\phi}(\Lambda_n)\}.$$

Reject if p-value is smaller than some significance level α.

The presence of the generalized time effects adds noise to the empirical cdf. In order to ensure the convergence of the empirical process that are used to compute the KS and CM statistics, we impose the following condition, which ensures that the underlying distribution is uniformly continuous.

Assumption 4.1 *(Bounded Density)*
$F_t(.)$ has a density $f_t(.)$ that is bounded, i.e. $\sup_{y\in\mathcal{Y}} |f_t(y)| < \infty$, $t = 1, 2$.[31]

Theorem 4.2 *Given that $\{(Y_i, X_i')\}_{i=1}^n$ is an iid sequence, $|\mathcal{X}| = K$, $P(\Delta X_i = 0) > 0$, $F_t(.|\Delta X_i = 0)$ is non-degenerate for $t = 1, 2$, and Assumption 4.1 holds, the procedure described in 1-4 for $KS_{n,\mathcal{Y}}(\Lambda_n)$ and $CM_{n,\phi}(\Lambda_n)$ to test H_0^2 (i) provides correct asymptotic size α and (ii) is consistent against any fixed alternative.*

The proof is given in Appendix B. The key difference between the above result and Theorem 4.1 is that we have to ensure that the functional Delta method applies to ensure that the empirical process converges, hence we impose Assumption 4.1. The intuition behind imposing this regularity condition is that by demeaning the variables, we are introducing asymptotically normal noise to the empirical process. Assumption 4.1 ensures that the empirical process converges nonetheless to a Brownian bridge, by allowing us to apply the functional delta method. From here, it is straightforward to show that the bootstrap empirical process converges to the same tight limit process as the empirical process. Then, we show that the bootstrap-adjusted tests have correct asymptotic size and are consistent against fixed alternatives.

[31] Since we only demean Y_{i2}, it is sufficient to impose the above condition for that time period. It is imposed on both time periods, since the choice of demeaning the variables in the second time period is arbitrary.

4.3.3 Extensions to $T > 2$

For the case where $T > 2$, there are two possible approaches. First, one could approach this problem as a multiple testing problem, where the hypothesis would be which two periods in the time series exhibit time homogeneity. In this situation, for the different pairs of time periods, the p-values of the statistics can be computed by the bootstrap procedure given above. A multiple-testing correction, as in Romano and Shaikh (2006), can then be applied to control the family-wise error rate. An alternative approach would be to test that all time periods have the same distribution. In this situation, one can apply the bootstrap procedure above to a convex combination of the statistics for the different pairs of time periods.

4.4 Testing Within-Period Identification

For within-period identification, we have the following setup

$$\begin{aligned} Y_{it} &= \xi_t(X_{it}, \mathcal{A}_i, \mathcal{U}_{it}) \\ F_{\mathcal{A}_i,\mathcal{U}_{it}|X_i}(.,.|\underline{x}) &= F_{\mathcal{A}_i,\mathcal{U}_{it}|h(X_i)}(.,.|h(\underline{x})) \quad \text{for } t = 1, 2, ..., T. \end{aligned} \tag{44}$$

By the finiteness of $\mathcal{X}^T$, $h(\underline{x})$ also has finite-support.[32] Let $\mathcal{H}$ denote the support of $h(\underline{x})$ and $|\mathcal{H}| = L$. Let $h_l \in \mathcal{H}$. For $l = 1, ..., L$, define $\mathcal{X}_l \equiv \{\underline{x} \in \mathcal{X}^T : h(\underline{x}) = h_l\}$. Let $k_l = |\mathcal{X}_l|$.

$$H_0^3 : F_{Y_{it}|X_i}(.|\underline{x}_1) = F_{Y_{it}|X_i}(.|\underline{x}_2) = ... = F_{Y_{it}|X_i}(.|\underline{x}_{k_l}), \ \ \underline{x} \in \mathcal{X}_l, \forall l = 1, ..., L, t = 1, 2, ..., T$$

Similar to Quessy and Ethier (2012), we define

$$\bar{F}_{t,l} = \frac{1}{k_l} \sum_{\underline{x} \in \mathcal{X}_l} F_{Y_{it}|X_i}(.|\underline{x}). \tag{45}$$

[32] For identification purposes, its support has to be smaller than $\mathcal{X}$, otherwise there would be no identification gain from the restriction as shown in Theorem **??**.

The KS and CM test statistics for the restrictions in each time period are given by

$$\begin{aligned} KS_{n,t} &= \sum_{l=1}^{L} P(h(X_i) = h_l) \sum_{\underline{x} \in \mathcal{X}_l} P(X_i = \underline{x} | h(X_i) = h_l) \| \sqrt{n} \{ F_{t,n}(.|\underline{x}) - \bar{F}_{t,l,n}(.) \} \|_{\infty, \mathcal{Y}} \\ CM_{n,t} &= \sum_{l=1}^{L} P(h(X_i) = h_l) \sum_{\underline{x} \in \mathcal{X}_l} P(X_i = \underline{x} | h(X_i) = h_l) \| \sqrt{n} \{ F_{t,n}(.|\underline{x}) - \bar{F}_{t,l,n}(.) \} \|_{2, \phi} \end{aligned}$$

Averaging the above statistics over t, we obtain the following,

$$\begin{aligned} KS_{n,\mathcal{Y}} &= \frac{1}{T} \sum_{t=1}^{T} KS_{n,t} \\ CM_{n,\phi} &= \frac{1}{T} \sum_{i=1}^{T} CM_{n,t}. \end{aligned}$$

Now the following bootstrap procedure is used to approximate the distribution of the above statistics.

1. Compute the statistics $KS_{n,\mathcal{Y}}$ and $CM_{n,\phi}$ for $\{\{Y_1, X_1\}, ..., \{Y_n, X_n\}\}$, hereinafter the original sample.

2. Resample n observations $\{(\hat{Y}_1, \hat{X}_1), ..., (\hat{Y}_n, \hat{X}_n)\}$ with replacement from the original sample. Compute the centered statistics

$$\begin{aligned} KS^b_{n,\mathcal{Y}} &= \frac{1}{T} \sum_{t=1}^{T} KS^b_{n,t} \\ CM^b_{n,\phi} &= \frac{1}{T} \sum_{t=1}^{T} CM^b_{n,t}, \end{aligned}$$

   where $KS^b_{n,t}$ and $CM^b_{n,t}$ are given below.

3. Repeat 1-2 B times.

4. Calculate the p-values of the tests with

$$\begin{aligned} p_{KS,n} &= \sum_{b=1}^{B} 1\{KS^b_{n,\mathcal{Y}}(\hat{\Lambda}_n) > KS_{n,\mathcal{Y}}(\Lambda_n)\} \\ p_{CM,n} &= \sum_{b=1}^{B} 1\{CM^b_{n,\phi}(\hat{\Lambda}_n) > CM_{n,\phi}(\Lambda_n)\}. \end{aligned}$$

   Reject if p-value is smaller than some significance level α.

For event A_i, let $\hat{P}_n^b(A_i)$ be the empirical probability of A_i in the b^{th} bootstrap sample. $KS_{n,t}^b$ and $CM_{n,t}^b$ are given as follows:

$$
\begin{aligned}
& KS_{n,t}^b \\
= & \sum_{l=1}^{L} \hat{P}_n^b(h(X_i)=h_l) \sum_{\underline{x}\in\mathcal{X}_l} \hat{P}_n^b(X_i=\underline{x}|h(X_i)=h_l) \|\sqrt{n}\{\hat{F}_{1,n}(.|\underline{x}) - \hat{\bar{F}}_{1,l,n}(.|\underline{x}) - (F_{1,n}(.|\underline{x}) - \bar{F}_{l,n}(.|\underline{x}))\}\|_{\infty,\mathcal{Y}} \\
& CM_{n,t}^b \\
= & \sum_{l=1}^{L} \hat{P}_n^b(h(X_i)=h_l) \sum_{\underline{x}\in\mathcal{X}_l} \hat{P}_n^b(X_i=\underline{x}|h(X_i)=h_l) \|\sqrt{n}\{\hat{F}_{t,n}(.|\underline{x}) - \hat{\bar{F}}_{t,l,n}(.|\underline{x}) - (F_{1,n}(.|\underline{x}) - \bar{F}_{l,n}(.|\underline{x}))\}\|_{2,\phi}.
\end{aligned}
$$

The following theorem shows that the bootstrap-adjusted tests are justified asymptotically.

Theorem 4.3 *Given $\{(Y_i, X_i')\}_{i=1}^n$ is an iid sequence, $|\mathcal{X}| = K$, $P(X_i = \underline{x}) > 0$ for all $\underline{x} \in \mathcal{X}^T$, and $F_{Y_{it}|X_i}(.)$ is nondegenerate for all t, the procedure described in 1-4 for $KS_{n,\mathcal{Y}}$ and $CM_{n,\phi}$ to test H_0^3 (i) provides correct asymptotic size α and (ii) is consistent against fixed alternatives.*

The proof is in Appendix B. The convergence of the empirical and bootstrap empirical processes to a tight Brownian bridge follows from results in Van der Vaart and Wellner (2000). The remainder of the proof follows by similar arguments to Theorem **??**.

4.5 Monte Carlo Study

The baseline model is from Evdokimov (2010), but is adapted to have a binary regressor.

$$
\begin{aligned}
Y_{it} &= m(X_{it}, \mathcal{A}_i) + \mathcal{U}_{it} \\
m(x,a) &= 2a + (2+a)(2x-1)^3 \\
X_{it} &\sim \text{i.i.d.Bernoulli(0.5)}, \\
\mathcal{A}_i &= \frac{\rho}{\sqrt{T}} \sum_{t=1}^{T} \sqrt{12}(X_{it} - 0.5) + \sqrt{1-\rho^2}\psi_i \\
\psi_i &\sim \text{i.i.d.N(0,1)} \\
\epsilon_{it} &\sim N(0,1).
\end{aligned}
$$

where $\rho = 0.5$. Note that the above model exhibits time homogeneity without a time effect. Now we also include the following three variants of the above model for our simulations:

(A) Time Homogeneity with no Trend

$\mathcal{U}_{it} = (1 + X_{it})\epsilon_{it}$

$Y_{it} = m(X_{it}, \mathcal{A}_i) + \mathcal{U}_{it}$, $t = 1, 2$

(B) Time Homogeneity up to a Time Effect

$Y_{it} = m(X_{it}, \mathcal{A}_i) + \mathcal{U}_{it} + \lambda_t$, where $\lambda_1 = 0$, $\lambda_2 = 0.5$

(C) Time Homogeneity up to a Generalized Time Effect

$Y_{it} = m(X_{it}, \mathcal{A}_i) + \mathcal{U}_{it} + \lambda_t(X_{it})$, where $\lambda_1(0) = \lambda_2(0) = 0$, $\lambda_2(0) = -0.5$, $\lambda_2(1) = 0.5$

(D) Time Heterogeneity with Exclusion Restriction $F_{\mathcal{A}_i, \mathcal{U}_{it}|X_i}(.) = F_{\mathcal{A}_i, \mathcal{U}_{it}|X_{i1}}(.)$

$\mathcal{A}_i = \rho\sqrt{\frac{12}{T}}(X_{i1} - 0.5) + \sqrt{1 - \rho^2}\psi_i$

$\mathcal{U}_{i1} = (1 + X_{i1})\epsilon_{i1}$

$\mathcal{U}_{i2} = (1 + X_{i1})(\epsilon_{i2} + \lambda_2)\sigma_2$, where $\lambda_2 = 0.5$, $\sigma_2 = 1.5$ $Y_{it} = m(X_{it}, \mathcal{A}_i) + \mathcal{U}_{it}$, $t = 1, 2$

Under each model, the behavior of the bootstrap procedure for the following statistics is examined, where we use the notation introduced in sections 4.3 and 4.4.

$$
\begin{aligned}
KS^{nt}_{n,\mathcal{Y}} &= \|F_{1,n}(.|\Delta X_i = 0) - F_{2,n}(.|\Delta X_i = 0)\|_{\infty,\mathcal{Y}} && (46)\\
KS^{pt}_{n,\mathcal{Y}} &= \|F_{1,n}(.|\Delta X_i = 0) - F_{2,n}(. + \lambda_n|\Delta X_i = 0)\|_{\infty,\mathcal{Y}} && (47)\\
KS^{gt}_{n,\mathcal{Y}} &= \|F_{1,n}(.|\Delta X_i = 0) - F_{2,n}(., \Lambda_n|\Delta X_i = 0)\|_{\infty,\mathcal{Y}} && (48)\\
KS^{excl}_{n,\mathcal{Y}} &= P_n(X_{i1} = 0)\|F_{1,n}(.|X_i = (0,0)) - F_{1,n}(.|X_i = (0,1))\|_{\infty,\mathcal{Y}} \\
&+ P_n(X_{i1} = 1)\|F_{1,n}(.|X_i = (1,0)) - F_{1,n}(.|X_i = (1,1))\|_{\infty,\mathcal{Y}} && (49)\\
CM^{nt}_{n,\phi} &= \|F_{1,n}(.|\Delta X_i = 0) - F_{2,n}(.|\Delta X_i = 0)\|_{2,\phi} && (50)\\
CM^{pt}_{n,\phi} &= \|F_{1,n}(.|\Delta X_i = 0) - F_{2,n}(. + \lambda_n|\Delta X_i = 0)\|_{2,\phi} && (51)\\
CM^{gt}_{n,\phi} &= \|F_{1,n}(.|\Delta X_i = 0) - F_{2,n}(., \Lambda|\Delta X_i = 0)\|_{2,\phi} && (52)\\
CM^{excl}_{n,\phi} &= P_n(X_{i1} = 0)\|F_{1,n}(.|X_i = (0,0)) - F_{1,n}(.|X_i = (0,1))\|_{2,\phi} \\
&+ P_n(X_{i1} = 1)\|F_{1,n}(.|X_i = (1,0)) - F_{1,n}(.|X_i = (1,1))\|_{2,\phi}. && (53)
\end{aligned}
$$

where ϕ is the standard normal density.

