Time-Varying Linear Transformation Models with Fixed Effects and Endogeneity for Short Panels (preliminary and incomplete, see link for latest version)*

Irene Botosaru, Chris Muris, and Senay Sokullu†

August 16, 2021

This paper considers a class of fixed-T nonlinear panel models with time-varying link function, fixed effects, and endogenous regressors. We establish sufficient conditions for the identification of the regression coefficients, the time-varying link function, the distribution of the counterfactual outcomes, and certain (time-varying) average partial effects. We propose estimators for the regression coefficient, the link function, the average partial effects, and study their asymptotic properties. We show the relevance of our model by obtaining new results for a nonlinear version of the canonical dynamic panel data model.

*We would like to thank Xavier D'Haultfoeuille, Jean-Pierre Florens, Thierry Magnac for comments and suggestions, and participants at various conferences and seminars.

†Contact: Botosaru and Muris are at McMaster University, Department of Economics, Canada, and Sokullu is at Bristol University, School of Economics, UK. Emails: botosari@mcmaster.ca, muerisc@mcmaster.ca, senay.sokullu@bristol.ac.uk. We gratefully acknowledge financial support from the Social Sciences and Humanities Research Council of Canada under grants IG 435-2021-0778. This research was undertaken, in part, thanks to funding from the Canada Research Chairs Program.

1 Introduction

We consider a nonlinear panel model with endogeneity, where the outcome for individual i at time t can be written as a time-varying transformation of a latent linear variable with endogenous regressors. That is, the observed outcome is specified as:

$$Y_{it} = h_t\left(Y_{it}^*\right) = h_t(\alpha_i + \overline{X}_{it}\bar{\beta} + U_{it}),\ i = 1, \ldots, n,\ t = 1, \ldots, T, \tag{1}$$

where $Y_{it} \in \mathcal{Y}_t \subseteq \mathbb{R}$ is a continuous random variable, $h_t : \mathbb{R} \to \mathcal{Y}_t$ is an unknown, strictly monotonic transformation function that varies with t, $\alpha_i \in \mathbb{R}$ is an unobserved individual effect, $\bar{\beta} \in \mathbb{R}^{k+1}$ is a vector of coefficients,[1] $U_{it} \in \mathbb{R}$ is the stochastic error, and

$$\overline{X}_{it} = (X_{0it}, X_{1it}, \ldots, X_{kit}) \in \mathcal{X}_t \subseteq \mathbb{R}^{k+1}$$

is a vector of explanatory variables that are endogenous, i.e.

$$E\left(U_{it}|\overline{X}_{it}\right) \neq 0 \text{ for all } t = 1, \ldots, T. \tag{2}$$

The individual effect α_i is a fixed effect in the sense that it is allowed to be arbitrarily correlated with $\overline{X}_{it}$ and that its distribution is arbitrary. The distribution of U_{it} is left unspecified except for a conditional mean restriction in (5) below. Our interest lies in the identification and estimation of $\bar{\beta}$ and $h_t,\ t = 1, \ldots, T$. We will also show that identification of $\bar{\beta}$ and h_t obtains identification of the distribution of the counterfactual outcome, and, as such of the average structural function and certain partial effects, all of which can be time-varying.

Our main example is a dynamic panel model with outcome equation:

$$Y_{i0} = h_0\left(\phi_i, \widetilde{X}_{i0}\right), \tag{3}$$

$$Y_{it} = h_t\left(\alpha_i + \widetilde{X}_{it}\widetilde{\beta} + \rho Y_{i,t-1} + U_{it}\right), t = 1, \ldots, T, \tag{4}$$

where ϕ_i captures additional individual-specific unobserved heterogeneity from the

[1] In Appendix B, we allow for a nonparametric function of the covariates, e.g. $\rho\left(\overline{X}_{it}\right)$.

initial condition. This specification is nested by (1) by setting $\overline{X}_{it} = \left(\widetilde{X}_{it}, Y_{i,t-1}\right)$,$\overline{\beta} = \left(\widetilde{\beta}, \rho\right)$, and, in turn, nests the linear dynamic panel model when $h_t(v) = v$ for $t \geq 1$. To the best of our knowledge, this is the first paper showing identification of $\left(\widetilde{\beta}, \rho\right)$, h_t, $t \geq 1$, of the distribution of the counterfactual outcomes for this class of models.

Our proposed solution involves the existence of instrumental variables, $Z_{it} \in \mathcal{Z}_t \subseteq \mathbb{R}^q$, that are continuous random variables and that satisfy the following mean independence condition. For any $z_t \in \mathcal{Z}_t$,

$$E\left(U_{it} - U_{it-1} | Z_{it} = z_t\right) = 0 \text{ for all } t = 1, \ldots, T. \tag{5}$$

The conditional mean restriction above allows for heteroskedasticity in the errors, and, in particular, it allows the instrumental variable at time t to affect the level of the errors at t as long as it does so in a time homogeneous way.[2]

Relative contribution. We are not aware of any existing work on fixed-T, fixed effects nonlinear panel models that allows for time-varying link functions *and* endogenous regressors. Thus, to the best of our knowledge, our identification and estimation results are novel. Our paper contributes to at least three literatures: nonlinear panel models; dynamic panel models; and transformation models.

First, our results contribute to the literature on nonlinear panel models with fixed-effects and fixed-T. Within this literature, Abrevaya (1999) considered the outcome equation (1) and proposed an estimator for the regression coefficient. Botosaru et al. (2021) studied identification and estimation of the time-varying link function. In these papers, results are developed under the assumption that the regressors X_{it} are strictly exogenous. We do not invoke such an exogeneity assumption.

That it is challenging to deal with endogenous regressors in nonlinear panel models with fixed effects and fixed-T is evident from reviews of the literature in Arellano and Honoré (2001); Arellano and Bonhomme (2011). A notable exception is Altonji and Matzkin (2005), who consider an outcome equation that nests ours, and who

[2] In contrast, in nonseparable models this type of conditional mean restriction is not sufficient for identification of the structural parameters or of the partial effects. In general stronger assumptions, such as independence between U_t and X_t for each and all t is usually maintained in those models.

also allow for endogenous regressors. Altonji and Matzkin (2005) make progress by imposing restrictions on the distribution of (α_i, X_i). In contrast, we obtain identification without imposing such restrictions.

Within the nonlinear panel literature described above, Botosaru et al. (2021) is closest in spirit. The main differences with the specification in Botosaru et al. (2021) are that the link function in (1) is assumed to be strictly monotonic, and that the covariates are allowed to be endogenous.

We also contribute to the literature on dynamic panel models. As a special case of our general result, we analyze a nonlinear version of the linear dynamic panel model by including in (1) the lagged dependent variable $Y_{i,t-1}$ in the vector of regressors X_{it}. There is a large literature on linear dynamic panels, see Bun and Sarafidis (2015) for a review. These models are very popular in applied practice. For example, the early key contributions by Arellano and Bond (1991); Blundell and Bond (1998) have $32,279$ resp. $22,236$ Google Scholar citations at the time of writing. That the combination of a dynamic structure and a nonlinear structure is difficult to handle is clear from the literature on the dynamic binary choice model with fixed effects, cf. Honoré and Kyriazidou (2000). For example, in Honoré and Kyriazidou (2000), no deterministic time trend can be accommodated. In contrast, our model allows for the transformation functions to vary over time in an arbitrary fashion.

We contribute to the literature on panel transformation models with endogenous regressors. We use insights from the nonparametric instrumental variables (NPIV) literature to solve an endogeneity issue, extending previous work by Florens et al. (2012), Féve and Florens (2014), and Florens and Sokullu (2017) to nonlinear panel models with fixed effects. The main difference with Féve and Florens (2014) is that our specification allows for a time-varying link function, which allows for the observed and unobserved covariates to impact the outcomes differently over time. The analysis in Florens et al. (2012) and Florens and Sokullu (2017) applies to cross-sectional data, so the link function there does not vary over time and there are no fixed effects (see also Féve and Florens (2010)). On the other hand, both Florens and Sokullu (2017) and Féve and Florens (2014) allow for a nonparametric function of observed endogenous covariates, i.e. $\rho(X_{it})$ instead of $X_{it}\bar{\beta}$. In Appendix B, we explain how our analysis can be extended to allow for this possibility. Note that in this case, our framework nests that of Féve and Florens (2014).

Other related works have addressed the problem of endogeneity in transformation models via arguments based on special regressors, e.g. Chiappori et al. (2015), and control functions, e.g. Vanhems and Van Keilegom (2019). These papers consider a cross-sectional set-up, so the transformation function is not indexed by time and, importantly, there are no fixed effects. We consider a panel data setting and our identification argument uses instrumental variables.[3]

We adopt an inverse problem approach to derive sufficient conditions for the identification of $\bar{\beta}$ and h_t. As such, we use concepts from the NPIV literature such as invertibility of an operator, completeness, and measurable separability. We derive sufficient conditions for these high-level assumptions for particular models nested by our framework, such as the nonlinear dynamic panel model. Our proposed estimator is based on Tikhonov regularization and follows closely the procedure in Florens and Sokullu (2017).