Note that the statistics with nt super-script test time homogeneity with no trend, pt time

homogeneity with a 'pure' time effect, gt time homogeneity with a generalized time effect, $\lambda_t(X_{it})$, and $excl$ the exclusion restriction $F_{\mathcal{A}_i,\mathcal{U}_{it}|X_i} = F_{\mathcal{A}_i,\mathcal{U}_{it}|X_{i1}}$.

Under the baseline model (A), all variants of time homogeneity are not violated, whereas the exclusion restriction is. Under Model (B), both the exclusion restriction and time homogeneity with no trend are violated. Under Models (C) and (D), only time homogeneity up to a generalized time effect and the exclusion restriction, respectively, are not violated. Table 1 reports the coverage probabilities of all of the above statistics under the four different models, when $n = 1000, 1500, 2000$, and $T = 2$, where we perform 1000 simulation replications (S).

The CM statistics follow our asymptotic results in finite samples, since they control size when their respective null is true, and reject with high probability when their respective null does not hold. The KS statistic for the time homogeneity with no trend also performs well in finite samples. However, for time homogeneity with both pure and generalized time effect, the KS statistic is significantly under-sized. However, as n increases, the size properties improve. As for the exclusion restriction, the KS statistic tends to over-reject.[33]

5 Empirical Illustration: Returns to Schooling

The standard function for estimating returns to schooling is Mincer's human capital earnings function, given by

$$Y = \alpha + \beta S + \gamma E + \delta E^2 + \mathcal{U} \tag{54}$$

where Y is log earnings, S is years of completed education, and E is the number of years an individual has worked. Since potential experience is used instead, where $E = Age - S - 6$, many researchers just control for S and Age. Angrist and Newey (1991) examine fixed effects estimation of Mincer's equation using a subsample of the national longitudinal survey of youth (NLSY), since they observe changes in schooling status for 20% of their

[33]For $n = 3000$, under Model (D), its finite-sample performance matches the asymptotic results, and it exhibits good size control.

sample.[34] They also propose a test of the over-identifying restrictions of the linear fixed effects model and reject it for Mincer's equation. In the following, we will test time homogeneity up to a location time effect, which may be viewed as a nonparametric test of the fixed-effects assumption in the presence of time effects. The model is given by the following equation,

$$\begin{aligned} Y_{it} &= \xi(S_{it}, \mathcal{A}_i, \mathcal{U}_{it}) + \lambda_t \\ \mathcal{U}_{it}|S_i, \mathcal{A}_i &\stackrel{d}{=} \mathcal{U}_{i1}|S_i, \mathcal{A}_i. \end{aligned}$$

Our results indicate that we cannot reject the time homogeneity assumption. Hence, we cannot reject the fixed effects assumption nonparametrically. Since our nonparametric estimates of the APE indicate violations of the linear model, we conjecture that the rejection of the over-identification test in Angrist and Newey (1991) is due to linearity.

5.1 Data and Methodology

Angrist and Newey (1991) use a NLSY random subsample of young men from 1983-1987. The young men are aged 18-26 in 1983. 20% of the subsample exhibits changes in schooling. We use a revised version of this sample with 1087 young men. The descriptive statistics are reported in Table 2.[35]

The linear specification is the most widely used specification of Mincer's equation. Card (1999) however points out that there is no economic justification for the linear specification and cites empirical findings of possible nonlinearities in the relationship between schooling and earnings. In this spirit, we propose the following model exhibiting time homogeneity up to a location time effect

$$\begin{aligned} Y_{it} &= \xi(S_{it}, \mathcal{A}_i, \mathcal{U}_{it}) + \lambda_t \\ \mathcal{U}_{it}|S_i, \mathcal{A}_i &\stackrel{d}{=} \mathcal{U}_{i1}|S_i, \mathcal{A}_i. \end{aligned}$$

[34] Panel data methods are not often used for the identification of returns to schooling, since it is unlikely to observe changes in schooling over the short span of a micro-panel. For instance, Chamberlain (1984) includes schooling as a time-invariant control variable in the union-wage example.

[35] Angrist and Newey (1991) had 1045 in their sample. Despite the difference between their original sample and the revised one used here, the descriptive statistics are quite similar. Hence, this difference cannot be driving the difference in our results. We also replicate their estimation results in Table 5.

where Y_{it} is log earnings, and we normalize λ_1 to zero. As noted above, this model allows us to identify the effect of schooling on earnings using within-group variation.

The model implies that individuals that do not change their schooling status, i.e. stayers, should have the same distribution for log earnings across time, once appropriately demeaned. Formally, the testable implication is given as follows

$$F_{Y_{i1}|\Delta S_i}(.|0) \quad = \quad F_{Y_{i2}|\Delta S_i}(.+\lambda_2|0). \tag{55}$$

Now we test the above implication for the following year pairs, 1983-84, 1984-85, 1985-86, and 1986-87, using the following KS and CM statistics

$$\begin{aligned} KS_{n,\mathcal{Y}} &= \sup_{y\in\mathcal{Y}} \left| \frac{\sum_{i=1}^n (1\{Y_{i1}\le y\} - 1\{Y_{i2}-\lambda_n \le y\})1\{\Delta S_i = 0\}}{\sum_{i=1}^n 1\{\Delta S_i = 0\}} \right| \\ CM_{n,\phi} &= \int \left(\frac{\sum_{i=1}^n (1\{Y_{i1}\le y\} - 1\{Y_{i2}-\lambda_n \le y\})1\{\Delta S_i = 0\}}{\sum_{i=1}^n 1\{\Delta S_i = 0\}} \right)^2 \phi(y)dy, \end{aligned}$$

where $\mathcal{Y}$[36] is the support of Y_{it} and ϕ is the standard normal density, but any other density may also be used. $\lambda_n = \sum_{i=1}^n (Y_{i2}-Y_{i1})1\{\Delta S_i = 0\}/\sum_{i=1}^n 1\{\Delta S_i = 0\}$. The p-values for both statistics were obtained using the bootstrap procedure outlined above.

The above model has another implication. All stayer subpopulations regardless of the years of schooling completed must exhibit the same mean shift across time. Formally,

$$E[Y_{i2}-Y_{i1}|S_i=(s,s)] \quad = \quad E[Y_{i2}-Y_{i1}|S_i=(s',s')] \;\; \forall s,s' \in \mathcal{S}, s\neq s', \tag{56}$$

where $\mathcal{S}$ denotes the support of S_{it}. This implication can be tested by an F-test, which we also report.

For every year pair, the APEs for movers are then estimated as follows

$$\hat{\beta}(s\to s+1|S_i=(s,s+1)) \quad = \quad \frac{\sum_{i=1}^n (Y_{i2}-Y_{i1})1\{S_i=(s,s+1)\}}{\sum_{i=1}^n 1\{S_i=(s,s+1)\}} - \hat{\lambda}_{2,n}, \tag{57}$$

[36] To implement the KS statistic in practice, I search for the maximum over a fine grid between the minimum and maximum value of Y_{it} in the sample.

where $\hat{\lambda}_{2,n} = \sum_{i=1}^{n}(Y_{i2} - Y_{i1})1\{\Delta S_i = 0\}/\sum_{i=1}^{n} 1\{\Delta S_i = 0\}$. The APE for all movers[37] is given by

$$\hat{\beta}(\Delta S_i = 1) \quad = \quad \sum_{s \in \mathcal{S}} P_n(S_i = (s, s+1)|\Delta S_i = 1)\hat{\beta}(s \to s+1|S_i = (s, s+1)), \quad (58)$$

The standard errors are computed under the assumption of cross-sectional independence.

5.2 Results and Discussion

Table 3 shows the p-values for the bootstrap-adjusted KS and CM tests as well as the F test for time homogeneity up to a time effect. We use 500 bootstrap replications to adjust the p-value of the KS and CM statistics. The p-values indicate that we cannot reject the fixed-effects assumption for any of the year-pairs. Comparing our findings to Angrist and Newey (1991), we conjecture that the rejection in Angrist and Newey (1991) is due to misspecification of the linear model rather than a violation of the fixed-effects assumption, especially given our estimates of the APE, which we give below.

The estimates of the APE for mover subpopulations are reported in Table 4. Similar to Angrist and Newey (1991), our APE estimates are not significant, except in 1985-1986. Our results also provides evidence against the implication of the linear model that the APE for identified subpopulations is constant across time. Furthermore, it is important to note here that the object we are estimating is not a *ceteris paribus* effect. Recall that $E = Age - S - 6$. Since everyone's age increases across time, individuals that increase their schooling are forgoing a year of potential experience, while individuals that do not increase their schooling gain a year of potential experience. Hence, the estimate of the APE here identifies the effect of schooling corrected for the opportunity cost of forgoing one year of potential experience.

Now the significance of the APE in 1985-86 is driven solely by individuals that completed their bachelor degree, i.e. 16 years of schooling. This is in line with the empirical evidence quoted in Card (1999) that the changes in schooling status have particularly significant effects on earnings around the completion of terminal degrees.

[37] Since schooling only increases by unit increments only, $\Delta S_i = 1$ characterizes all mover subpopulations.

6 Concluding Remarks

The paper at hand contributes to the nonparametric identification literature for nonseparable panel data models in two main ways. It attempts to characterize the trade-off between different assumptions that achieve identification of the APE for a subpopulation by quasi-differencing. It also provides a menu of testable identifying assumptions and proposes tests for these assumptions that are asymptotically valid and perform well in finite samples. The empirical illustration shows that testing identifying assumptions, such as the time homogeneity assumption, nonparametrically may aid the empirical researcher in justifying the choice of an identification strategy, even if a parametric model is used for estimation and inference.

The identifying assumptions given above can be used to identify other aspects of the distribution of the outcome variable due to changes in variables of interest. This will be addressed in future work. The inclusion of continuous control variables, such as macroeconomic indicators, is another important direction for future work. Finally, the identifying assumptions hold if the appropriate control variables are in place. Future work will examine a method that would allow the empirical researcher to select the control variables that aid identification.

A Proofs of Section 3 Results

Proof (Theorem 3.1)

Note that

$$\begin{aligned}
& E[Y_{i2}^x - Y_{i1}|X_i = (x,x')] - E[Y_{i2} - Y_{i1}|X_i = (x,x)] \\
= & \int (\xi_2(x,a,u_2) - \xi_1(x,a,u_1))(dF_{\mathcal{A}_i,\mathcal{U}_{i1},\mathcal{U}_{i2}|X_i}(a,u_1,u_2|(x,x')) - dF_{\mathcal{A}_i,\mathcal{U}_{i1},\mathcal{U}_{i2}|X_i}(a,u_1,u_2|(x,x))) \\
= & \int (\xi_2(x,a,u_2) - \xi_1(x,a,u_1)) \\
& \times (f_{\mathcal{A}_i,\mathcal{U}_{i1},\mathcal{U}_{i2}|X_i}(a,u_1,u_2|(x,x')) - f_{\mathcal{A}_i,\mathcal{U}_{i1},\mathcal{U}_{i2}|X_i}(a,u_1,u_2|(x,x)))d(a,u_1,u_2)
\end{aligned}$$

Let $\lambda(x) = \xi_2(x,a,u) - \xi_1(x,a,u)$, where $\lambda(x)$ is the component of the difference between the structural function that only depends on the regressors. Now adding and subtracting $\lambda(x)$ to the integrand, we obtain the following,

$$\begin{aligned}
& E[Y_{i2}^x - Y_{i1}|X_i = (x,x')] - E[Y_{i2} - Y_{i1}|X_i = (x,x)] \\
= & \int (\xi_2(x,a,u_2) - \xi_1(x,a,u_1) - \lambda(x)) \\
& \times (f_{\mathcal{A}_i,\mathcal{U}_{i1},\mathcal{U}_{i2}|X_i}(a,u_1,u_2|(x,x')) - f_{\mathcal{A}_i,\mathcal{U}_{i1},\mathcal{U}_{i2}|X_i}(a,u_1,u_2|(x,x)))d(a,u_1,u_2) \\
= & \int (\xi_2(x,a,u_2) - \xi_1(x,a,u_1) - \lambda(x)) \\
& \times \left(\frac{f_{\mathcal{A}_i,\mathcal{U}_{i1},\mathcal{U}_{i2}|X_i}(a,u_1,u_2|(x,x'))}{f_{\mathcal{A}_i,\mathcal{U}_{i1},\mathcal{U}_{i2}|X_i}(a,u_1,u_2|(x,x))} - 1 \right) f_{\mathcal{A}_i,\mathcal{U}_{i1},\mathcal{U}_{i2}|X_i}(a,u_1,u_2|(x,x))d(a,u_1,u_2).
\end{aligned}$$

The above is equal to zero if

$$(\xi_2(x,a,u_2) - \xi_1(x,a,u_1) - \lambda(x)) \left(\frac{f_{\mathcal{A}_i,\mathcal{U}_{i1},\mathcal{U}_{i2}|X_i}(a,u_1,u_2|(x,x'))}{f_{\mathcal{A}_i,\mathcal{U}_{i1},\mathcal{U}_{i2}|X_i}(a,u_1,u_2|(x,x))} - 1 \right) = 0. \; \forall a,u_1,u_2$$

□

Proof (Variation 3.1)

Under time homogeneity, the condition in Theorem 3.1 simplifies to

$$(\xi_2(x,a,u) - \xi_1(x,a,u) - \lambda(x)) \left(f_{\mathcal{A}_i,\mathcal{U}_{i1}|X_i}(a,u|(x,x')) - f_{\mathcal{A}_i,\mathcal{U}_{i1}|X_i}(a,u|(x,x)) \right) = 0. \quad \forall a,u$$

Under unrestricted individual heterogeneity, the second term is not equal to zero. Hence, the above is zero if

$$\xi_2(x,a,u) - \xi_1(x,a,u) = \lambda(x) \; \forall a,u$$

Without loss of generality, we can write the above as $\xi_t(x,a,u) = \xi(x,a,u) + \lambda_t(x)$. □

Proof (Variation 3.2)

Recall that $E[Y_{i2}^x - Y_{i1}|X_i = (x,x')] = E[Y_{i2} - Y_{i1}|X_i = (x,x)]$ can be written as

$$\int (\xi_2(x,a,u_2) - \xi_1(x,a,u_1) - \lambda(x)) \times \left(f_{\mathcal{A}_i,\mathcal{U}_{i1},\mathcal{U}_{i2}|X_i}(a,u_1,u_2|(x,x')) - f_{\mathcal{A}_i,\mathcal{U}_{i1},\mathcal{U}_{i2}|X_i}(a,u_1,u_2|(x,x)) \right) d(a,u_1,u_2) = 0.$$

By the conditions of this variation, $\xi_2(x,a,u_2) = \xi_1(x,a,u_2) + \lambda(x)$. Plugging this into the integrand of the previous equation, it follows that

$$\int (\xi_1(x,a,u_2) - \xi_1(x,a,u_1)) \times (f_{\mathcal{A}_i,\mathcal{U}_{i1},\mathcal{U}_{i2}|X_i}(a,u_1,u_2|(x,x')) - f_{\mathcal{A}_i,\mathcal{U}_{i1},\mathcal{U}_{i2}|X_i}(a,u_1,u_2|(x,x))) = 0$$

Under unrestricted individual heterogeneity and $f_{\mathcal{A}_i,\mathcal{U}_{i1},\mathcal{U}_{i2}|X_i}(.) > 0$ by Assumption 3.2, the above equals to zero if each term is zero. Hence, for $\underline{x} \in \{(x,x),(x,x')\}$,

$$\int (\xi_1(x,a,u_2) f_{\mathcal{A}_i,\mathcal{U}_{i2}|X_i}(a,u_2|\underline{x}) - \xi_1(x,a,u_1) f_{\mathcal{A}_i,\mathcal{U}_{i1}|X_i}(a,u_1|\underline{x})) d(a,u_1,u_2) = 0$$

The above holds if

$$\xi_1(x,a,u) f_{\mathcal{A}_i,\mathcal{U}_{i2}|X_i}(a,u|\underline{x}) = \xi_1(x,a,u) f_{\mathcal{A}_i,\mathcal{U}_{i1}|X_i}(a,u|\underline{x}) \; \forall a,u$$

which holds if $f_{\mathcal{A}_i,\mathcal{U}_{i1}|X_i}(a,u|\underline{x}) = f_{\mathcal{A}_i,\mathcal{U}_{i2}|X_i}(a,u|\underline{x}) \; \forall a,u$, which implies time homogeneity,

i.e. $\mathcal{U}_{i1}|X_i, \mathcal{A}_i \stackrel{d}{=} \mathcal{U}_{i2}|X_i, \mathcal{A}_i$. □

Proof (Variation 3.3)

Recall the condition in Theorem 3.1

$$\begin{aligned}&(\xi_2(x,a,u_2) - \xi_1(x,a,u_1) - \lambda(x)) \\ &\times \left(f_{\mathcal{A}_i,\mathcal{U}_{i1},\mathcal{U}_{i2}|X_i}(a,u_1,u_2|(x,x')) - f_{\mathcal{A}_i,\mathcal{U}_{i1},\mathcal{U}_{i2}|X_i}(a,u_1,u_2|(x,x))\right) = 0.\end{aligned}$$

Clearly, without restrictions on time heterogeneity and the structural function, the above is true if

$$f_{\mathcal{A}_i,\mathcal{U}_{i1},\mathcal{U}_{i2}|X_i}(a,u_1,u_2|(x,x')) \;=\; f_{\mathcal{A}_i,\mathcal{U}_{i1},\mathcal{U}_{i2}|X_i}(a,u_1,u_2|(x,x)) \;\; \forall a,u_1,u_2.$$

□

Proof (Variation 3.4) For notational convenience, we set $\xi_1(x,a,u) = \xi(x,a,u)$. By Assumption 3.1(iii), Fubini's theorem applies as follows,

$$\begin{aligned}&\int (\xi(x,a,u_2) - \xi(x,a,u_1) - \lambda(x)) d(F_{\mathcal{A}_i,\mathcal{U}_{i1},\mathcal{U}_{i2}|X_i}(a,u_1,u_2|(x,x')) - F_{\mathcal{A}_i,\mathcal{U}_{i1},\mathcal{U}_{i2}|X_i}(a,u_1,u_2|(x,x))) \\ = &\int (m_2(x,a,X_i) - m_1(x,a,X_i) - \lambda(x)) d(F_{\mathcal{A}_i|X_i}(a|(x,x')) - F_{\mathcal{A}_i|X_i}(a|(x,x))),\end{aligned}$$

where $m_t(x,a,\underline{x}) = \int \xi(x,a,u) dF_{\mathcal{U}_{it}|X_i,\mathcal{A}_i}(u|\underline{x},a)$.

Under arbitrary individual heterogeneity and $f_{\mathcal{A}_i|X_i}(.|.) > 0$, the above is zero if

$$(m_2(x,a,\underline{x}) - m_1(x,a,\underline{x}) - \lambda(x)) \;=\; 0 \; \forall a,\underline{x}, \tag{59}$$

By Assumption 3.3 (i), (ii) and the dominated convergence theorem, $\partial(m_2(x,a,\underline{x}) - m_1(x,a,\underline{x}))/\partial a$ exists and is continuous. Now note that

$$\begin{aligned}&m_2(x',a',\underline{x}') - m_1(x',a',\underline{x}') - (m_2(x,a,\underline{x}) - m_1(x,a,\underline{x})) \\ = \;& m_2(x',a',\underline{x}') - m_1(x',a',\underline{x}') - (m_2(x',a',\underline{x}) - m_1(x',a',\underline{x})) \\ + \;& m_2(x',a',\underline{x}) - m_1(x',a',\underline{x}) - (m_2(x',a,\underline{x}) - m_1(x',a,\underline{x})) \\ + \;& m_2(x',a,\underline{x}) - m_1(x',a,\underline{x}) - (m_2(x,a,\underline{x}) - m_1(x,a,\underline{x}))\end{aligned}$$

$$
\begin{aligned}
&= m_2(x', a', \underline{x}') - m_1(x', a', \underline{x}') - (m_2(x', a', \underline{x}) - m_1(x', a', \underline{x})) \\
&+ \int_a^{a'} \frac{\partial(m_2(x', a, \underline{x}) - m_1(x', a, \underline{x}))}{\partial a} da \\
&+ m_2(x', a, \underline{x}) - m_1(x', a, \underline{x}) - (m_2(x, a, \underline{x}) - m_1(x, a, \underline{x})) \\
&\quad \forall x, x', a, a', \underline{x}, \underline{x}'
\end{aligned} \tag{60}
$$

where the last equality follows by the first fundamental theorem of calculus and the continuity of $\partial(m_2(x', a, \underline{x}) - m_1(x', a, \underline{x}))/\partial a$. Now if the following conditions hold

$$
\frac{\partial(m_2(x, a, \underline{x}) - m_1(x, a, \underline{x}))}{\partial a} = 0 \quad \forall a, \underline{x} \tag{61}
$$

$$
\frac{\bar{\partial}(m_2(x, a, \underline{x}) - m_1(x, a, \underline{x}))}{\bar{\partial}(\underline{x}, \underline{x}')} = 0 \quad \forall a, \underline{x}, \underline{x}' \tag{62}
$$

where $\bar{\partial} g(x)/\bar{\partial}(x, x') = g(x) - g(x')$, (60) yields the following

$$
m_2(x', a', \underline{x}') - m_1(x', a', \underline{x}') = m_2(x', a, \underline{x}) - m_1(x', a, \underline{x}) \quad \forall x, x', a, a', \underline{x}, \underline{x}' \tag{63}
$$

since $a \neq a'$ and $\underline{x} \neq \underline{x}'$ for some a' and $\underline{x}'$, (63) implies (59) by the arbitrariness of $\lambda(x)$. Note that we can re-write (61) as follows

$$
\frac{\partial(m_2(x, a, \underline{x}) - m_1(x, a, \underline{x}))}{\partial a} = \int (\xi(x, a, u_2) - \xi(x, a, u_1)) dF_{\mathcal{U}_{i1}, \mathcal{U}_{i2}|X_i, \mathcal{A}_i}(u_1, u_2|\underline{x}, a)
$$

By Assumption 3.3(ii) and the dominated convergence theorem,

$$
\begin{aligned}
&\frac{\partial(m_2(x, a, \underline{x}) - m_1(x, a, \underline{x}))}{\partial a} \\
&= \int \frac{\partial(\xi(x, a, u_2) - \xi(x, a, u_1)) f_{\mathcal{U}_{i1}, \mathcal{U}_{i2}|X_i, \mathcal{A}_i}(u_1, u_2|\underline{x}, a))}{\partial a} d(u_1, u_2), \\
&= \int \frac{\partial(\xi(x, a, u_2) - \xi(x, a, u_1))}{\partial a} f_{\mathcal{U}_{i1}, \mathcal{U}_{i2}|X_i, \mathcal{A}_i}(u_1, u_2|\underline{x}, a) d(u_1, u_2) \\
&\quad + \int (\xi(x, a, u_2) - \xi(x, a, u_1)) f^a_{\mathcal{U}_{i1}, \mathcal{U}_{i2}|X_i, \mathcal{A}_i}(u_1, u_2|\underline{x}, a) d(u_1, u_2) \\
&= \int \frac{\partial(\xi(x, a, u_2) - \xi(x, a, u_1))}{\partial a} f_{\mathcal{U}_{i1}, \mathcal{U}_{i2}|X_i, \mathcal{A}_i}(u_1, u_2|\underline{x}, a) d(u_1, u_2) \\
&+ \int (\xi(x, a, u_2) - \xi(x, a, u_1)) \frac{f^a_{\mathcal{U}_{i1}, \mathcal{U}_{i2}|X_i, \mathcal{A}_i}(u_1, u_2|\underline{x}, a)}{f_{\mathcal{U}_{i1}, \mathcal{U}_{i2}|X_i, \mathcal{A}_i}(u_1, u_2|\underline{x}, a)} f_{\mathcal{U}_{i1}, \mathcal{U}_{i2}|X_i, \mathcal{A}_i}(u_1, u_2|\underline{x}, a) d(u_1, u_2)
\end{aligned} \tag{64}
$$

Now the above equals to zero if

$$\begin{aligned} & \frac{\partial(\xi(x,a,u_2)-\xi(x,a,u_1))}{\partial a} f_{\mathcal{U}_{i1},\mathcal{U}_{i2}|X_i,\mathcal{A}_i}(u_1,u_2|\underline{x},a) \\ + \quad & (\xi(x,a,u_2)-\xi(x,a,u_1)) f^a_{\mathcal{U}_{i1},\mathcal{U}_{i2}|X_i,\mathcal{A}_i}(u_1,u_2|\underline{x},a)=0 \quad \forall a,u_1,u_2,\underline{x} \end{aligned} \tag{65}$$

The above is true if the following conditions hold

$$\frac{\partial(\xi(x,a,u_2)-\xi(x,a,u_1))}{\partial a} = 0 \ \forall a,u_1,u_2 \tag{66}$$

$$f^a_{\mathcal{U}_{i1},\mathcal{U}_{i2}|X_i,\mathcal{A}_i}(u_1,u_2|\underline{x},a) = 0 \ \forall a,u_1,u_2,\underline{x}. \tag{67}$$

As for (62), let $\bar{f}^z(z)=\bar{\partial}f(z)/\bar{\partial}(z,z')$.

$$\frac{\bar{\partial}(m_2(x,a,\underline{x})-m_1(x,a,\underline{x}))}{\bar{\partial}\underline{x}} = \int (\xi(x,a,u_2)-\xi(x,a,u_1)) \bar{f}^{\underline{x}}_{\mathcal{U}_{i1},\mathcal{U}_{i2}|X_i,\mathcal{A}_i}(u_1,u_2|\underline{x},a) d(u_1,u_2)$$

The above equals to zero if

$$(\xi(x,a,u_2)-\xi(x,a,u_1)) \bar{f}^{\underline{x}}_{\mathcal{U}_{i1},\mathcal{U}_{i2}|X_i,\mathcal{A}_i}(u_1,u_2|\underline{x},a) = 0, \ \forall a,u_1,u_2,\underline{x} \tag{68}$$

since $(\xi(x,a,u_2)-\xi(x,a,u_1)) \neq 0$ for some a,u_1,u_2, it follows that the above is true iff

$$\bar{f}^{\underline{x}}_{\mathcal{U}_{i1},\mathcal{U}_{i2}|X_i,\mathcal{A}_i}(u_2|\underline{x},a) = 0 \tag{69}$$