Finally, there is a growing number of working papers deriving conditions for the identification of marginal and partial effects for dynamic discrete choice models, e.g., Aguirregabiria and Carro (2021), Aguirregabiria et al. (2021), Davezies et al. (2021), Dobronyi et al. (2021), Liu et al. (2021), Pakel and Weidner (2021). None covers our specification with unknown and time-varying transformation. Botosaru and Muris (2017) consider partial effects for the class of models where the outcome equation is as in (1) with exogenous X_t. Chernozhukov et al. (2013) consider a non-separable outcome equation, but do not allow for endogenous regressors, arbitrary time-varyingness, and they require boundedness of the dependent variable and discreteness of the regressors. An important point made by recent papers starting with Botosaru and Muris (2017) is that, even in nonlinear models, average partial effects can be identified without identification of the distribution of the fixed effects.

Notation. For a random variable V with support $\mathcal{V}$, we let $L^2_{\mathcal{V}}$ denote the space of functions $g : \mathcal{V} \to \mathbb{R}$ such that $E\left|g\left(V\right)\right|^2 < \infty$. We denote by g^{-1} the inverse of an

[3]Other related work, set in a cross-sectional setting, can be found in the literature review in e.g. Birke et al. (2017).

arbitrary, invertible function $g : \mathbb{R} \to \mathbb{R}$. We use $\otimes$ to denote the tensor product. We let $\mathcal{C} : \mathbb{R} \to \mathbb{R}$ denote a bounded, continuous, symmetric, univariate kernel function of order m, i.e. $\int \mathcal{C}(u)du = 1$, $\int u^j \mathcal{C}(u)du = 0$ for all $j = 1, ..., m-1$, and $\int u^m \mathcal{C}(u)du < \infty$ and $\int \mathcal{C}^2(u)du < \infty$. We let $\mathcal{K} : \mathbb{R}^d \to \mathbb{R}$ denote a multivariate kernel function defined as the product kernel $\mathcal{K}(w) = \prod_{k=1}^{d} \mathcal{C}(w_k)$. For an operator K between two Hilbert spaces, we denote by $\mathcal{R}(K)$ the range of the operator and by $\mathcal{R}(K)^{\perp}$ its orthogonal complement .

2 Identification

Our identification results require at least two time periods, so we will let $T = 2$ in what follows. We drop the i subscript in this section.

Assumption 1. *For each $t = 1, 2$, $h_t : \mathbb{R} \to \mathcal{Y}_t$ is strictly monotonic.*

Assumption 1 allows us to work with the inverse of h_t, $t = 1, 2$. Such shape restrictions are quite common in the literature on transformation models.

Assumption 2. *(i) The first element of $\bar{\beta}$ is normalized to 1, i.e. $\bar{\beta} = (1, \beta)$, $\beta \in \mathbb{R}^k$. (ii) $E\left(h_1^{-1}(Y_1)\right) = 0$.*

Without parametric restrictions on h_t and on the distribution of U_t, the outcome in (1) follows a semiparametric single index specification. Therefore, both a scale normalization, 2(i), and a location normalization, 2(ii), are needed for identification, see also e.g. Horowitz (2009).

Letting $X_{0t} \in \mathbb{R}$ denote the covariate associated with the normalized coefficient from Assumption 2(i) and

$$X_t \equiv (X_{1t}, X_{2t}, \ldots, X_{kt}) \in \mathbb{R}^k,$$

the outcome equation (1) can then be written as

$$Y_t = h_t\left(\alpha + X_{0t} + X_t\beta + U_t\right).$$

Assumption 3. *There exist random variables $Z \in \mathcal{Z}$ such that for any $z \in \mathcal{Z}$,*

$$E\left(U_2 - U_1 \middle| Z = z\right) = 0.$$

In Assumption 3 the instrumental variables are time-invariant. This assumption is made for convenience since our identification strategy can easily be extended to accommodate time-varying instrumental variables. As the assumption is made on the difference over time in the errors, it does not require that $E\left(U_t|Z\right) = 0$, thus allowing Z to enter the outcome equation, as long as it does so in a time-homogeneous way.

Let $\Delta X \equiv X_2 - X_1$, $\Delta X_0 \equiv X_{02} - X_{01}$, and $r\left(z\right) \equiv E\left(\Delta X_0 \middle| Z = z\right)$. Assumptions 1, 2, and 3, obtain that, for any $z \in \mathcal{Z}$,

$$E\left(h_2^{-1}\left(Y_2\right) - h_1^{-1}\left(Y_1\right) - \Delta X \beta \middle| Z = z\right) = r\left(z\right). \tag{6}$$

Equation (6) is an integral equation for the parameters of interest, $h_1^{-1}, h_2^{-1}, \beta$. This shows that the three parameters can be characterized by the functional equation $K\left(h_1^{-1}, h_2^{-1}, \beta\right) = r$, where $K : \mathcal{H}_1 \to \mathcal{H}_2$ is a multilinear integral operator and $\mathcal{H}_1$,$\mathcal{H}_2$ are function spaces defined below.

Assumption 4. *(i) $\mathcal{H}_1 = L^2_{\mathcal{Y}_1} \otimes L^2_{\mathcal{Y}_2} \otimes \mathbb{R}^k$ and $\mathcal{H}_2 = L^2_{\mathcal{Z}}$; (ii) $r \in L^2_{\mathcal{Z}}$; and (iii) The joint distribution of $\left(Y_1, Y_2, \Delta X, Z\right)$ is dominated by the product of its marginal distributions, and its density is square integrable w.r.t. the product of marginals.*

A few remarks are in order. First, Assumption 4 allows us to define K as the conditional expectation operator:

$$K : L^2_{\mathcal{Y}_1} \otimes L^2_{\mathcal{Y}_2} \otimes \mathbb{R}^k \to L^2_{\mathcal{Z}}, \tag{7}$$

that maps

$$\left(h_1^{-1}, h_2^{-1}, \beta\right) \mapsto E\left(h_2^{-1}\left(Y_2\right) \middle| Z = .\right) - E\left(h_1^{-1}\left(Y_1\right) \middle| Z = .\right) - E\left(\Delta X \beta \middle| Z = .\right),$$

where we used the linearity property of the expectation operators. Second, Assumption 4(i) and (ii) are satisfied provided that the variance of $U_2 - U_1$ and the variance of ΔX each is finite. Assumption 4(iii) guarantees that the conditional density $f_{Y_1,Y_2,\Delta X|Z}$ is well-defined and that the operator K is Hilbert-Schmidt.

Given correct model specification, $r \in \mathcal{R}(K)$ so that the functional equation in (6) has at least one solution for $\left(h_1^{-1}, h_2^{-1}, \beta\right)$. The assumption below guarantees uniqueness of the solution.

Assumption 5. *K is injective, i.e. for any $\left(\delta_1^{-1}, \delta_2^{-1}, b\right) \in L^2_{\mathcal{Y}_1} \otimes L^2_{\mathcal{Y}_2} \otimes \mathbb{R}^k$,*

$$K\left(\delta_1^{-1}, \delta_2^{-1}, b\right) = 0 \ a.s. \implies \left(\delta_1^{-1}, \delta_2^{-1}, b\right) = (0,0,0) \ \ a.s.$$

Injectivity of a linear integral operator is a commonly invoked assumption in the NPIV literature. In our set-up, we require injectivity of a multilinear operator.[4] However, using the linearity of the expectation operator, we can state necessary and sufficient assumptions for Assumption 5 that are in line with the literature that works with linear integral operators. To this end, define the following linear operators:

$$K_{y_t} : L^2_{\mathcal{Y}_t} \to L^2_{\mathcal{Z}} : h_t^{-1} \mapsto E\left(h_t^{-1}(Y_t)\middle| Z = .\right), \ t = 1, 2, \tag{8}$$

$$K_x : \mathbb{R}^k \to L^2_{\mathcal{Z}} : \beta \mapsto E\left(\Delta X \beta | Z = .\right). \tag{9}$$

Assumption 6. *(i) Each operator K_{y_1}, K_{y_2}, K_x is injective; (ii) $\mathcal{R}(K_{y_1}) \cap \mathcal{R}(K_{y_2}) \cap \mathcal{R}(K_x) = \{0\}$.*

Assumption 6(i) is equivalent to assuming that each Y_t is strongly identified by Z, i.e. for any $\gamma \in L^2_{\mathcal{Y}_t}$,

$$E\left(\gamma(Y_t)| Z\right) = 0 \text{ a.s. } F_Z \implies \gamma(Y_t) = 0 \text{ a.s. } F_{Y_t}, t = 1, 2,$$

and that the matrix $E\left[E\left(\Delta X|Z\right) E\left(\Delta X'|Z\right)\right]$ has full rank. The strong identification assumption is the standard completeness assumption usually invoked in the NPIV literature. If each Y_t and Z are continuous, have the same dimension, and support equal to a rectangle then the completeness condition holds generically in the sense

[4] Florens et al. (2012) use a bilinear operator in the proof of their Theorem 2.1.

of Andrews (2017), see also Chen et al. (2014), Newey and Powell (2003). It may be possible to consider weaker sufficient conditions by adapting Proposition 2.2 in d'Haultfoeuille (2010) to the case of square integrable functions. Other papers that provide sufficient conditions for completeness are d'Haultfoeuille (2011), Andrews (2017), and Hu and Shiu (2018).