We obtain (i) by noting that (66) is true if

$$\partial\xi^a(x,a,u)/\partial u = 0. \ \forall a,u$$

Now (67) and (69) are true if $f_{\mathcal{U}_{it}|X_i,\mathcal{A}_i}(.) = f_{\mathcal{U}_{it}}(.)$, which is equivalent to (ii) by Assumption 3.3(i). □

Proof (Theorem 3.2)

By (iii) $\underline{x}^c_t = \underline{x}^c_\tau = x$,

$$\begin{aligned} E[Y_{i\tau}-Y_{it}|X_i=\underline{x}^c] &\overset{(i)}{=} \int (\xi_\tau(x,a,u)-\xi_t(x,a,u)) dF_{\mathcal{A}_i,\mathcal{U}_{it}|X_i}(a,u|\underline{x}^c) \\ &\overset{(ii)}{=} \lambda_\tau(h(\underline{x}^c)) \end{aligned} \tag{70}$$

Note that $h(\underline{x}) = h(\underline{x}^c)$.

$$\begin{aligned} E[Y_{i\tau} - Y_{it}|X_i = \underline{x}] &\overset{(i)}{=} \int (\xi_\tau(\underline{x}_\tau, a, u) - \xi_t(\underline{x}_t, a, u))dF_{\mathcal{A}_i,\mathcal{U}_{it}|X_i}(a, u|\underline{x}) \\ &= \int (\xi_t(\underline{x}_\tau, a, u) - \xi_t(\underline{x}_t, a, u))dF_{\mathcal{A}_i,\mathcal{U}_{it}|X_i}(a, u|\underline{x}) + \lambda_\tau(h(\underline{x})) \\ &= \beta_t(\underline{x}_t \to \underline{x}_\tau|X_i = \underline{x}) + \lambda_\tau(h(\underline{x})) \end{aligned}$$

Thus, by (iii)

$$\beta_t(\underline{x}_t \to \underline{x}_\tau|X_i = \underline{x}) = E[Y_{i\tau} - Y_{it}|X_i = \underline{x}] - E[Y_{i\tau} - Y_{it}|X_i = \underline{x}^c]$$

□

Proof (Theorem 3.3)

$$\begin{aligned} & E[Y_{it}|X_i = \underline{x}] - E[Y_{it}|X_i = \underline{x}^c] \\ = & \int \xi_t(\underline{x}_t, a, u)dF_{\mathcal{A}_i,\mathcal{U}_{it}|X_i}(a, u|\underline{x}) - \int \xi_t(\underline{x}_t^c, a, u)dF_{\mathcal{A}_i,\mathcal{U}_{it}|X_i}(a, u|\underline{x}^c) \\ \overset{\text{(i) \& (ii)}}{=} & \int (\xi_t(\underline{x}_t, a, u) - \xi_t(\underline{x}_t^c, a, u))dF_{\mathcal{A}_i,\mathcal{U}_{it}|h(X_i)}(a, u|h(\underline{x})) \\ = & \beta_t(\underline{x}_t^c \to \underline{x}_t|X_i = \underline{x}) \\ = & \beta_t(\underline{x}_t^c \to \underline{x}_t|X_i = \underline{x}^c) \end{aligned}$$

□

B Proofs of Section 4 Results

Proof (Theorem 4.1)

In order to show (i) and (ii), we have to show that the following empirical and bootstrap empirical processes converge to the same tight Brownian bridge. Let $\mathbb{G}_{n|\Delta X_i=0}$ and $\hat{\mathbb{G}}_{n|\Delta X_i=0}$ denote the empirical and bootstrap empirical processes, respectively.

$$\begin{aligned} \mathbb{G}_{n|\Delta X_i=0} &= \sqrt{n}(F_{1,n}(.|\Delta X_i = 0) - F_1(.|\Delta X_i = 0) - (F_{2,n}(.|\Delta X_i = 0) - F_2(.|\Delta X_i = 0))) \\ \hat{\mathbb{G}}_{n|\Delta X_i=0} &= \sqrt{n}(\hat{F}_{1,n}(.|\Delta X_i = 0) - F_{1,n}(.|\Delta X_i = 0) - (\hat{F}_{2,n}(.|\Delta X_i = 0) - F_{2,n}(.|\Delta X_i = 0))), \end{aligned}$$

where $\hat{F}_{t,n} = \sum_{i=1}^{n} M_{ni} 1\{Y_{it} \leq y\} 1\{\Delta X_i = 0\} / \sum_{i=1}^{n} M_{ni} 1\{\Delta X_i = 0\}$.

Since the unconditional cdfs trivially fulfill the Donsker property, it follows that they converge to a tight Brownian bridge

$$\sqrt{n} \begin{pmatrix} F_{1,n}(.) - F_1(.) \\ F_{2,n}(.) - F_2(.) \end{pmatrix} \rightsquigarrow \begin{pmatrix} \mathbb{G}_1 \\ \mathbb{G}_2 \end{pmatrix}, \tag{71}$$

where $\mathbb{G}_1$ and $\mathbb{G}_2$ are each a tight Brownian bridge on $L^\infty(\mathcal{Y})$.

Since $P(\Delta X_i = 0) > 0$ and $\{(Y_i, X_i')\}_{i=1}^n$ is an i.i.d. sequence, Lemma C.3 applies, which implies that

$$\sqrt{n} \begin{pmatrix} F_{1,n}(.|\Delta X_i) - F_1(.|\Delta X_i) \\ F_{2,n}(.|\Delta X_i = 0) - F_2(.|\Delta X_i = 0) \end{pmatrix} \rightsquigarrow \begin{pmatrix} \frac{\mathbb{G}_1}{P(\Delta X_i = 0)} + \mathcal{Z} F_1(., \Delta X_i = 0) \\ \frac{\mathbb{G}_2}{P(\Delta X_i = 0)} + \mathcal{Z} F_2(., \Delta X_i = 0) \end{pmatrix} \equiv \begin{pmatrix} \mathbb{G}_{1|\Delta X=0} \\ \mathbb{G}_{2|\Delta X=0} \end{pmatrix}$$

$$\sqrt{n} \begin{pmatrix} \hat{F}_{1,n}(.|\Delta X_i) - F_{1,n}(.|\Delta X_i) \\ \hat{F}_{2,n}(.|\Delta X_i = 0) - F_{2,n}(.|\Delta X_i = 0) \end{pmatrix} \rightsquigarrow \begin{pmatrix} \mathbb{G}_{1|\Delta X=0} \\ \mathbb{G}_{2|\Delta X=0} \end{pmatrix}$$

where $\sqrt{n}(P_n(\Delta X_i = 0) - P(\Delta X_i = 0)) \rightsquigarrow \mathcal{Z}$, where $P_n(\Delta X_i = 0) = \sum_{i=1}^n 1\{\Delta X_i = 0\}/n$. By the tightness of $\mathbb{G}_t$ for $t = 1, 2$ and $\mathcal{Z}$, it follows that $\mathbb{G}_{t|\Delta X=0}$ is a tight Brownian bridge for $t = 1, 2$. By continuous mapping theorem, it follows that

$$\begin{aligned} \mathbb{G}_{n|\Delta X_i=0} &\rightsquigarrow \frac{\mathbb{G}_1 - \mathbb{G}_2}{P(\Delta X_i = 0)} + \mathcal{Z}(F_1(., \Delta X_i = 0) - F_2(., \Delta X_i = 0)) \equiv \mathbb{H} \\ \hat{\mathbb{G}}_{n|\Delta X_i=0} &\rightsquigarrow \mathbb{H}. \end{aligned}$$

where $(\mathbb{G}_1 - \mathbb{G}_2)$ is a tight Brownian bridge in $\mathcal{L}^\infty(\mathcal{F})$, where $\mathcal{F} = \{1\{Y_{i1} \leq y\} - 1\{Y_{i2} \leq y\} : y \in \mathcal{Y}\}$. Thus, $\mathbb{H}$ is a tight Brownian bridge.

Since $\|.\|_{\infty,\mathcal{Y}}$ and $\|.\|_{2,\phi}$ are continuous, convex functionals, it follows that $\|\mathbb{H}\|_{\infty,\mathcal{Y}}$ and $\|\mathbb{H}\|_{2,\phi}$ has absolutely continuous and strictly increasing distribution on its support $[0, \infty)$, except possibly at zero, by Theorem 11.1 in Davydov, Lifshits, and Smorodina (1998). Since $F_1(.|\Delta X_i = 0)$ and $F_2(.|\Delta X_i = 0)$ are non-degenerate, then $P(\|\mathbb{H}\|_{\infty,\mathcal{Y}} = 0) = 0$ and $P(\|\mathbb{H}\|_{2,\phi} = 0) = 0$. Thus, both $\|\mathbb{H}\|_{\infty,\mathcal{Y}}$ and $\|\mathbb{H}\|_{2,\phi}$ have absolutely continuous distributions on $[0, \infty)$.

Now the critical values of the bootstrap-adjusted KS and CM tests are given by

$$\begin{aligned}\hat{c}_n^{KS} &= \inf\{t : \hat{P}_n(\|\hat{\mathbb{G}}_{n|\Delta X_i=0}\|_{\infty,\mathcal{Y}} > t) \leq \alpha\},\\ \hat{c}_n^{CM} &= \inf\{t : \hat{P}_n(\|\hat{\mathbb{G}}_{n|\Delta X_i=0}\|_{2,\phi} > t) \leq \alpha\},\end{aligned}$$

where $\hat{P}_n$ is the bootstrap probability measure for the sample. Given the above, it follows that

$$\begin{aligned}\hat{c}_n^{KS} \xrightarrow{p} c^{KS} &= \inf\{t : P(\|\mathbb{H}\|_{\infty,\mathcal{Y}} > t) \leq \alpha\},\\ \hat{c}_n^{CM} \xrightarrow{p} \tilde{c}^{CM} &= \inf\{t : P(\|\mathbb{H}\|_{2,\phi} > t) \leq \alpha\}.\end{aligned}$$

Thus, under the null, the bootstrap-adjusted statistics have correct asymptotic size. Hence, we have shown (i). By the tightness of the limit process, it follows that $\hat{c}_n^{KS}$ and $\hat{c}_n^{CM}$ are bounded in probability. Thus, the tests are consistent against any fixed alternative, which proves (ii). □

Proof (Theorem 4.2)

In order to show (i) and (ii), we first have to show that the underlying empirical and bootstrap empirical processes, given below, converge to the same tight Brownian bridge. Once this is established, the proofs of (i) and (ii) follow by similar arguments to Theorem 4.1.

We define the empirical and bootstrap empirical processes, $\mathbb{G}_{n|\Delta X_i=0}(\Lambda_n)$ and $\hat{\mathbb{G}}_{n|\Delta X_i=0}(\hat{\Lambda}_n)$

$$\begin{aligned}\mathbb{G}_{n|\Delta X_i=0}(\Lambda_n) &= \sqrt{n}(F_{1,n}(.|\Delta X_i = 0) - F_1(.|\Delta X_i = 0) - (F_{2,n}(., \Lambda_n|\Delta X_i = 0) - F_2(., \Lambda|\Delta X_i = 0)))\\ \hat{\mathbb{G}}_{n|\Delta X_i=0}(\hat{\Lambda}_n) &= \sqrt{n}(\hat{F}_{1,n}(.|\Delta X_i = 0) - F_{1,n}(.|\Delta X_i = 0) - (\hat{F}_{2,n}(., \hat{\Lambda}_n|\Delta X_i = 0) - F_{2,n}(., \Lambda_n|\Delta X_i = 0)))\end{aligned}$$

Now note that

$$\begin{aligned}F_{2,n}(.|\Delta X_i = 0) &= \sum_{k=1}^{K} P_n(X_{i1} = X_{i2} = x^k) F_{2,n}(. + \lambda(x)|X_{i1} = X_{i2} = x^k)\\ F_2(.|\Delta X_i = 0) &= \sum_{k=1}^{K} P(X_{i1} = X_{i2} = x^k) F_2(. + \lambda(x)|X_{i1} = X_{i2} = x^k) \qquad (72)\end{aligned}$$

Since Assumption 4.1 holds, it follows that, for $x \in \mathcal{X}$, $F_2(. + \lambda(x)|X_{i1} = X_{i2} = x)$ is

Hadamard differentiable tangentially to $\mathbb{D} \times \mathcal{Y}$ by Lemma D.1, where $\mathbb{D} = \{g \in \mathcal{L}^\infty(\mathcal{F}) : g \text{ is } \rho_2\text{-uniformly continuous}\}$, where ρ_2 is the variance metric. The Hadamard derivative is given by $\phi'_{F_2(.|x),\lambda(x)}(g,\epsilon) = g(.+\lambda(x)) + \epsilon f_2(.+\lambda(x)|X_{i1} = X_{i2} = x)$, where the subscript $F_2(.|x)$ denotes $F_2(.|X_{i1} = X_{i2} = x)$. Now $F_1(.|X_{i1} = X_{i2} = x) - F_2(.+\lambda(x)|X_{i1} = X_{i2} = x)$ is trivially Hadamard differentiable tangentially to $\mathbb{D} \times \mathcal{L}^\infty(\mathcal{Y}) \times \mathcal{Y}$.

Let $F_{t|x}(.) = F_t(.|X_{i1} = X_{i2} = x)$ and $F_{t|x,n}(.)$ be the sample analogue thereof. Now noting that

$$\sqrt{n}\begin{pmatrix} F_{1|.,n}(.) - F_{1|.}(.) \\ F_{2|.,n}(.) - F_{2|.}(.) \\ \Lambda_n - \Lambda \end{pmatrix} \rightsquigarrow \begin{pmatrix} \mathbb{G}_{1|.} \\ \mathbb{G}_{2|.} \\ \mathcal{E} \end{pmatrix} \tag{73}$$

where $\mathbb{G}_{1|.}$ and $\mathbb{G}_{2|.}$ are each $K \times 1$ tight Brownian bridges on $\{\mathcal{L}^\infty(\mathcal{Y})\}^K$ and $\mathcal{E}$ is a K-dimensional normal random vector. For the following, we define $\mathcal{E}(x)$ as follows, $\sqrt{n}(\lambda_n(x) - \lambda(x)) \xrightarrow{p} \mathcal{E}(x)$.

By Theorem 3.9.4 in Van der Vaart and Wellner (2000), it follows that

$$\sqrt{n}\left(F_{1|.,n}(.) - F_{2|.,n}(.+\lambda_n(.)) - (F_{1|.}(.) - F_{2|.}(.+\lambda(.)))\right) \mapsto \mathbb{G}_{1,2|.}, \tag{74}$$

where $\mathbb{G}_{1,2|x} = \mathbb{G}_{1|x} - \phi'_{F_{2|x},\lambda(x)}(\mathbb{G}_{2|x}, \mathcal{E}(x))$, which is a tight Brownian bridge.

To show the weak convergence of the bootstrap empirical process, given below, we have to check that the conditions in Theorem 3.9.11 in Van der Vaart and Wellner (2000),

$$\sqrt{n}\left(\hat{F}_{1|.,n}(.) - \hat{F}_{2|.,n}(.+\hat{\lambda}_n(.)) - (F_{1|.,n}(.) - F_{2|.,n}(.+\lambda_n(.)))\right) \mapsto \mathbb{G}_{1,2|.}. \tag{75}$$

The conditions in Theorem 3.9.11 include (a) Hadamard-differentiability tangentially to a subspace $\mathbb{D} \times \mathcal{L}^\infty(\mathcal{Y}) \times \mathcal{Y}$, (b) the underlying empirical processes converge to a separable limit, and (c) Condition (3.9.9), p. 378, in Van der Vaart and Wellner holds in outer probability. (a) follows from the above. Now (b) follows by (73) and tightness, since the latter implies separability. Finally, (c) is fulfilled if the conditions of Theorem 3.6.2 hold. We can consider $(\mathbb{G}_{1|.}, \mathbb{G}_{2|.}, \mathcal{E})$ as a tight Brownian bridge on $\mathcal{L}^\infty(\mathcal{F}) \times \mathcal{L}^\infty(\mathcal{F}) \times \mathcal{Y}$, where $\mathcal{F} = \{1\{y \leq t\} : t \in \mathcal{Y}\}$. Note that $\mathcal{E}$ is finite-dimensional, it suffices to show that Theorem 3.6.2 applies to $\mathcal{F}$. Since $\mathcal{F}$ is clearly Donsker and

$\sup_{f\in\mathcal{F}}|\int(f-\int f dF_t(y|X_{i1}=X_{i2}=x))^2 dF_t(y|X_{i1}=X_{i2}=x)|<\infty$, the conditions in Theorem 3.6.2. hold. Thus, (c) holds, which implies (75).

Now we relate (74) and (75) to $\mathbb{G}_{n|\Delta X_i=0}$ and $\hat{\mathbb{G}}_{n|\Delta X_i=0}$, respectively. Let $F_{1|x,n}(.)\equiv F_{1,n}(.|X_{i1}=X_{i2}=x)$.

$$
\begin{aligned}
& \mathbb{G}_{n|\Delta X_i=0} \\
= \;& \sqrt{n}(F_{1,n}(.|\Delta X_i)-F_{2,n}(.,\Lambda_n|\Delta X_i=0)-(F_1(.|\Delta X_i=0)-F_2(.,\Lambda|\Delta X_i=0))) \\
= \;& \sqrt{n}\sum_{k=1}^{K}P_n(X_{i1}=X_{i2}=x^k|\Delta X_i=0)\left(F_{1|x^k,n}(.)-F_{2|x^k,n}(.+\lambda_n(x^k))-(F_{1|x^k}(.)-F_{2|x^k}(.+\lambda(x^k)))\right) \\
= \;& \sqrt{n}\sum_{k=1}^{K}P(X_{i1}=X_{i2}=x^k|\Delta X_i=0)\left(F_{1|x^k,n}(.)-F_{2|x^k,n}(.+\lambda_n(x^k))-(F_{1|x^k}(.)-F_{2|x^k}(.+\lambda(x^k)))\right) \\
+ \;& \sum_{k=1}^{K}(P_n(X_{i1}=X_{i2}=x^k|\Delta X_i=0)-P(X_{i1}=X_{i2}=x^k|\Delta X_i=0)) \\
& \times\sqrt{n}\left(F_{1|x^k,n}(.)-F_{2|x^k,n}(.+\lambda_n(x^k))-(F_{1|x^k}(.)-F_{2|x^k}(.+\lambda(x^k)))\right) \qquad (76) \\
\rightsquigarrow \;& \sum_{k=1}^{K}P(X_{i1}=X_{i2}=x^k|\Delta X_i=0)(\mathbb{G}_{1|x^k}-\phi'_{F_2,\lambda(x^k)}(\mathbb{G}_{2|x^k},\mathcal{E}(x^k)))\equiv\mathbb{H}(\Lambda) \qquad (77)
\end{aligned}
$$

since the second term of the last equality converges to zero in probability by the weak convergence of (76) and $P_n(X_{i1}=X_{i2}=x^k|\Delta X_i=0)\xrightarrow{p}P(X_{i1}=X_{i2}=x^k|\Delta X_i=0)$ for all $k=1,2,...,K$. Hence, we have shown that the empirical process, $\mathbb{G}_{n|\Delta X_i=0}$ converges to a tight Brownian bridge.

Now the bootstrap empirical process can be decomposed as follows

$$
\begin{aligned}
& \hat{\mathbb{G}}_{n|\Delta X_i=0} \\
= \;& \sqrt{n}\left(\hat{F}_{1,n}(.|\Delta X_i=0)-\hat{F}_{2,n}(.,\hat{\Lambda}_n|\Delta X_i=0)-(F_{1,n}(.|\Delta X_i=0)-F_{2,n}(.,\Lambda_n|\Delta X_i=0))\right) \\
= \;& \sqrt{n}\sum_{k=1}^{K}\hat{P}_n(X_{i1}=X_{i2}=x^k|\Delta X_i=0) \\
& \times\left(\hat{F}_{1|x^k,n}(.+\hat{\lambda}_n(x^k))-\hat{F}_{2|x^k,n}(.)-(F_{1|x^k,n}(.+\lambda_n(x^k))-F_{2|x^k,n}(.))\right)
\end{aligned}
$$

$$
\begin{aligned}
&= \sqrt{n}\sum_{k=1}^{K} P(X_{i1}=X_{i2}=x^k|\Delta X_i=0) \\
&\quad \times\left(\hat{F}_{1|x^k,n}(.)-\hat{F}_{2|x^k,n}(.+\hat{\lambda}_n(x^k))-(F_{1|x^k,n}(.)-F_{2|x^k,n}(.+\lambda_n(x^k)))\right) \\
&+ \sum_{k=1}^{K}(\hat{P}_n(X_{i1}=X_{i2}=x^k|\Delta X_i=0)-P(X_{i1}=X_{i2}=x^k|\Delta X_i=0)) \\
&\quad \times\sqrt{n}\left(\hat{F}_{1|x^k,n}(.)-\hat{F}_{2|x^k,n}(.+\hat{\lambda}_n(x^k))-(F_{1|x^k,n}(.)-F_{2|x^k,n}(.+\lambda_n(x^k)))\right) \\
&\rightsquigarrow \mathbb{H}(\Lambda) \qquad (78)
\end{aligned}
$$

where the first term of the last equality follows by continuous mapping theorem and (**??**). The second term converges to zero by (75) and $(\hat{P}_n(X_{i1}=X_{i2}=x^k|\Delta X_i=0)-P(X_{i1}=X_{i2}=x^k|\Delta X_i=0))\xrightarrow{p}0$ by Lemma C.1. Thus, the bootstrap empirical process converges to the same tight Brownian bridge as the empirical process. (i) and (ii) follow by the same arguments as Theorem 4.1. □

Proof (Theorem 4.3)

We first have to show that the underlying empirical and bootstrap empirical processes converge to the same tight Brownian bridge. Let $m_l=|\{\underline{x}\in\mathcal{X}_l:\underline{x}\in\mathcal{X}\}|$, where $|S|$ denotes the cardinality of a set S. Our statistics can be written as follows:

$$
\begin{aligned}
KS_{n,\mathcal{Y}} &= \frac{1}{T}\sum_{t=1}^{T}\sum_{l=1}^{L}P_n(h(X_i)=h_l)\sum_{j=1}^{m_l}P_n(X_i=\underline{x}_j|h(X_i)=h_l)\|\sqrt{n}\{F_{t,n}(.|\underline{x}_j)-\bar{F}_{t,n,l}(.)\}\|_{\infty,\mathcal{Y}} \\
CM_{n,\phi} &= \frac{1}{T}\sum_{t=1}^{T}\sum_{l=1}^{L}P_n(h(X_i)=h_l)\sum_{j=1}^{m_l}P_n(X_i=\underline{x}_j|h(X_i)=h_l)\|\sqrt{n}\{F_{t,n}(.|\underline{x}_j)-\bar{F}_{t,n,l}(.)\}\|_{2,\phi}.
\end{aligned}
$$

Let $(\zeta(.,\underline{x}))_{\underline{x}\in\mathcal{X}^T}$ be the vector that contains the elements $\{\zeta(.,\underline{x}):\underline{x}\in\mathcal{X}^T\}$.

$$
\begin{pmatrix}
\sqrt{n}\left(F_{1,n}(.|X_i=\underline{x})-F_1(.|X_i=\underline{x})\right)_{\underline{x}\in\mathcal{X}^T} \\
\sqrt{n}\left(F_{2,n}(.|X_i=\underline{x})-F_2(.|X_i=\underline{x})\right)_{\underline{x}\in\mathcal{X}^T} \\
\cdots \\
\sqrt{n}\left(F_{T,n}(.|X_i=\underline{x})-F_T(.|X_i=\underline{x})\right)_{\underline{x}\in\mathcal{X}^T}
\end{pmatrix}
\rightsquigarrow \mathbb{G} \qquad (79)
$$

Since $T<\infty$ and $|\mathcal{X}^T|<\infty$, the joint distribution of the centered empirical conditional cdfs converges to a tight Brownian bridge. Now note that a linear combination of the

above yields the empirical process that we use to construct our statistics.

$$\left(\left(\begin{array}{c} \sqrt{n}(F_{1,n}(.|\underline{x}_j) - \bar{F}_{1,n,l}(.) - (F_1(.|\underline{x}_j) - \bar{F}_{1,l}(.))) \\ ... \\ \sqrt{n}(F_{T,n}(.|\underline{x}_j) - \bar{F}_{T,n,l}(.) - (F_T(.|\underline{x}_j) - \bar{F}_{T,l}(.))) \end{array} \right)_{j=1,2,...,m_l} \right)_{l=1,2,...,L} \rightsquigarrow \mathbb{H},$$

where $\mathbb{H} \equiv ((\mathbb{H}_{j,l})_{j=1,..,m_l})_{l=1,..,L}$. Note that the above process is a $(T\sum_{l=1}^{L} m_l)\times 1$ vector of functionals. Since all of the above processes are defined on a Donsker class, the bootstrap empirical process also converges to the same limit process by Theorem 3.6.1 in Van der Vaart and Wellner (2000).

$$\left(\left(\begin{array}{c} \sqrt{n}(\hat{F}_{1,n}(.|\underline{x}_j) - \hat{\bar{F}}_{1,n,l}(.) - (F_{1,n}(.|\underline{x}_j) - \bar{F}_{1,n,l}(.))) \\ ... \\ \sqrt{n}(\hat{F}_{T,n}(.|\underline{x}_j) - \hat{\bar{F}}_{T,n,l}(.) - (F_{T,n}(.|\underline{x}_j) - \bar{F}_{T,n,l}(.))) \end{array} \right)_{j=1,2,...,m_l} \right)_{l=1,2,...,L} \rightsquigarrow \mathbb{H}.$$

Now we give the limiting statistics as follows,

$$\begin{aligned} KS_{\mathcal{Y}} &= \frac{1}{T}\sum_{t=1}^{T}\sum_{l=1}^{L} P(h(X_i) = h_l) \sum_{j=1}^{m_l} P(X_i = \underline{x}|h(X_i) = h_l)\|\mathbb{H}_{j,l}\|_{\infty,\mathcal{Y}} \\ CM_{\phi} &= \frac{1}{T}\sum_{t=1}^{T}\sum_{l=1}^{L} P(h(X_i) = h_l) \sum_{j=1}^{m_l} P_n(X_i = \underline{x}|h(X_i) = h_l)\|\mathbb{H}_{j,l}\|_{2,\phi}. \end{aligned} \tag{80}$$

Since the above is a linear combination of convex continuous functionals, it follows that Theorem 11.1 in Davydov, Lifshits, and Smorodina (1998) applies. Thus, the distributions of $KS_{\mathcal{Y}}$ and CM_{ϕ} are absolutely continuous and strictly increasing on $(0,\infty)$. Since $F_t(.|X_i = \underline{x})$ is non-degenerate for $\underline{x} \in \mathcal{X}^T$ and $t = 1, 2..., T$, it follows that the $P(KS_{\mathcal{Y}} = 0) = 0$ and $P(CM_{\phi} = 0) = 0$. Hence, it follows that their distribution is absolutely continuous on $[0,\infty)$. Now it remains to show that $KS_{n,\mathcal{Y}}$ and $CM_{n,\phi}$ converge to $KS_{\mathcal{Y}}$ and CM_{ϕ}, respectively.