Assumption 6(ii) is implied by $Y_1, Y_2, \Delta X$ each being strongly identified by Z, and by measurable separability of $(Y_1, Y_2, \Delta X)$.[5] Measurable separability is a high-level assumption that rules out a linear relationship between $Y_1, Y_2,$ and ΔX. The assumption fails if there exists an *additive* functional relationship between Y_1, Y_2, and ΔX, see, e.g., Newey et al. (1999). Lemma (1) in Appendix A establishes low-level assumptions for measurable separability.

Theorem 1 (Identification)**.** *Suppose that* $(Y_1, Y_2, X_{01}, X_{02}, X_1, X_2, Z)$ *follow the model described by 1, 2, and 5. Let assumptions 1, 2, 3, 4, and either 5 or 6 hold. Then* h_1, h_2, β *are identified.*

Proof. The proof can be found in Appendix B.1. □

2.1 Illustration: A nonlinear dynamic panel model

The nonlinear panel model introduced above nests a nonlinear version of the canonical linear panel data model, which plays an important role in our Monte Carlo study in Section 6 and in our empirical illustration in Section 7.

Consider the outcome equations in (3) and (4). Setting $h_t(v) = v$ for $t \geq 1$ obtains the outcome equation of the standard dynamic panel model. The period-0 equation has its own ϕ_i that can capture i-specific terms regarding to the initial condition, the history of a given unit i, and a period-0 error term.

[5]The random variables $(Y_1, Y_2, \Delta X)$ are measurably separable when for any $\delta_t \in L^2_{Y_t}$ and any $b \in \mathbb{R}^k$ if

$$\delta_2(Y_2) - \delta_1(Y_1) - \Delta X b = 0 \text{ a.s. } F_{Y_1, Y_2, \Delta X},$$

then there exist constants $c_t \in \mathbb{R}$, $t = 1, 2$, such that

$$\delta_t(Y_t) = c_t \text{ a.s. } F_{Y_t}, \; t = 1, 2.$$

In the linear version of this model, estimation via differences is problematic because $\Delta Y_{i,t-1}$ is correlated with ΔU_{it}. Naturally, this problem carries over to the nonlinear generalization. We can use internal instruments to address this endogeneity issue. Internal instruments are available under a strict exogeneity assumption and restrictions on the serial correlation in U_{it}.

Assumption 7. *For each t, $E[U_{it}]$ does not depend on t, and*

$$U_{it} \perp \left(\widetilde{X}_{i0,\cdots,}\widetilde{X}_{iT}, \alpha_i, \phi_i, U_{i1}, \cdots U_{i,t-1}\right).$$

This assumption is stronger than necessary: serial independence in the errors and the strict exogeneity condition on the regressors can be relaxed to a form of mean-independence, However, given the nonlinear nature of our model, it will be convenient to maintain statistical independence.

To place the nonlinear panel model in the notation of the general specification above, set

$$X_{it} = \left(\widetilde{X}_{it}, Y_{i,t-1}\right),$$
$$\overline{\beta} = \left(\widetilde{\beta}, \rho\right)$$

and assume that $\widetilde{X}_{it}$ is non-empty. Then we can rewrite the nonlinear dynamic panel model as

$$Y_{it} = h_t\left(\alpha_i + X_{it}\overline{\beta} + U_{it}\right),\ i = 1, \cdots, n,\ t = 1, \cdots, T, \tag{10}$$

and, with three periods of data on Y_i and two periods on X_i, we can use as instruments $Z_i = \left(Y_{i0}, \widetilde{X}_{i1}, \widetilde{X}_{i2}\right)$ for the difference $U_{i2} - U_{i1}$. We have the following result:

Theorem 2. *Suppose that $\left(Y_0, Y_1, Y_2, \widetilde{X}_1, \widetilde{X}_2\right)$ follow the nonlinear dynamic panel model above and that Assumptions 1, 3, 4, either 5 or 6, and 7 hold. Then h_1, h_2, $\widetilde{\beta}$, and ρ are identified.*

Proof. This follows immediately from Theorem 1 once we verify that Assumption 7 implies Assumption 3, i.e. that

$$E\left(U_{i,2}-U_{i,1}\right| Y_0, X_1, X_2) = 0.$$

But this follows immediately from the fact that $(U_{i1}, U_{i2}) \perp (\phi_i, X_{i1}, X_{i2})$. □

3 Partial effects

Building on results in Botosaru and Muris (2017) (Section 3.2), we show that identification of the structural parameters (β, h_1, h_2) implies identification of certain partial effects. The main differences between the current setting and that of Botosaru and Muris (2017) are that, here, (i) the transformation function is invertible, and (ii) the regressors are endogenous. Below, we first show that the distribution of the counterfactual outcome at t is identified, and then we show identification of the average partial effects.

Denote the *counterfactual outcome* by

$$Y_{it}(x) = h_t\left(\alpha_i + x\beta + U_{it}\right),$$

which is the outcome of person i at time t had $X_{it} = x$, while holding α_i, U_{it} fixed. Note that $Y_{it} = Y_{it}(X_{it})$. The distribution of the counterfactual outcome at time t and for any values (y, x) is defined as:

$$P\left(Y_{it}(x) \leq y\right) = P\left(h_t\left(\alpha_i + x\beta + U_{it}\right) \leq y\right).$$

Since (Y_{it}, X_{it}) are observed, and (h_t, β) have been identified, this counterfactual quantity is also identified, and given by:

$$\begin{aligned}P\left(Y_{it}\left(x\right)\leq y\right)&=P\left(h_t\left(\alpha_i+x\beta+U_{it}\right)\leq y\right)\\&=P\left(\alpha_i+x\beta+U_{it}\leq h_t^{-1}\left(y\right)\right)\\&=P\left(\alpha_i+X_{it}\beta+U_{it}\leq h_t^{-1}\left(y\right)-\left(x-X_{it}\right)\beta\right)\\&=P\left(h_t\left(\alpha_i+X_{it}\beta+U_{it}\right)\leq h_t\left(h_t^{-1}\left(y\right)-\left(x-X_{it}\right)\beta\right)\right)\\&=P\left(Y_{it}\leq h_t\left(h_t^{-1}\left(y\right)-\left(x-X_{it}\right)\beta\right)\right).\end{aligned}$$

The first equality uses the outcome equation for our model; the second uses strict increasingness of h_t; the third adds $\left(X_{it}-x\right)\beta$ on both sides of the inequality; the fourth applies the strictly monotone function to both sides; and the final equality substitutes the observed Y_{it}. The final form resembles the expression in the changes-in-changes estimator in Athey and Imbens (2006).

In our empirical illustration, we build on this result to obtain regressor effects. Rather than looking at a fixed value of the regressors x, we will look at a counterfactual value of the covariates that change the kth covariate by 1 unit. This counterfactual value of the regressors can be written as $X_{it}+e_k$, where e_k is the unit vector of the appropriate length, with a 1 in entry k, and zeros elsewhere. Following the sequence of equalities above, we obtain

$$P\left(Y_{it}\left(X_{it}+e_k\right)\leq y\right)=P\left(Y_{it}\leq h_t\left(h_t^{-1}\left(y\right)-\beta_k\right)\right),$$

where β_k is the value of the kth coefficient. Then, the difference in distributions

$$\begin{aligned}\tau_{k,t}\left(y\right)&\equiv P\left(Y_{it}\left(X_{it}+e_k\right)\leq y\right)-P\left(Y_{it}\leq y\right)\\&=P\left(Y_{it}\leq h_t\left(h_t^{-1}\left(y\right)-\beta_k\right)\right)-P\left(Y_{it}\leq y\right)\end{aligned}\tag{11}$$

is the partial effect for regressor k.

In our empirical illustration, we are interested in the average effect rather than the distribution of the counterfactual. This can be obtained from taking expectations in (11), or directly from

$$\delta_{k,t} \equiv E\left[Y_{it}\left(X_{it}+e_k\right)-Y_{it}\right] \tag{12}$$

$$= E\left[h_t\left(h_t^{-1}\left(Y_{it}\right)+\beta_k\right)\right]-E\left[Y_{it}\right]. \tag{13}$$

The final expression depends only on observable or identified quantities, and is therefore identified. We call $\delta_{k,t}$ the average partial effect (APE), which shows the average change in the outcome at time t when the kth covariate changes by one unit at time t. The expression in (12) has the advantage that it suggests an estimator for APE, namely the sample analog of the right hand side.

4 Estimation

Our identification argument naturally gives rise to a system of three normal equations based on the three linear operators K_{y_1}, K_{y_2}, K_x and their adjoints. However, the resulting system of normal equations is unnecessarily complicated.[6] Instead, we follow Florens and Sokullu (2017) and work with the following bilinear operator:

$$K_y : \tilde{L}^2_{\mathcal{Y}_1} \otimes L^2_{\mathcal{Y}_2} \to L^2_{\mathcal{Z}} : \left(h_1^{-1}, h_2^{-1}\right) \mapsto K_{y_2} h_2^{-1} - K_{y_1} h_1^{-1},$$

where $\tilde{L}^2_{\mathcal{Y}_1} \equiv \left\{h_1^{-1} \in L^2_{\mathcal{Y}_1} : E\left(h_1^{-1}\left(Y_1\right)=0\right)\right\}$,[7] so that Assumption 2(ii) holds. We define the following dual operators or adjoints:

$$K_y^* : L^2_{\mathcal{Z}} \to \tilde{L}^2_{\mathcal{Y}_1} \otimes L^2_{\mathcal{Y}_2} : \psi \mapsto \begin{pmatrix} E\left[\psi\left(Z\right)\middle| Y_2 = \cdot\right] \\ -\mathbb{P}E\left[\psi\left(Z\right)\middle| Y_1 = \cdot\right] \end{pmatrix},$$

$$K_x^* : L^2_{\mathcal{Z}} \to \mathbb{R}^k : \psi \mapsto E\left[\psi\left(Z\right)\Delta X\right],$$

where $\mathbb{P}$ is the operator that projects functions from $L^2_{\mathcal{Y}_1}$ to $\tilde{L}^2_{\mathcal{Y}_1}$.

[6] We show in Appendix A that the system of three normal equations based on the three linear operators is identical to the one that we use with two normal equations based on a bilinear operator.

[7] It is possible to show that this operator is injective given Assumption 6.

We can then write (6) as:

$$K_y\left(h_1^{-1}, h_2^{-1}\right) + K_x\beta = r, \tag{14}$$

and we can project the problem in (14) onto the parameter spaces, $\tilde{L}^2_{\mathcal{Y}_1} \otimes L^2_{\mathcal{Y}_2}$ and $\mathbb{R}^k$, using the dual operators above. The functions $\left(h_1^{-1}, h_2^{-1}, \beta\right)$ are then characterized as solutions to the following system of normal equations:

$$K_y^* K_y\left(h_1^{-1}, h_2^{-1}\right) = K_y^* r + K_y^* K_x \beta, \tag{15}$$

$$K_x^* K_y\left(h_1^{-1}, h_2^{-1}\right) = K_x^* r + K_x^* K_x \beta. \tag{16}$$

Letting I be the identity operator in $L^2_{\mathcal{Y}_1} \otimes L^2_{\mathcal{Y}_2}$, $P_x \equiv K_x \left(K_x^* K_x\right)^{-1} K_x^*$ be the orthogonal projection operator onto the closure of the range of K_x, and $P_y \equiv K_y\left(K_y^* K_y\right)^{-1} K_y^*$ be the orthogonal projection operator onto the closure of the range of K_y, the above linear system is equivalent to:

$$K_y^*\left(I - P_x\right) r = K_y^*\left(I - P_x\right) K_y \left(h_1^{-1}, h_2^{-1}\right), \tag{17}$$

$$K_x^*\left(I - P_y\right) r = K_x^*\left(I - P_y\right) K_x \beta. \tag{18}$$

The parameters of interest can in principle be obtained from equations (17) and (18), after replacing the operators, their adjoints, and r by their sample analogues, call them $\hat{K}_x$, $\hat{K}_y$, $\hat{r}$. For example,

$$\begin{aligned} \left(\hat{h}_1^{-1}, \hat{h}_2^{-1}\right)'_{naive} &= \left(\hat{K}_y^*\left(I - \hat{P}_x\right)\hat{K}_y\right)^{-1} \hat{K}_y^*\left(I - \hat{P}_x\right)\hat{r}, & (19)\\ \hat{\beta}_{naive} &= \left(\hat{K}_x^*\left(I - \hat{P}_y\right)\hat{K}_x\right)^{-1} \hat{K}_x^*\left(I - \hat{P}_y\right)\hat{r}. & (20) \end{aligned}$$

It is well known in the literature on inverse problems, see, e.g. Carrasco and Florens (2011), Centorrino et al. (2017), Florens and Sokullu (2017), Babii and Florens (2020), that estimating $\left(h_1^{-1}, h_2^{-1}\right)$ and β by naively inverting the sample analogues of (17) and (18) as in (19) and (20) is a statistically ill-posed problem, in the sense that the naive estimators are not stable with respect to estimation error in $\hat{K}_y$ and

$\hat{P}_y$. Tikhonov regularization is then used in order to smooth out discontinuities due to inversion.

Letting γ_n be a regularization parameter, such that $\gamma_n \to 0$ at a rate defined in Assumption (12) below, the regularized estimators $\left(\hat{h}_1^{-1}, \hat{h}_2^{-1}\right)$, $\hat{\beta}$ are given by:

$$\left(\hat{h}_1^{-1}, \hat{h}_2^{-1}\right)' = \left(\gamma_n I + \hat{K}_y^* \left(I - \hat{P}_x\right) \hat{K}_y\right)^{-1} \hat{K}_y^* \left(I - \hat{P}_x\right) \Delta X_0, \tag{21}$$

$$\hat{\beta} = \left(\hat{K}_x^* \left(I - \hat{P}_y^{\gamma_n}\right) \hat{K}_x\right)^{-1} \hat{K}_x^* \left(I - \hat{P}_y^{\gamma_n}\right) \hat{r}, \tag{22}$$

where $\hat{P}_y^{\gamma_n} \equiv \hat{K}_y \left(\gamma_n I + \hat{K}_y^* \hat{K}_y\right)^{-1} \hat{K}_y^*$ is the regularized projection operator P_y. Note that, although estimation of β is also affected by regularization, we show that $\hat{\beta}$ is $\sqrt{n}$-consistent and asymptotically normal. Additionally, note that a single regularization parameter is introduced. Although it is possible to allow for two different regularization parameters, one for h_1^{-1} and one for h_2^{-1}, we would need them to converge to zero at the same rate when deriving the asymptotic properties of our estimators in the next section. Hence, for the sake of exposition, we assume that the two regularization parameters are equal.

Letting $\{Y_{1i}, Y_{2i}, \Delta X_{i0} = X_{02i} - X_{01i}, \Delta X_i = X_{2i} - X_{1i}, Z_i\}_{i=1}^n$ be a random sample from a population conformable to our assumptions in Section (2), we consider the following nonparametric estimators for the operators in (21) and (22):

$$\hat{K}_x \gamma (z) = \frac{1}{n b_z^q} \frac{1}{\hat{f}_Z(z)} \sum_{i=1}^{n} \Delta X_i' \gamma \mathcal{K} \left(\frac{Z_i - z}{b_z} \right), \text{ for all } \gamma \in \mathbb{R}^k,$$

$$\hat{K}_y (g_1, g_2)(z) = \frac{1}{\hat{f}_Z(z)} \left(\int g_2(y) \hat{f}_{Y_2,Z}(y,z)\, dy - \int g_1(y) \hat{f}_{Y_1,Z}(y,z)\, dy \right), \text{ for all } g_1, g_2 \in \tilde{L}^2_{\mathcal{Y}_1} \otimes L^2_{\mathcal{Y}_2},$$

$$\hat{K}_x^* g_3 (z) = \frac{1}{n b_z^q} \sum_{i=1}^{n} \Delta X_i \int g_3(z) \mathcal{K} \left(\frac{Z_i - z}{b_z} \right) dz, \text{ for all } g_3 \in L^2_{\mathcal{Z}},$$

$$\hat{K}_y^* g_4 (y_1, y_2) = \begin{pmatrix} \frac{1}{\hat{f}_{Y_2}(y_2)} \int g_4(z) \hat{f}_{Y_2,Z}(y_2,z)\, dz \\ -\frac{1}{\hat{f}_{Y_1}(y_1)} \int g_4(z) \hat{f}_{Y_1,Z}(y_1,z)\, dz \end{pmatrix}, \text{ for all } g_4 \in L^2_{\mathcal{Z}},$$

$$\hat{r}(z) = \frac{1}{n b_z^q} \frac{1}{\hat{f}_Z(z)} \sum_{i=1}^{n} \Delta X_{i0} \mathcal{K} \left(\frac{Z_i - z}{b_z} \right),$$

where $\mathcal{K}$ is a multivariate kernel function, b_z a bandwidth parameter that is assumed to be the same for each of the q components of Z and which approaches 0 as $n \to \infty$ at a rate specified in Assumption (12) below, and where:

$$\hat{f}_{Y_t,Z}(y,z) = \frac{1}{n b_{y_t} b_z^q} \sum_{i=1}^{n} \mathcal{C} \left(\frac{Y_{it} - y}{b_{y_t}} \right) \mathcal{K} \left(\frac{Z_i - z}{b_z} \right), t = 1, 2,$$

$$\hat{f}_{Y_t}(y) = \frac{1}{n b_{y_t}} \sum_{i=1}^{n} \mathcal{C} \left(\frac{Y_{it} - y}{b_{y_t}} \right), t = 1, 2,$$

$$\hat{f}_Z(z) = \frac{1}{n b_z^q} \sum_{i=1}^{n} \mathcal{K} \left(\frac{Z_i - z}{b_z} \right),$$

where $\mathcal{C}$ is a univariate kernel function and b_{y_t} is a bandwidth parameter that approaches 0 as $n \to \infty$ at a rate specified in Assumption (12) below.