Let T_n with norm $\|.\|$ denote either the KS or CM with their respective norms, and let

$\mathbb{H}_{n,j,l}$ denote the relevant empirical process

$$\begin{aligned}
T_n &= \frac{1}{T}\sum_{t=1}^{T}\sum_{l=1}^{L} P_n(h(X_i)=h_l)\sum_{j=1}^{m_l} P_n(X_i=\underline{x}_j|h(X_i)=h_l)\|\mathbb{H}_{n,j,l}\| \\
&= \frac{1}{T}\sum_{t=1}^{T}\sum_{l=1}^{L} P(h(X_i)=h_l)\sum_{j=1}^{m_l} P(X_i=\underline{x}|h(X_i)=h_l)\|\mathbb{H}_{n,j,l}\| \\
&+ \frac{1}{T}\sum_{t=1}^{T}\sum_{l=1}^{L} P_n(h(X_i)=h_l)\sum_{j=1}^{m_l}(P_n(X_i=\underline{x}|h(X_i)=h_l)-P(X_i=\underline{x}|h(X_i)=h_l))\|\mathbb{H}_{n,j,l}\| \\
&+ \frac{1}{T}\sum_{t=1}^{T}\sum_{l=1}^{L}(P_n(h(X_i)=h_l)-P(h(X_i)=h_l))\sum_{j=1}^{m_l} P_n(X_i=\underline{x}|h(X_i)=h_l)\|\mathbb{H}_{n,j,l}\| \\
&\rightsquigarrow T
\end{aligned}$$

where T equals $KS_{\mathcal{Y}}$ and CM_{ϕ} for the KS and CM statistics, respectively. The convergence follows since the latter two terms converge in probability to zero, since $(P_n(X_i = \underline{x}|h(X_i)=h_l) - P(X_i=\underline{x}|h(X_i)=h_l)) \xrightarrow{p} 0$ and $(P_n(h(X_i)=h_l) - P(h(X_i)=h_l)) \xrightarrow{p} 0$, and both terms are multiplied by $O_p(1)$ terms.

Now it follows that the tests based on the bootstrap critical values yield correct asymptotic size, which gives (i). Since the empirical processes on which the tests are based are all tight, the critical values of the tests are bounded in probability. Hence, the tests are consistent against any fixed alternatives, which yields (ii). □

C Supplementary Results: Testing Identifying Restrictions

Theorem C.1 *Given two random samples of size n and m from $F_X(.)$ and $F_Y(.)$, respectively, where $F_X(.)$ and $F_Y(.)$ are atomless distribution functions, $D_{n_1,m}$ is distribution-free, where*

$$D_{n,m} = \sup_{z\in\mathcal{Z}} |\hat{F}_X(z) - \hat{F}_Y(z)|, \tag{81}$$

Proof Note that $D_{n,m} = \max\{D^+_{n,m}, D^-_{n,m}\}$, where $D^+_{n,m} = \sup_{z\in\mathcal{Z}}[\hat{F}_X(z) - \hat{F}_Y(z)]$ and $D_{n,m} = \sup_{z\in\mathcal{Z}}[\hat{F}_Y(z) - \hat{F}_X(z)]$. Let $\{Z_i\}_{i=1}^{n+m}$ be the combined observations from both

samples, and let $Z_{(i)}$ be the i^{th} order statistic and add $Z_{(0)} = -\infty$ and $Z_{(n+m+1)} = \infty$. Also, $r_i \equiv \sum_{j=1}^{n}\{X_j \leq Z_{(i)}\}$ and $s_i \equiv \sum_{j=1}^{m}\{Y_j \leq Z_{(i)}\}$.

$$\begin{aligned}
D^{+}_{n,m} &= \max_{0\leq i\leq n+m} \sup_{Z_{(i)}\leq y<Z_{(i+1)}} \left[\hat{F}_1(y) - F_X(y) - (\hat{F}_2(y) - F_2(y)) + F_X(y) - F_2(y)\right] \\
&= \max_{0\leq i\leq n+m} \left[\frac{r_i}{n} - \inf_{Z_{(i)}\leq y<Z_{(i+1)}} F_X(y) - \left(\frac{s_{i-1}}{m} - \sup_{Z_{(i)}\leq y<Z_{(i+1)}} F_2(y)\right)\right] \\
&\quad + \max_{0\leq i\leq n+m} \left[\sup_{Z_{(i)}\leq y<Z_{(i+1)}} (F_X(y) - F_2(y))\right] \\
&= \max_{0\leq i\leq n+m} \left[\frac{r_i}{n} - F_X(Z_{(i)}) - \left(\frac{s_{i-1}}{m} - F_2(Z_{(i)})\right)\right] \\
&\quad + \max_{0\leq i\leq n+m} \left[F_X(Z_{(i)}) - F_2(Z_{(i)})\right] \\
&= \max[\max_{1\leq i\leq n+m} \left[\frac{r_i}{n} - F_X(Z_{(i)}) - \left(\frac{s_{i-1}}{m} - F_Y(Z_{(i)})\right)\right] \\
&\quad + \max_{1\leq i\leq n+m} \left[F_X(Z_{(i)}) - F_Y(Z_{(i)})\right], 0] \qquad (82)
\end{aligned}$$

By the probability-integral transform theorem and continuity, $F_X(Z_{(j)})$ and $F_Y(Z_{(j)})$ are both j^{th} order statistics from $U(0,1)$ variables, for $j = 1, 2$. hence $D^{+}_{n,m}$ is distribution-free. Similarly,

$$\begin{aligned}
D^{-}_{n,m} &= \max[\max_{1\leq i\leq n} \left[\frac{s_i}{m} - F_Y(Z_{(i)}) - (\frac{r_{i-1}}{n} - F_X(Z_{(i)}))\right] \\
&\quad - \min_{1\leq i\leq n} \left[(F_X(Z_{(i)}) - F_Y(Z_{(i)}))\right], 0]
\end{aligned}$$

$$\begin{aligned}
D_{n,m} &= \max[\max_{1\leq i\leq n} \left[\frac{s_i}{m} - F_Y(Z_{(i)}) - \left(\frac{r_{i-1}}{n} - F_X(Z_{(i)})\right)\right] \\
&\quad + \max[\max_{1\leq i\leq n} \left[F_X(Z_{(i)}) - F_Y(Z_{(i)})\right], \\
&\quad \max_{1\leq i\leq n+m} \left[\frac{r_i}{n} - F_X(Z_{(i)}) - \left(\frac{s_i}{m} - F_Y(Z_{(i)})\right)\right]] \\
&\quad + \max_{1\leq i\leq n+m} \left[(F_X(Z_{(i)}) - F_Y(Z_{(i)}))\right], 0]
\end{aligned}$$

□

Lemma C.1 *For an iid sequence of events,* $\{A_i\}_{i=1}^{n}$, $P(A_i) > 0$, $\hat{P}_n(A_i) \equiv \sum_{i=1}^{n} M_{ni} 1\{A\}/n \xrightarrow{p} P(A)$ *as* $n \to \infty$.

Proof Note that $M_{ni} \sim Bin(n, n^{-1})$ independent of $\{A_i\}_{i=1}^n$. Now,

$$E\left[\frac{\sum_{i=1}^n M_{ni} 1\{A_i\}}{n}\right] = \frac{\sum_{i=1}^n 1\{A_i\}}{n} \tag{83}$$

Now since M_{ni} is not iid across i, we cannot apply a law of large numbers directly. We use the Poissonized process, where $\{M_{N_n,i}\}_{i=1}^n$ are iid Poisson variables with mean 1.

$$\hat{P}_n(A_i) = \frac{\sum_{i=1}^n (M_{ni} - M_{N_n,i}) 1\{A_i\}}{n} + \frac{\sum_{i=1}^n M_{N_n,i} 1\{A_i\}}{n} \tag{84}$$

So we can apply a weak law of large numbers to the latter term to show convergence to $P(A_i)$, since given $\{(Y_i, X_i')\}_{i=1}^n$ and N_n

$$E\left[\frac{\sum_{i=1}^n M_{N_n,i} 1\{A_i\}}{n}\right] = \frac{\sum_{i=1}^n 1\{A_i\}}{n} \tag{85}$$

Since $E[N_n] = n$ and it is independent of $\{A_i\}_{i=1}^n$, by the law of large numbers

$$\frac{\sum_{i=1}^n M_{N_n,i} 1\{A_i\}}{N_n} \xrightarrow{p} P(A_i) \tag{86}$$

It remains to show that the residual term $\frac{\sum_{i=1}^n (M_{ni} - M_{N_n,i}) 1\{\Delta X_i = 0\}}{n}$ converges to zero in probability. From the proof of Theorem 3.6.2. in Van der Vaart and Wellner (2000), we have that

$$P\left(\max_{1 \le i \le n} |M_{N_n,i} - M_{n,i}| > 2\right) \to \epsilon \tag{87}$$

So we can write

$$\begin{aligned} & \frac{\sum_{i=1}^n |M_{N_n,i} - M_{ni}|}{n} 1\{\max_{1 \le i \le n} |M_{N_n,i} - M_{n,i}| > 2\} + \frac{\sum_{i=1}^n |M_{N_n,i} - M_{ni}|}{n} 1\{\max_{1 \le i \le n} |M_{N_n,i} - M_{n,i}| \le 2\} \\ = \; & o_p(1) + \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^{\infty} 1\{|M_{N_n,i} - M_{ni}| \le j\} 1\{\max_{1 \le i \le n} |M_{N_n,i} - M_{n,i}| \le 2\} \end{aligned} \tag{88}$$

Now we can write the difference between the bootstrap empirical process and the Pois-

sonized process as follows

$$\begin{aligned}
& \frac{\sum_{i=1}^{n}(M_{n,i}-M_{ni})1\{\Delta X_i=0\}}{n} \\
\leq\ & \frac{\sum_{i=1}^{n}|M_{n,i}-M_{ni}|1\{\Delta X_i=0\}}{n} \\
=\ & \frac{\sum_{i=1}^{n}|M_{N_n,i}-M_{ni}|}{n}1\{\max_{1\leq i\leq n}|M_{N_n,i}-M_{n,i}|>2\}1\{\Delta X_i=0\} \\
& +\frac{\sum_{i=1}^{n}|M_{N_n,i}-M_{ni}|}{n}1\{\max_{1\leq i\leq n}|M_{N_n,i}-M_{n,i}|\leq 2\}1\{\Delta X_i=0\}
\end{aligned}$$

$$\begin{aligned}
\leq\ & o_p(1)+\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{\infty}1\{|M_{N_n,i}-M_{ni}|\leq j\}1\{\max_{1\leq i\leq n}|M_{N_n,i}-M_{n,i}|\leq 2\} \\
=\ & o_p(1)+\frac{1}{n}\sum_{j=1}^{\infty}\frac{\#I_n^j}{n}\left(\frac{1}{\#I_n^j}\sum_{i\in I_n^j}1\{\Delta X_i=0\}\right) \\
\leq\ & o_p(1)+\frac{1}{n}\sum_{j=1}^{2}\frac{\#I_n^j}{n}\left(\frac{n}{\sum_{i=1}^{n}1\{|M_{N_n,i}-M_{n,i}|\geq j\}}\frac{\sum_{i=1}^{n}1\{\Delta X_i=0\}}{n}\right) \\
=\ & o_p(1)+\frac{O_p(n^{1/2})}{n}O_p(1)=o_p(1) \qquad (89)
\end{aligned}$$

where $I_n^j=\{i\in\{1,2,...,n\}:|M_{N_n,i}-M_{ni}|\geq j\}$. Now the penultimate equality follows from $\#I_n^j\leq|N_n-n|=O_p(n^{1/2})$ for all j. Now the second term follows by the law of large numbers and $P(|M_{N_n,i}-M_{n,i}|=0,\hat{\tau},\tilde{\tau})=1-2\epsilon$, such that $P(|M_{N_n,i}-M_{n,i}|=j)>0$ for $j=0,1,2$. □

Lemma C.2 *Given Assumption 4.1,* $\rho_2(s+\epsilon,s)\leq K|\epsilon|$.