The estimators $\left(\hat{h}_1^{-1}, \hat{h}_2^{-1}, \hat{\beta} \right)$ are then the solutions to (21) and (22), where the operators are replaced by their estimators defined above. Below, we describe how to implement the method and give an explicit expression for the estimators $\left(\hat{h}_1^{-1}, \hat{h}_2^{-1}, \hat{\beta}\right)$.

Given $\left(\hat{h}_1^{-1}, \hat{h}_2^{-1}, \hat{\beta}\right)$, the estimator for the APE $\delta_{k,t}$ defined in (13) is given by the sample analog of that expression, i.e.

$$\widehat{\delta}_{k,t} = \frac{1}{n}\sum_{i=1}^{n}\left[\widehat{h}_t\left(\widehat{h}_t^{-1}\left(Y_{it}\right) + \widehat{\beta}_k\right) - Y_{it}\right],\ t = 1, 2. \tag{23}$$

4.1 Implementation of the estimation method

The estimators $\left(\hat{h}_1^{-1}, \hat{h}_2^{-1}, \hat{\beta}\right)$ are constructed as follows.

Let A_{y_t}, $t = 1, 2$, and A_z be matrices with the (i, j) element given by:

$$A_{y_t}(i,j) = \frac{\mathcal{C}\left(\frac{Y_{ti}-Y_{tj}}{b_{y_t}}\right)}{\sum_{j=1}^{n}\mathcal{C}\left(\frac{Y_{ti}-Y_{tj}}{b_{y_t}}\right)},\ t = 1, 2,$$

$$A_z(i,j) = \frac{\mathcal{K}\left(\frac{Z_i-Z_j}{b_z}\right)}{\sum_{j=1}^{n}\mathcal{K}\left(\frac{Z_i-Z_j}{b_z}\right)},$$

for $i = 1, \ldots, n$. For the univariate kernel, we use the Gaussian kernel and for the bandwidths we set b_{y_t} and b_z equal to $n^{-1/5}$ times the standard deviations of Y_t and Z, respectively.

Then (21) can be written as:

$$\begin{pmatrix} \gamma_n h_2^{-1} + A_{y_2}\left(I - \hat{P}_x\right)A_z h_2^{-1} - A_{y_2}\left(I - \hat{P}_x\right)A_z h_1^{-1} \\ -\gamma_n h_1^{-1} + P_n A_{y_1}\left(I - \hat{P}_x\right)A_z h_2^{-1} - P_n A_{y_1}\left(I - \hat{P}_x\right)A_z h_1^{-1} \end{pmatrix} = \hat{R}, \tag{24}$$

where

$$\hat{R} \equiv \begin{pmatrix} A_{y_2}\left(I - \hat{P}_x\right)A_z \Delta X_0 \\ P_n A_{y_1}\left(I - \hat{P}_x\right)A_z \Delta X_0 \end{pmatrix},$$

and P_n is the $n \times n$ matrix with $\frac{n-1}{n}$ on the diagonal and $-\frac{1}{n}$ elsewhere. It is used to impose Assumption 2 by projecting onto the space of functions of Y_1 where the mean is 0. Moreover, $\left(I - \hat{P}_x\right)$ is given by

$$\left(I - \hat{P}_x\right) = A_z \Delta X \left(\frac{\Delta X'}{n} A_z \Delta X\right)^{-1} \frac{\Delta X'}{n}.$$

Then the estimators $\left(\hat{h}_2^{-1}, \hat{h}_1^{-1}\right)$ are given by:

$$\begin{pmatrix} \hat{h}_2^{-1} \\ \hat{h}_1^{-1} \end{pmatrix} = \begin{pmatrix} \gamma_n I + A_{y_2}\left(I - \hat{P}_x\right) A_z & -A_{y_2}\left(I - \hat{P}_x\right) A_z \\ P_n A_{y_1}\left(I - \hat{P}_x\right) A_z & -\left(\gamma_n I + P_n A_{y_1}\left(I - \hat{P}_x\right) A_z\right) \end{pmatrix}^{-1} \hat{R}. \quad (25)$$

Given $\left(\hat{h}_2^{-1}, \hat{h}_1^{-1}\right)$, $\widehat{\beta}$ can be obtained by:

$$\widehat{\beta} = \left(\hat{K}_x^* \hat{K}_x\right)^{-1} \hat{K}_x^* \left[\hat{K}_y\left(\hat{h}_1^{-1}, \hat{h}_2^{-1}\right) - \hat{r}\right].$$

We suggest choosing the regularization parameter γ_n that minimizes the squared norm of residuals, following Florens and Sokullu (2017).

5 Asymptotic Properties

In this section, we derive assumptions for the $\sqrt{n}$- asymptotic normality of $\hat{\beta}$ and for the rate of convergence of $\left(\hat{h}_1^{-1}, \hat{h}_2^{-1}\right)$.

In this section, a subscript of 0 will denote the true value of the parameter being estimated.

Assumption 8. *The operator K_y is compact.*

This assumption allows us to use singular value decomposition (SVD) of the operator K_y.

Definition 1. Let $\mathcal{T} : \mathcal{E} \mapsto \mathcal{F}$ be a compact operator and let $\{\lambda_j, \phi_j, \psi_j\}$ be the singular system $\mathcal{T}$ such that:

$$\mathcal{T}\phi_j = \lambda_j \psi_j \quad \text{and} \quad \mathcal{T}^* \psi_j = \lambda_j \phi_j,$$

where λ_j denotes the sequence of the nonzero singular values of the compact linear operator $\mathcal{T}$, and ϕ_j and ψ_j, for all $j \in \mathbb{N}$, are orthonormal sequences of functions in $\mathcal{E}$ and $\mathcal{F}$, respectively. The *singular value decomposition* for each function $\varphi \in \mathcal{E}$ can be written as:

$$\mathcal{T}\varphi = \sum_{j=1}^{\infty} \lambda_j \langle \varphi, \phi_j \rangle \psi_j.$$

Given the definition above let $\{\lambda_j, \phi_j, \psi_j\}$ for $j \geq 1$ be the singular system of the operator K_y and let $\left\{\mu_l, e_l, \tilde{\psi}_l\right\}$ for $l = 1, 2, \ldots, k$ be the singular system of the operator K_x, such that for each $\beta \in \mathbb{R}^k$ we can write:

$$K_x \beta = \sum_{l=1}^{k} \mu_l \langle \beta, e_l \rangle \tilde{\psi}_l.$$

Assumption 9. *Source Condition: There exists $\nu > 0$ and $\eta > 0$ such that:*

$$\sum_{j=1}^{\infty} \frac{\left\langle \left(h_1^{-1}, h_2^{-1}\right), \phi_j \right\rangle^2}{\lambda_j^{2\nu}} = \sum_{j=1}^{\infty} \frac{\left(\left\langle h_1^{-1}, \phi_{1,j} \right\rangle + \left\langle h_2^{-1}, \phi_{2,j} \right\rangle\right)^2}{\lambda_j^{2\nu}} < \infty,$$

and

$$\max_{l=1,\ldots,k} \sum_{j=1}^{\infty} \frac{\left\langle \tilde{\psi}_l, \psi_j \right\rangle^2}{\lambda_j^{2\eta}} < \infty.$$

Assumption 9 is a common assumption in the NPIV literature and it defines a regularity space for the parameters of interest. The first equation in Assumption 9 defines a regularity space for $\left(h_1^{-1}, h_2^{-1}\right)$, in other words, this assumption adds a smoothness condition on the unknown functions. The second equation in Assumption 9 is about collinearity between Y_1, Y_2 and ΔX. As it is pointed out in Florens et al. (2012), η can be interpreted as a degree of collinearity between Y_1, Y_2 and ΔX measured through a projection on the instruments Z. For instance, when $\eta = \infty$,

$\mathcal{R}(K_y)$ and $\mathcal{R}(K_x)$ are orthogonal to each other and the estimation of β is not affected by the existence of the nonparametric component as $K_y^* K_x$ and $K_x^* K_y$ vanish from the normal equations (15) and (16).

Assumption 10. *The parameters ν, η in Assumption 9 satisfy $\nu \leq 2$ and $\eta \leq 2$.*

Assumption 10 is for the sake of exposition and it is without loss of generality. In this paper, we solve the ill-posed inverse problem we encounter during estimation using Tikhonov regularization. Since Tikhonov regularization has a qualification of two, we cannot improve upon the rate of convergence when the functions we consider have regularity greater than 2, i.e., $\nu, \eta > 2$. Hence, under this assumption during the derivation of the rates, we can simply write ν or η instead of $\min\{\nu, 2\}$ or $\min\{\eta, 2\}$.

Assumption 11. *Let s be the minimum between the order of the kernel used in estimation and the order of the differentiability of densities $f(Y_1, Y_2, Z), f(\Delta X, Z)$ and $f(\Delta X_0, Z)$ and assume that $s \geq 2$ and*

$$\begin{aligned}
\left\| \hat{K}_y - K_y \right\|^2 &= O_p\left(\frac{1}{n b_n^{q+1}} + b_n^{2s} \right), \\
\left\| \hat{K}_y^* - K_y^* \right\|^2 &= O_p\left(\frac{1}{n b_n^{q+1}} + b_n^{2s} \right), \\
\left\| \hat{K}_y^* \hat{r} - \hat{K}_y^* \hat{K}_y \left(h_{1,0}^{-1}, h_{2,0}^{-1} \right) \right\|^2 &= O_p\left(\frac{1}{n} + b_n^{2s} \right), \\
\left\| \hat{r} - r_0 \right\|^2 &= O_p\left(\frac{1}{n b_n^{q}} + b_n^{2s} \right),
\end{aligned}$$

where q is the dimension of the instrument vector Z and $b_{y_1} = b_{y_2} = b_z = b_n$ is the bandwidth.

Assumption 11 is a high-level assumption on the convergence rate of the estimated operators. Preliminary conditions leading to these rates have been studied in Darolles et al. (2011). Note that we set the bandwidths to be equal for exposition reasons. Below we state the rates we need for the smoothing parameters to converge to zero to obtain our final result.

Assumption 12. $\lim_{n\to\infty} \gamma_n \to 0$, $\lim_{n\to\infty} b_n^{2s} \to 0$, $\lim_{n\to\infty} nb_n^{q+1} \to \infty$, $\lim_{n\to\infty} n\gamma_n \to 0$, $\lim_{n\to\infty} n\gamma_n b_n^{2s} \to 0$, $\lim_{N\to\infty} \frac{\gamma_n}{b_n^{q+1}} \to 0$.

Assumption 13. $\mathcal{R}\left(K_y\right)^{\perp} = \mathcal{N}\left(K_y^*\right) \neq \{0\}$.

Assumption 13 implies that there exists an element ψ_j defined by the SVD of K_y such that $\psi_j \in \mathcal{R}(K_y)^{\perp}$. For example, this condition is satisfied in the joint nondegenerate normal case, i.e, if $(Y_1, Y_2, \Delta X, Z)$ is jointly distributed as a nondegenerate normal distribution. In such a case, the null space of K_y^* is $\{0\}$ if the range of the covariance with $(Y_1, Y_2, \Delta X)$ and Z is equal to the dimension of Z.

Assumption 14. *For $\theta > 0$, we have:* $\mathbb{E}\left[\left|U_2 - U_1\right|^{2+\theta} |Z\right] = c$, *for any $c \in \mathbb{R}$, and* $\mathbb{E}\left[\left|\left(I - P_y\right) K_x\right|^{2+\theta}\right] < \infty$.

Assumption 14 gives the conditions needed to satisfy the Liapounoff condition to apply the Liapounoff central limit theorem to obtain asymptotic normality of our estimators.

Using equation (22) we can show that:

$$\sqrt{n}\left(\hat{\beta} - \beta_0\right) = \hat{M}_\gamma^{-1}\left\{\sqrt{n}\left[K_x^*\left(I - P_y\right)\hat{E}\left(U_2 - U_1 | Z\right)\right] + O_p(1)\right\},$$

where

$$\hat{M}_\gamma \equiv \hat{K}_x^* \hat{K}_y\left(\gamma_n I + \hat{K}_y^* \hat{K}_y\right)^{-1} \hat{K}_y^* \hat{K}_x - \hat{K}_x^* \hat{K}_x,$$
$$\hat{E}\left(U_2 - U_1 | Z\right) \equiv r - \hat{K}_y\left(h_1^{-1}, h_2^{-1}\right) + \hat{K}_x \beta.$$

This decomposition is useful for the following result.

Theorem 3. *Assume that $Var\left(U_2 - U_1 | Z\right) = \sigma^2$. Moreover let Assumptions 9, 10, 11, 12, 13 and 14 hold. Then:*

$$\left\|\left(\hat{h}_1^{-1}, \hat{h}_2^{-1}\right)' - \left(h_{1,0}^{-1}, h_{2,0}^{-1}\right)'\right\|_{L^2}^2 = O_p\left(\frac{1}{\gamma_n^2}\left(\frac{1}{n} + b_n^{2s}\right) + \frac{1}{\gamma_n^2}\left(\frac{1}{nb_n^{q+1}} + b_n^{2s}\right)\gamma_n^\nu + \gamma_n^\nu\right),$$

and

$$\sqrt{n}\left(\hat{\beta}-\beta_0\right)\to\mathcal{N}\left(0,V\right),$$

where

$$V\equiv\sigma^2M^{-1}\left[\sum_j E\left(\Delta X\psi_j\right)E\left(\Delta X\psi_j\right)'\right]M^{-1},\psi\in\mathcal{R}\left(K_y\right)^{\perp},$$
$$M\equiv K_x^*K_y\left(K_y^*K_y\right)^{-1}K_y^*K_x-K_x^*K_x.$$

Proof. The proof can be found in Appendix B. □

Theorem 3 shows that a $\sqrt{n}$-convergence rate and asymptotic normality for $\hat{\beta}$ can be obtained, as well as showing the convergence rate of $\left(\hat{h}_1^{-1},\hat{h}_2^{-1}\right)$. Note that the estimator for h_t^{-1} is not necessarily monotone in its argument. We can make the estimator monotone by rearrangement. The weak convergence result obtained remains valid for the estimator obtained by rearrangement since the rearragement operator is Hadamard differentiable, see Chernozhukov et al. (2010).

Corollary 1. *Let assumptions 9 to 12 hold, and assume that* $s\geq 2\left(q+1\right)$ *and* $\gamma_n\sim n^{-\frac{3}{8}}$. *Then*

$$\left\|\left(\hat{h}_1^{-1},\hat{h}_2^{-1}\right)'-\left(h_{1,0}^{-1},h_{2,0}^{-1}\right)'\right\|_{L^2}^2=O_p\left(n^{-1/4}\right).$$

Proof. The proof can be found in the Appendix. □

Consider now the limiting distribution of the estimator of the APE defined in (23) above. The APE is characterized by the moment condition

$$E\left[h_t\left(h_t^{-1}\left(Y_{it}\right)-\beta_k\right)-Y_{it}-\delta_{k,t}\right]=0.$$

Then, given a random sample $\{Y_{it}\}_{i=1}^n$ and estimators $\hat{\beta},\hat{h}_t$, the APE $\delta_{k,t}$ can be estimated by the zero of the estimating equation below:

$$\frac{1}{n}\sum_{i=1}^{n}\left(\hat{h}_t\left(\hat{h}_t^{-1}\left(Y_{it}\right)-\hat{\beta}_k\right)-Y_{it}-\delta_{k,t}\right)=0.$$

This shows that $\hat{\delta}_{k,t}$ is a plug-in two-step Z-estimator. That this estimator can be shown to be $\sqrt{n}$-asymptotically normal should be no surprise given the regularity conditions on h_t, the way that $\delta_{k,t}$ enters the estimating equation, and the rate results on $\hat{\beta}$ and $\hat{h}_t^{-1}$.

A general result on two-step Z-estimators can be found in Chen et al. (2003). In that paper, Theorems 1 and 2 state sufficient high-level conditions under which $\hat{\delta}_{k,t}$ can be shown to be consistent and $\sqrt{n}$-asymptotically normal.[8] Primitive conditions for those high-level assumptions will be provided in due course, for now we note that it is well known that checking the conditions in Chen et al. (2003) can be quite difficult particularly when kernel estimators are used. For example, for a class of transformation models with exogenous regressors where the objects of interest are estimated via kernel estimators, Colling and Van Keilegom (2019, 2020) provide low-level assumptions that satisfy the conditions of Theorems 1 and 2 in Chen et al. (2003). We conjecture that primitive conditions similar to those there can be derived for our set-up as well. Instead, here we make high-level assumptions as in Theorem 2 in Chen et al. (2003), in order to state our result on the $\sqrt{n}$-asymptotic normality of $\hat{\delta}_{k,t}$. Our simulation studies in Section 6 confirm that $\hat{\delta}_{k,t}$ is $\sqrt{n}$-asymptotically normal, e.g. Figures 3 and 4.