Proof (Lemma C.2)

$$\begin{aligned}
\rho_2(s+\epsilon,s) &= E[(\mathbb{G}(s+\epsilon)-\mathbb{G}(s))^2] \\
&= E[\mathbb{G}(s+\epsilon)^2|\epsilon]+E[\mathbb{G}(s)^2]-2E[\mathbb{G}(s+\epsilon)\mathbb{G}(s)] \\
&= |F(s+\epsilon)(1-F(s+\epsilon))+F(s)(1-F(s))-2\{F(\{s+\epsilon\}\wedge s)-F(s+\epsilon)F(s)\}| \\
&\leq |F(s+\epsilon)(1-F(s+\epsilon))-\{F(\{s+\epsilon\}\wedge s)-F(s+\epsilon)F(s)\}| \qquad (90) \\
&+ |F(s)(1-F(s))-\{F(\{s+\epsilon\}\wedge s)-F(s+\epsilon)F(s)\}| \qquad (91)
\end{aligned}$$

Now for (90)

$$\begin{aligned} & |F(s+\epsilon)(1-F(s+\epsilon)) - \{F(\{s+\epsilon\}\wedge s) - F(s+\epsilon)F(s)\}| \\ = \;& |F(s+\epsilon)(1-F(s+\epsilon)) - F(\{s+\epsilon\}\wedge s)(1-F(\{s+\epsilon\}\vee s))\}| \end{aligned}$$

$$\begin{aligned} \leq \;& |F(s+\epsilon)(F(\{s+\epsilon\}\vee s) - F(\{s+\epsilon\}\vee s)) + |(F(s+\epsilon) - F(\{s+\epsilon\}\wedge s))(1-F(\{s+\epsilon\}\vee s))| \\ \leq \;& |F(\{s+\epsilon\}\vee s) - F(s+\epsilon)| + |F(s+\epsilon) - F(\{s+\epsilon\}\wedge s)| \\ \leq \;& f(s)\{|\max(s+\epsilon, s) - (s+\epsilon)| + |s+\epsilon - \min(s+\epsilon, s)| \\ = \;& f(s)\{|\max(0,-\epsilon)| + |min(0,\epsilon)|\} \\ \leq \;& 2|f(s)||\epsilon| \end{aligned}$$

where $f(s)$ is the density of F. Similar manipulation of (91) yields

$$|F(s)(1-F(s)) - \{F(\{s+\epsilon\}\wedge s) - F(s+\epsilon)F(s)\}| \;\leq\; 2|f(s)||\epsilon_n|$$

Assumption 4.1 delivers the result with $K = 4\sup_{s\in\mathbb{R}}|f(s)|$. □

We need to introduce some notation for the following lemma. Let $\mathbb{G}_n \equiv \sqrt{n}(F_n(.) - F(.))$. $\mathbb{G}_{n_A} \equiv \sqrt{n}(F_n(.|A_i) - F(.|A_i))$ and $\hat{\mathbb{G}}_{n_A} \equiv \sqrt{n}(\hat{F}_n(.|A_i) - F_n(.|A_i))$. $\mathbb{G}_n \rightsquigarrow \mathbb{G}$, a $\mathbb{P}$-Brownian bridge in $\mathcal{L}^\infty(\mathcal{Y})$, where $\mathcal{Y}$ is the support of Y_i. Clearly,

Lemma C.3 *Given $\{Y_i, A_i\}_{i=1}^n$ an iid sequence, where A_i is an event with $P(A_i) > 0$, then*

(i) $\mathbb{G}_{n|A} \rightsquigarrow \mathbb{G}_A = \frac{\mathbb{H}}{P(A_i)} + \mathcal{Z}F(., A_i)$

(ii) $\hat{\mathbb{G}}_{n|A} \rightsquigarrow \mathbb{G}_A$

where $\mathbb{H}$ is $\mathbb{P}$-Brownian bridge in $\mathcal{L}^\infty(\mathcal{F})$, where $\mathcal{F} = \{1\{y \leq t\}1\{A_i\} : t \in \mathcal{Y}\}$, and $\sqrt{n}((P_n(A_i))^{-1} - (P(A_i))^{-1}) \rightsquigarrow \mathcal{Z}$.

Proof For (i), let $P_n(A) \equiv \sum_{i=1}^n 1\{A_i\}$.

$$\begin{aligned}
\mathbb{G}_{n_A} &= \sqrt{n}(F_n(.|A) - F(.|A)) \\
&= \sqrt{n}\left(\frac{F_n(.,A) - F(.,A)}{P_n(A)} + F(.,A)\left(\frac{1}{P_n(A)} - \frac{1}{P(A)}\right)\right) \\
&\rightsquigarrow \frac{\mathbb{H}}{P(A_i)} + \mathcal{Z}F(.,A_i) \equiv \mathbb{G}_A
\end{aligned}$$

The result for the first term follows by the Donsker property of $\mathcal{F}$ and Slutzky's theorem. The latter trivially follows by $P(A_i) > 0$ and the continuous mapping theorem.

For (ii), let $\hat{P}_n(A) \equiv \sum_{i=1}^n M_{ni} 1\{A_i\}/n$

$$\begin{aligned}
\hat{\mathbb{G}}_{n_A} &= \sqrt{n}(\hat{F}_n(.|A_i) - F_n(.|A_i)) \\
&= \sqrt{n}\left(\frac{\hat{F}_n(.,A_i)}{\hat{P}_n(A_i)} - \frac{F_n(.,A_i)}{P_n(A_i)}\right) \\
&= \sqrt{n}\left(\frac{\hat{F}_n(.,A_i) - F(.,A_i)}{\hat{P}_n(A_i)} + F_n(.,A_i)\left(\frac{1}{\hat{P}_n(A_i)} - \frac{1}{P_n(A_i)}\right)\right) \\
&= \frac{\hat{\mathbb{G}}_n}{\hat{P}_n(A_i)} + F_n(.,A_i)\sqrt{n}\left(\frac{1}{\hat{P}_n(A_i)} - \frac{1}{P_n(A_i)}\right) \\
&\rightsquigarrow \frac{\mathbb{H}}{P(A_i)} - F(.,A_i)\mathcal{Z} \qquad (92)
\end{aligned}$$

where the weak convergence of the first term follows by the Donsker property of $\mathcal{F}$, which implies that Theorem 3.6.1 of Van der Vaart and Wellner (2000) applies. The second term follows by the Glivenko-Cantelli property of $F(.,A_i)$ and the continuous mapping theorem. □

D Paired-Sample Problem

We observe an iid sequence $\{Y_{i1}, Y_{i2}\}_{i=1}^n$. We are interested in testing the equality of the distribution of the demeaned variables $W_{i,n} = \{Y_{i1} - \lambda_n, Y_{i2}\}$, where $\lambda_n = \bar{Y}_{1,n} = \sum_{i=1}^n (Y_{i1} - Y_{i2})/n$. Now we will introduce some notation. Let $F_1(.)$ and $F_2(.)$ denote the distributions of Y_{i1} and Y_{i2}, respectively. The following are their empirical versions

$$F_{1,n}(t, \lambda_n) = \frac{1}{n}\sum_{i=1}^n 1\{Y_{i1} - \lambda_n \leq t\} \qquad F_{2,n} = \frac{1}{n}\sum_{i=1}^n 1\{Y_{i2} \leq t\} \qquad (93)$$

and their bootstrap empirical versions

$$\hat{F}_{1,n}(t+\hat{\lambda}_n) = \frac{1}{n}\sum_{i=1}^{n} M_{ni} 1\{Y_{i1} - \hat{\lambda}_n \leq t\} \qquad \hat{F}_{2,n}(t) = \frac{1}{n}\sum_{i=1}^{n} M_{ni} 1\{Y_{i2} \leq t\} \tag{94}$$

Now we need to show weak convergence of the following empirical process

$$\mathbb{H}_n(\lambda_n) \quad = \quad \sqrt{n}(F_{1,n}(.+\lambda_n) - F_{2,n}(.) - (F_1(.+\lambda) - F_2(.))) \tag{95}$$

and its bootstrap empirical counterpart,

$$\hat{\mathbb{H}}_n(\hat{\lambda}_n) \quad = \quad \sqrt{n}(\hat{F}_{1,n}(.+\hat{\lambda}) - \hat{F}_{2,n}(.) - (F_{1,n}(.,\lambda) - F_{2,n}(.))). \tag{96}$$

Now to show weak convergence of the above process, we need to show that the delta method applies here to the above empirical process, such that we can use the convergence of the following:

$$\sqrt{n}\begin{pmatrix} F_{1,n}(.) - F_1(.) \\ F_{2,n}(.) - F_2(.) \\ \lambda_n - \lambda \end{pmatrix} \quad \rightsquigarrow \quad \begin{pmatrix} \mathbb{G}_1 \\ \mathbb{G}_2 \\ \mathcal{E} \end{pmatrix} \tag{97}$$

The main complication is due to the evaluation of the distribution function at Y_{i1} after imposing a location shift. Let $\mathcal{Y}$ denote the support Y_{i1}. We first proceed to showing that the map $\phi : \mathcal{L}^\infty(\mathcal{Y}) \times \mathcal{Y} \mapsto \mathcal{L}^\infty(\mathcal{Y})$, which is given below, is Hadamard differentiable. Let G denote some distribution and γ a location shift.

$$\phi(G,\gamma) \quad = \quad G(.-\gamma) \tag{98}$$

We need to impose that the underlying distribution has a bounded density. For a Gaussian process, $\mathbb{G}$, define $\rho_2(s,t) = E|\mathbb{G}(s) - \mathbb{G}(t)|^2$ and $\mathbb{D} \equiv \{h \in \mathcal{L}^\infty(\mathcal{Y}) : \text{h is } \rho_2\text{-uniformly continuous}\}$.

Lemma D.1 *Assume that the distribution function F has a bounded density, $\phi(F,\lambda) : \mathcal{L}^\infty(\mathcal{Y}) \times \mathcal{Y} \mapsto \mathcal{L}^\infty(\mathcal{Y})$ is Hadamard differentiable at (F,λ) tangentially to $\mathbb{D} \times \mathcal{Y}$ with the*

following derivative

$$\phi'_{F,\lambda}(g,\epsilon) \quad = \quad g(.-\lambda) - \epsilon f(.-\lambda),$$

for $(g,\epsilon) \in \mathbb{D} \times \mathbb{R}$.

Proof Note that $\phi'_{F,\lambda}(g,\epsilon)$ is clearly linear and continuous in g and ϵ. Now we need to show that

$$\frac{\phi(F+\tau_n g_n, \lambda+\tau_n\epsilon_n) - \phi(F,\lambda)}{\tau_n} \quad \to \quad \phi'_{F,\lambda}(g,\epsilon) \qquad n \to \infty \tag{99}$$

for all converging sequences $\tau_n \searrow 0$, $g_n \to g \in \mathbb{D}$, and $\epsilon_n \to \epsilon \in \mathbb{R}$.

$$\begin{aligned}
& \left\| \frac{\phi(F+\tau_n g_n, \lambda+\tau_n\epsilon_n) - \phi(F,\lambda)}{\tau_n} - \phi'_{F,\lambda}(g,\epsilon) \right\|_\infty \\
= & \left\| \frac{(F+\tau_n g_n)(.-\lambda-\tau_n\epsilon_n) - F(.-\lambda)}{\tau_n} - (g(.-\lambda) - \epsilon f(.-\lambda)) \right\|_\infty \\
\leq & \left\| \frac{F(.-\lambda-\tau_n\epsilon_n) - F(.-\lambda)}{\tau_n} + \epsilon f(.-\lambda) \right\|_\infty + \|g_n(.-\lambda-\tau_n\epsilon_n) - g(.-\lambda)\|_\infty
\end{aligned} \tag{100}$$

For the first term of (100),

$$\begin{aligned}
& \left\| \frac{F(.-\lambda-\tau_n\epsilon_n) - F(.-\lambda)}{\tau_n} + \epsilon f(.-\lambda) \right\|_\infty \\
\leq & \left\| \frac{F(.-\lambda-\tau_n\epsilon_n) - F(.-\lambda)}{\tau_n} - \frac{F(.-\lambda-\tau_n\epsilon) - F(.-\lambda)}{\tau_n} \right\|_\infty \\
+ & \left\| \frac{F(.-\lambda-\tau_n\epsilon) - F(.-\lambda)}{\tau_n} + \epsilon f(.-\lambda) \right\|_\infty
\end{aligned}$$

$$\begin{aligned}
\leq & \sup_{t\in\mathbb{R}} |f(t)||\epsilon_n - \epsilon| + |\epsilon| \sup_{t\in\mathbb{R}} \left| \frac{F(t-\lambda-\tau_n\epsilon) - F(t-\lambda)}{\epsilon\tau_n} - f(t-\lambda) \right| \\
= & \sup_{t\in\mathbb{R}} |f(t)||\epsilon_n - \epsilon| + |\epsilon| \sup_{t\in\mathbb{R}} |f(w) - f(t-\lambda)| \;\; w \in (t-\lambda-\tau_n\epsilon, t-\lambda) \qquad (101) \\
\leq & \sup_{t\in\mathbb{R}} |f(t)||\epsilon_n - \epsilon| + |\epsilon\delta| \qquad (102)
\end{aligned}$$

The penultimate equality follows by the Mean Value Theorem.[38] The second term of the

[38]By the Mean Value Theorem, uniform continuity is equivalent to uniform differentiability.

last equality follows by uniform continuity implied by Assumption 4.1, which implies that for $|w-t-\lambda| < |\tau_n \epsilon| < \nu$, there is a δ that bounds the second term in (101) uniformly in t. Now as $\tau_n \to 0$ and $\epsilon_n \to \epsilon$, both terms converge to zero.

As for the second term of (100)

$$\begin{aligned}
& \|g_n(.-\lambda-\tau_n\epsilon_n) - g(.-\lambda)\|_\infty \\
\leq\ & \|g_n(.-\lambda-\tau_n\epsilon_n) - g(.-\lambda-\tau_n\epsilon_n)\|_\infty + \sup_{t\in\mathbb{R}} |g(t-\lambda-\tau_n\epsilon_n) - g(t+\lambda)| \\
\leq\ & \|g_n(.-\lambda-\tau_n\epsilon_n) - g(.-\lambda-\tau_n\epsilon_n)\|_\infty + \sup_{\rho_2(t-\lambda-\tau_n\epsilon_n, t-\lambda)} |g(t-\lambda-\tau_n\epsilon_n) - g(t-\lambda)| \\
\to\ & 0 \qquad\qquad \tau_n \to 0, \epsilon_n \to \epsilon, g_n \to g
\end{aligned} \tag{103}$$

where the first term follows by weak convergence for all $g \in \mathbb{D}$. As for the latter term, by Lemma C.2 $\tau_n \to 0$ implies that $\rho_2(t-\lambda-\tau_n\epsilon_n, t-\lambda) \to 0$. Since $g \in \mathbb{D}$, which is uniformly ρ_2-continuous, the result follows. □

Theorem D.1 *Assume that the distribution function F_1 has bounded density,*

(i) $\mathbb{H}_n(\lambda_n) \rightsquigarrow \mathbb{H}(\lambda)$,

(ii) $\hat{\mathbb{H}}_n(\hat{\lambda}_n) \rightsquigarrow \mathbb{H}(\lambda)$

where $\mathbb{H}(\lambda)$ is a tight Brownian bridge in $L^\infty(\mathcal{Y})$.