Using the notation in Chen et al. (2003) and assuming that h_t is differentiable

[8] See also van der Vaart and Wellner (2007).

on its support, define the following objects:

$$
\begin{aligned}
M\left(\delta_{k,t},h_t,\beta_k\right) &\equiv E\left[m\left(\delta_{k,t},h_t,\beta_k\right)\right] \\
&\equiv E\left[h_t\left(h_t^{-1}\left(Y_t\right)-\beta_k\right)-Y_t-\delta_{k,t}\right], \\
M_n\left(\delta_{k,t},h_t,\beta_k\right) &\equiv \frac{1}{n}\sum_{i=1}^{n}\left(h_t\left(h_t^{-1}\left(Y_{it}\right)-\beta_k\right)-Y_{it}-\delta_{k,t}\right), \\
\Gamma_1\left(\delta_{k,t},h_t,\beta_k\right) &\equiv \frac{\partial}{\partial\delta_{k,t}}M\left(\delta_{k,t},h_t,\beta_k\right)=-1, \\
\Gamma_2\left(\delta_{k,t},h_t,\beta_k\right)\left[\bar{h}_t-h_t\right] &\equiv \left.\frac{d}{d\gamma}M\left(\delta_{k,t},h_t+\gamma\left(\bar{h}_t-h_t\right),\beta_k\right)\right|_{\gamma=0} \\
&= E\left[\left(1-\frac{h_t'\left(h_t^{-1}\left(Y_t\right)-\beta_k\right)}{h_t^2\left(Y_t\right)}\right)\left[\bar{h}_t\left(Y_t\right)-h_t\left(Y_t\right)\right]\right], \\
\Gamma_3\left(\delta_{k,t},h_t,\beta_k\right) &\equiv \frac{\partial}{\partial\beta_k}M\left(\delta_{k,t},h_t,\beta_k\right) \\
&= -E\left[h_t'\left(h_t^{-1}\left(Y_t\right)-\beta_k\right)\right],
\end{aligned}
$$

where h_t' is the first derivative of h_t with respect to its argument.

Theorem 4. *Let the assumptions of Corollary 1 hold, and assume that (i) h_t is continuously differentiable on its support, and Lipschitz continuous with a uniformly bounded derivative for $t=1,2$; (ii) the density of Y_t is bounded away from zero and is bounded from above for $t=1,2$; (iii) for $t=1,2$,*

$$
\begin{aligned}
&\left\|M\left(\delta_{k,t},h_t,\beta_k\right)-M\left(\delta_{k,t},h_{t0},\beta_{k0}\right)-\Gamma_2\left(\delta_{k,t},h_{t0},\beta_{k0}\right)\left[h_t-h_{t0}\right]-\Gamma_3\left(\delta_{k,t},h_{t0},\beta_{k0}\right)\right\| \\
&\qquad\leq c\left(\left\|h_t-h_{t0}\right\|_{L^2}^2+\left\|\beta_k-\beta_{k0}\right\|^2\right);
\end{aligned}
$$

(iv) for $t=1,2$, and some finite matrix V_1,

$$
\sqrt{n}\left(M_n\left(\delta_{k,t,0},h_{t0},\beta_{k0}\right)+\Gamma_2\left(\delta_{k,t,0},h_{t0},\beta_{k0}\right)\left[\hat{h}_t-h_{t0}\right]+\Gamma_3\left(\delta_{k,t,0},h_{t0},\beta_{k0}\right)\right)\to\mathcal{N}\left(0,V_1\right).
$$

Then $t=1,2$,

$$
\sqrt{n}\left(\hat{\delta}_{k,t}-\delta_{k,t}\right)\to\mathcal{N}\left(0,V_1\right).
$$

Proof. The proof can be found in the Appendix. □

6 Simulation

In this section we illustrate the small sample performance of our proposed estimator through Monte Carlo simulations. We consider two data generating processes (DGP). In the first one we simulate a panel model with fixed effects (introduced in Section 1) and in the second one we simulate a dynamic panel model as in Equations (3) and (4).

6.1 DGP1

We consider the case of $T = 2$.

Let

$$
\begin{aligned}
(Z_1, Z_2) &\sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right), \\
\xi &\sim \mathcal{U}\left[0, 1\right], \\
(\omega_1, \omega_2) &\sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_\omega^2 & 0 \\ 0 & \sigma_\omega^2 \end{pmatrix}\right), \quad \sigma_\omega^2 = 0.5, \\
(U_1, U_2) &\sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_u^2 & 0 \\ 0 & \sigma_u^2 \end{pmatrix}\right), \quad \sigma_u^2 = 0.6,
\end{aligned}
$$

so that

$$
\begin{aligned}
X_{01} &= 0.7Z_1 + 0.5U_1 + \xi, \\
X_{02} &= 0.8Z_2 + 0.4U_2 + \xi + 20, \\
X_1 &= 0.8Z_1 + 0.7Z_2 + \omega_1 + U_1, \\
X_2 &= 0.7Z_1 + 0.8Z_2 + \omega_2 + U_2.
\end{aligned}
$$

Additionally, let

$$\alpha \sim \mathcal{N}(0,1) + \frac{1}{2}(X_1 + X_2),$$

and $h_1(s) = s$, $h_2(s) = \log(s)$, $\beta = 1$, so that

$$Y_{i1} = \alpha_i + X_{i01} + \beta X_{i1} + U_{i1},$$
$$Y_{i2} = \log(\alpha_i + X_{i02} + \beta X_{i2} + U_{i2}),$$

for $i = 1, \ldots, n$.

We simulate the model 500 times for sample sizes $n \in \{100, 200, 500, 1000\}$. We estimate the functions h_1, h_2 and the finite dimensional parameter β following the method described in Section 4. We impose monotonicity of the infinite dimensional parameters by the use of rearrangement. We choose the regularization parameter so as to minimize the squared norm of residuals, following Florens and Sokullu (2017). Figures 1 and 2 show the estimated functions $\hat{h}_1^{-1}$ and $\hat{h}_2^{-1}$, respectively. The light gray shaded area shows the estimated curves obtained at each draw plotted pointwise, dark gray dots show the pointwise average across simulations of the estimated functions, i.e. $\frac{1}{500}\sum_{s=1}^{500} \hat{h}_{s,t}(y_t^*), t = 1, 2$, whereas the black dots show the true (pointwise) function. Table 3 shows the mean, standard error, and root mean square error (RMSE) of $\hat{\beta}$ for different sample sizes. As expected, both bias and standard deviation decreases with increasing sample size.

After obtaining $\hat{h}_1^{-1}$, $\hat{h}_2^{-1}$and $\hat{\beta}$, we compute $\hat{\delta}_{k,2}$ as in (23). Table 2 shows the mean, standard error, and RMSE of estimated average partial effects for different sample sizes as well as the true average partial effect at $t = 2$ which is calculated using true values of h_1, h_2, and β. We show two different figures, one for $n = 500$ (Figure 3) and one for $n = 1000$ (Figure 4), which provide suggestive evidence that our estimator of the APE attains $\sqrt{n}$-asymptotic normality.

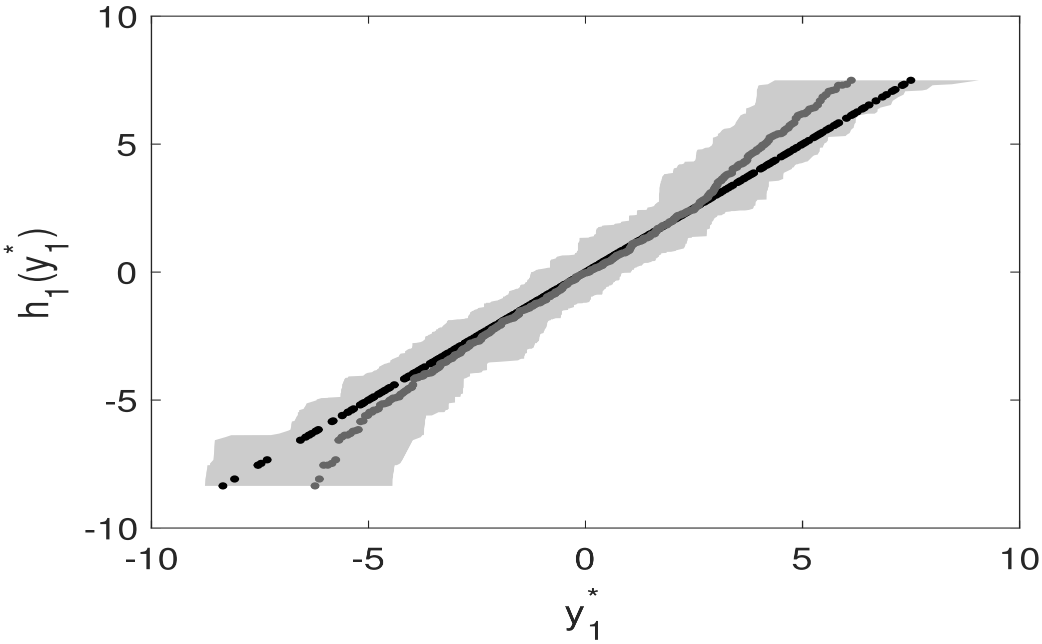

Figure 1: **Simulation result with** 500 **draws for** h_1**, monotonicity imposed by rearrangement**, $n = 500$.