Proof (i) Note that

$$\sqrt{n}\begin{pmatrix} F_{1,n}(.) - F_1(.) \\ F_{2,n}(.) - F_2(.) \\ \lambda_n - \lambda \end{pmatrix} \rightsquigarrow \begin{pmatrix} \mathbb{G}_1 \\ \mathbb{G}_2 \\ \mathcal{E} \end{pmatrix} \tag{104}$$

where $\mathbb{G}_1$ and $\mathbb{G}_2$ are tight Brownian bridges in $L^\infty(\mathcal{Y})$ and $\mathcal{E}$ is a normal scalar random variable.

Now let $\phi(F, \lambda) = F_1(.-\lambda)$. By Lemma D.1, $\phi : \mathcal{L}^\infty(\mathcal{Y}) \times \mathbb{R} \mapsto L^\infty(\mathbb{R})$ is Hadamard differentiable at F, λ tangentially to $\mathbb{D} \times \mathbb{R}$. Thus, by Theorem 3.9.4. in Van der Vaart and Wellner (2000), the delta method applies and

$$\sqrt{n}\begin{pmatrix} \phi(F_{1,n}, \lambda_n) - \phi(F_1, \lambda) \\ F_{2,n} - F_2 \end{pmatrix} \rightsquigarrow \begin{pmatrix} \phi'_{F,\lambda}(\mathbb{G}_1, \mathcal{E}) \\ \mathbb{G}_2 \end{pmatrix}$$

The result follows by the continuous mapping theorem.

(ii) It suffices to check the conditions of Theorem 3.9.11 in Van der Vaart and Wellner (2000), which are that the empirical process in (104), where $(\mathbb{G}_1, \mathbb{G}_2, \mathcal{E})$ are separable and in $\mathbb{D} \times L^\infty(\mathcal{Y}) \times \mathcal{Y}$. Since $\mathbb{G}_1$ and $\mathbb{G}_2$ are tight Brownian bridges and $\mathcal{E}$ is a scalar normal random variable, which implies that the limiting process is separable and by assumption it is in $\mathbb{D} \times \mathcal{L}^\infty(\mathcal{Y}) \times \mathcal{Y}$.

Furthermore, we need condition (3.9.9), p. 378, in Van der Vaart and Wellner (2000) to hold in probability. Theorem 3.6.2 implies that (3.9.9) holds almost surely, and hence in probability, if $\mathcal{F}$ is Donsker and $\|P(f - Pf)\|_{\mathcal{F}} < \infty$. Since $\mathcal{F} = \{1\{y_1 \leq t\} - 1\{y_2 \leq t : t \in \mathcal{Y}\}$, it is clearly Donsker. The second condition is also trivially fulfilled, since $\sup_{t \in \mathcal{Y}} |\int \{1\{y_1 \leq t\} - 1\{y_2 \leq t\} - (F_1(t) - F_2(t))\} dF(y)| \leq 2$. □

References

[1] Joseph Altonji and Rosa Matzkin. Cross-section and panel data estimators for non-separable models with endogenous regrssors. *Econometrica*, 73(3):1053–1102, 2005.

[2] T.W. Anderson. On the distribution of the two-sample cramervon mises criterion. *The Annals of Mathematical Statistics*, 33(3):1148–1159, 1962.

[3] D.W.K. Andrews. A conditional kolmogorov test. *Econometrica*, 65(5):1097–1128, 1997.

[4] Joshua Angrist. Treatment heterogeneity in theory and practice. *The Economic Journal*, 114(494):C52–C83, 2004.

[5] Manuel Arellano. *Panel Data Econometrics*. 2003.

[6] Susan Athey and Guido Imbens. Identification and inference of nonlinear difference-in-difference models. *Econometrica*, 74(2):431–497, 2006.

[7] C. Alan Bester and Christian Hansen. Identification of marginal effects in a nonparametric correlated random effects model. *Journal of Business and Economic Statistics*, 27(2):235–250, 2009.

[8] Ivan A. Canay, Andres Santos, and Azeem M. Shaikh.

[9] David Card. The causal effect of education on earnings. *Handbook of Labor Economics*, 1999.

[10] Gary Chamberlain. Panel data. *Handbook of Econometrics*, 2:1247–1318.

[11] Gary Chamberlain. Multivariate regression models for panel data. *Journal of Econometrics*, 18(1):5–46, 1982.

[12] Gary Chamberlain. Binary response models for panel data: Identifcation and information. *Econometrica*, 78(1):159–168, 2010.

[13] Victor Chernozhukov, Ivan Fernandez-Val, Jinyong Hahn, and Whitney Newey. Average and quantile effects in nonseparable panel data models. *Econometrica, forthcoming*, 2010.

[14] Y.A. Davydov, M.A. Lifshits, and N.V. Smorodina. *Local Properties of Distributions of Stochastic Functionals*. Providence: American Mathematical Society, 1998.

[15] Kirill Evdokimov. Identification and estimation of a nonparametric panel data model with unobserved heterogeneity. *Department of Economics, Princeton University*, 2010.

[16] Kirill Evdokimov. Nonparametric identification of a nonlinear panel model with application to duration analysis with multiple spells. *Department of Economics, Princeton University*, 2011.

[17] Ivan Fernandez-Val. Fixed effects estimation of structural parameters and marginal effects in panel probit models. *Journal of Econometrics*, 150:71–85, 2009.

[18] Jean Gibbons. *Nonparametric Statistical Inference*. 1985.

[19] Bryan Graham and James Powell. Identification and estimation of average partial effects in 'irregular' correlated random coefficient panel data models. *Econometrica*, 80(5):2105–2152.

[20] James Heckman and T.E. MaCurdy. A life cycle model of female labor supply. *Review of Economic Studies*, 47:47–74, 1980.

[21] James Heckman and T.E. MaCurdy. Corrigendum on: A life cycle model of female labor supply. *Review of Economic Studies*, 49:659–660, 1982.

[22] James Heckman and Richard Robb. Alternative methods for evaluating the impact of interventions: An overview. *Journal of Econometrics*, pages 239–267, 1985.

[23] Stefan Hoderlein and Halbert White. Nonparametric identification of nonseparable panel data models with generalized fixed effects. *Unpublished Manuscipt*, 2009.

[24] B. Honore and E. Kyriazydou. Panel discrete choice models with lagged dependent variables. *Econometrica*, 68(4):839–874, 2000.

[25] D.R. Hyslop. State dependence, serial correlation and heterogeneity in intertemporal labor force participation of married women. *Econometrica*, 67(6, page=), 1999.

[26] Thierry Magnac. Panel binary variables and sufficiency: Generalizing conditional logit. *Econometrica*, 76(6):1859–1876.

[27] Yair Mundlak. On the pooling of time series and cross section data. *Econometrica*, 1978.

[28] Jean-Francois Quessy and Fancois Ethier. Cramer-von mises and characteristic fuunction tests for the two and k-samlpe problems with dependent data. *Computational Statistics and Data Analysis*, 56(6):2097–2111, 2012.

[29] Joseph P. Romano and Azeem M. Shaikh. Stepup procedures for control of generalizations of the familywise error rate. *Annals of Statistics*, 2006.

Table 1: Monte Carlo Results: KS and CM Tests, $S = 1000$

	n	1000			1500			2000		
	α	0.025	0.05	0.10	0.025	0.05	0.10	0.025	0.05	0.10
		Model (A)								
KS	nt	0.024	0.052	0.110	0.026	0.054	0.104	0.036	0.068	0.103
	pt	0.013	0.022	0.055	0.020	0.041	0.068	0.020	0.031	0.074
	gt	0.009	0.021	0.048	0.016	0.022	0.047	0.012	0.023	0.045
	$excl$	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
CM	nt	0.025	0.053	0.113	0.026	0.046	0.086	0.029	0.045	0.116
	pt	0.022	0.042	0.091	0.026	0.052	0.084	0.028	0.052	0.094
	gt	0.024	0.049	0.094	0.024	0.039	0.082	0.031	0.048	0.091
	$excl$	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		Model (B)								
KS	nt	0.985	0.996	0.998	0.998	1.000	1.000	1.000	1.000	1.000
	pt	0.014	0.023	0.054	0.020	0.041	0.067	0.020	0.030	0.074
	gt	0.009	0.021	0.046	0.015	0.022	0.046	0.011	0.021	0.044
	$excl$	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
CM	nt	0.861	0.919	0.951	0.971	0.991	0.997	0.993	0.998	1.000
	pt	0.023	0.043	0.090	0.030	0.053	0.081	0.028	0.050	0.092
	gt	0.025	0.052	0.095	0.030	0.043	0.083	0.030	0.047	0.092
	$excl$	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		Model (C)								
KS	nt	0.979	0.987	0.997	1.000	1.000	1.000	1.000	1.000	1.000
	pt	0.941	0.966	0.982	0.992	0.996	0.998	1.000	1.000	1.000
	gt	0.009	0.021	0.048	0.015	0.020	0.045	0.012	0.023	0.045
	$excl$	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
CM	nt	0.428	0.588	0.775	0.669	0.805	0.921	0.871	0.937	0.977
	pt	0.474	0.626	0.762	0.734	0.848	0.930	0.893	0.952	0.979
	gt	0.025	0.048	0.096	0.030	0.039	0.081	0.031	0.048	0.092
	$excl$	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		Model (D)								
KS	nt	0.990	0.998	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	pt	0.672	0.771	0.861	0.894	0.941	0.975	0.974	0.986	0.995
	gt	0.416	0.521	0.658	0.731	0.821	0.920	0.901	0.954	0.985
	$excl$	0.038	0.073	0.124	0.042	0.064	0.103	0.036	0.065	0.125
CM	nt	0.965	0.979	0.994	1.000	1.000	1.000	1.000	1.000	1.000
	pt	0.047	0.073	0.121	0.048	0.092	0.177	0.073	0.125	0.212
	gt	0.051	0.111	0.203	0.110	0.185	0.309	0.157	0.259	0.423
	$excl$	0.035	0.064	0.116	0.030	0.051	0.103	0.025	0.048	0.099
	s.e.	0.005	0.007	0.009	0.005	0.007	0.009	0.005	0.007	0.009

Table 2: Descriptive Statistics

	1983	1984	1985	1986	1987
Race	0.12				
Age	21.84				
	(2.22)				
HGC	12.34	12.45	12.57	12.57	12.61
	(1.77)	(1.83)	(1.94)	(1.94)	(1.98)
South	0.29	0.30	0.30	0.30	0.30
Urban	0.76	0.77	0.76	0.77	0.76
Log Hourly Wage	6.31	6.39	6.50	6.61	6.72
	(0.48)	(0.49)	(0.49)	(0.49)	(0.50)

Table 3: P-Values for Testing the Time Homogeneity up to a Time Effect

	KS	CM	F
LOG Hourly Wage			
1983-84	0.50	0.74	0.11
1984-85	0.11	0.87	0.09
1985-86	0.40	0.72	0.25
1986-87	0.43	0.17	0.46

Table 4: APE Results for Log Hourly Wage

Grade		Subs	APE	S.E.	t-Stat
1983	1984				
11	12	26	0.052	0.089	0.576
12	13	24	-0.120	0.133	-0.904
13	14	28	-0.069	0.061	-1.124
14	15	19	-0.037	0.035	-1.045
15	16	20	0.141	0.131	1.073
All Movers		122	-0.012	0.043	-0.267
1984	1985				
11	12	6	-0.010	0.032	-0.318
12	13	14	-0.080	0.067	-1.191
13	14	12	0.295	0.247	1.194
14	15	17	0.011	0.063	0.178
15	16	17	0.263	0.111	2.379
All Movers		73	0.095	0.055	1.723
1985	1986				
11	12	3	0.457	0.278	1.644
12	13	12	0.148	0.113	1.311
13	14	11	0.206	0.170	1.214
14	15	14	0.052	0.106	0.496
15	16	11	0.601	0.112	5.345
All Movers		58	0.226	0.060	3.737
1986	1987				
13	14	9	0.030	0.119	0.250
14	15	8	-0.125	0.168	-0.743
15	16	11	0.167	0.099	1.695
All Movers		41	-0.012	0.068	-0.179

Note: Results are listed for subpopulations with subsample size of 5 and above.

Table 5: Returns to Schooling: ANACOVA Results

	Full Sample			1983-1984		
	RF	X	A	RF	X	A
$Grade$	-0.0688	0.0772	0.0714	-0.2343	-0.0196	-0.0097
	(0.1308)	(0.0152)	(0.0148)	(0.1502)	(0.0194)	(0.0184)
$Grade^2$	-0.0070	-0.0005		-0.0013	0.0006	
	(0.0048)	(0.0004)		(0.0060)	(0.0005)	
Age^2	-0.0030		-0.0001	-0.0023		0.0006
	(0.0005)		(0.0003)	(0.0009)		(0.0004)
$Grade * Age$	0.0134			0.0113		
	(0.0015)			(0.0032)		
$Union$	0.1397	0.1423	0.1424	0.1642	0.1642	0.1643
	(0.0162)	(0.0163)	(0.0163)	(0.0152)	(0.0152)	(0.0152)