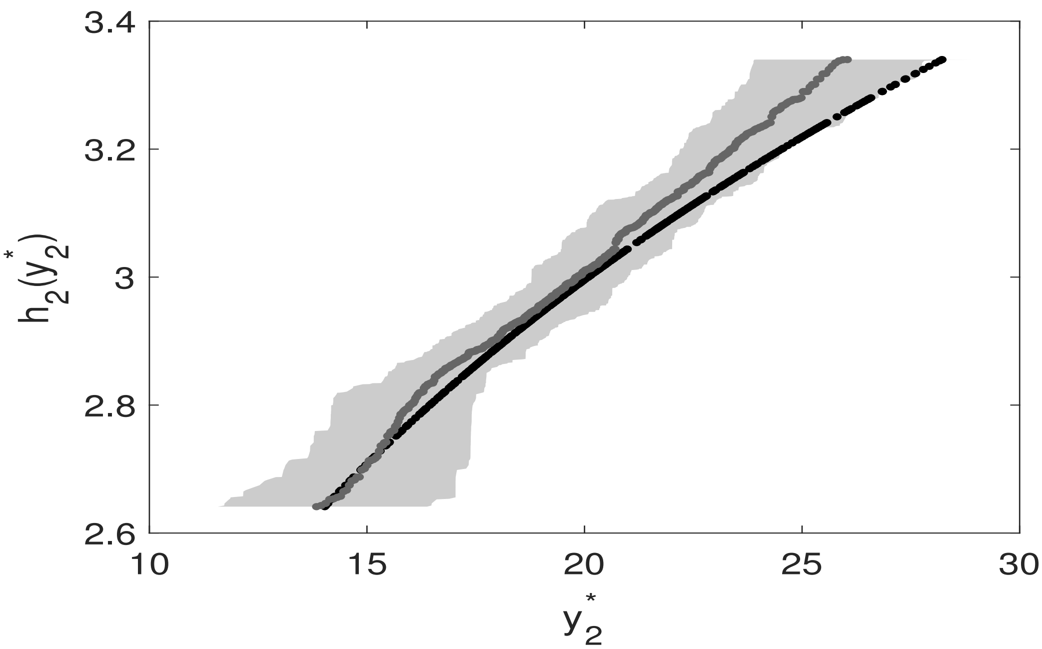

Figure 2: **Simulation result with** 500 **draws for** h_2**, monotonicity imposed by rearrangement,** $n = 500$**.**

Table 1: Estimation results for β

	Mean	Std. Err
$n = 100$	0.8614	0.2767
$n = 200$	0.9696	0.2145
$n = 500$	1.0363	0.1736
$n = 1000$	1.0583	0.1326

Table 2: Estimation results for APE

	Mean	Std. Err	RMSE	True APE
$n = 100$	0.0589	0.0165	0.0186	0.0505
$n = 200$	0.0612	0.0113	0.0154	0.0506
$n = 500$	0.0607	0.0084	0.0131	0.0506
$n = 1000$	0.0597	0.0062	0.0109	0.0506

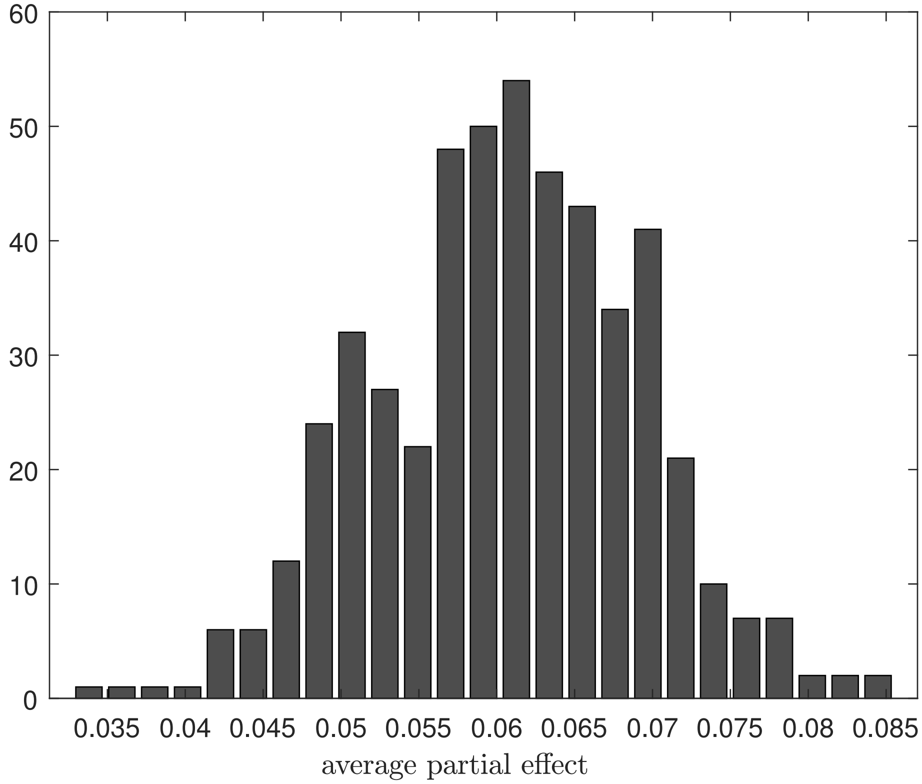

Figure 3: **Histogram of $\hat{APE}$ for n=500.**

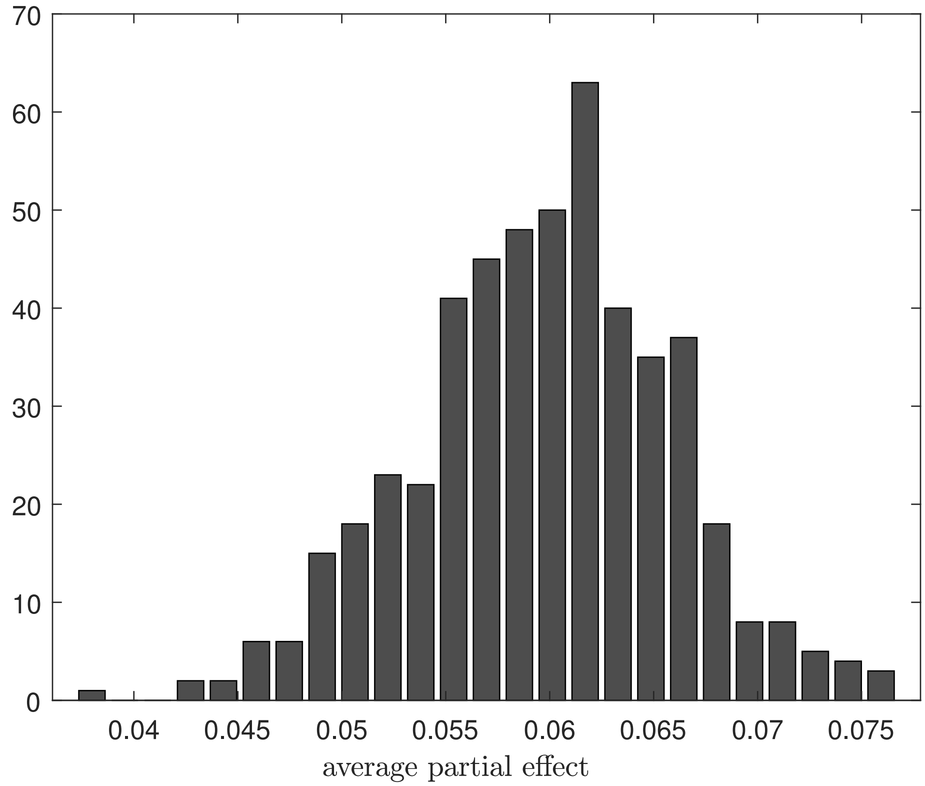

Figure 4: **Histogram of $A\hat{P}E$ for n=1000.**

6.2 DGP2

In this DGP, we simulate a dynamic nonlinear panel model as following:

$$Y_0 \sim \mathcal{N}(0,1)$$

$$X \sim \mathcal{N}(0, \mathcal{I}_T\sigma^2), \quad U \sim \mathcal{N}(0, \mathcal{I}_T\sigma^2), \quad \alpha \sim \mathcal{N}(0, \sigma^2)$$

where $\sigma^2 = (1-\beta^2)/3$. Moreover, $Y^*_{it} = \alpha_i + X_{it} + \beta Y_{i,t-1} + U_{it}$, where we set $\beta = 0.6$. We generate the model up to $T = 3$.

$$Y_{it} = h_t(Y^*_{it}) = Y^*_{it} \quad for \quad t = 1, 2$$

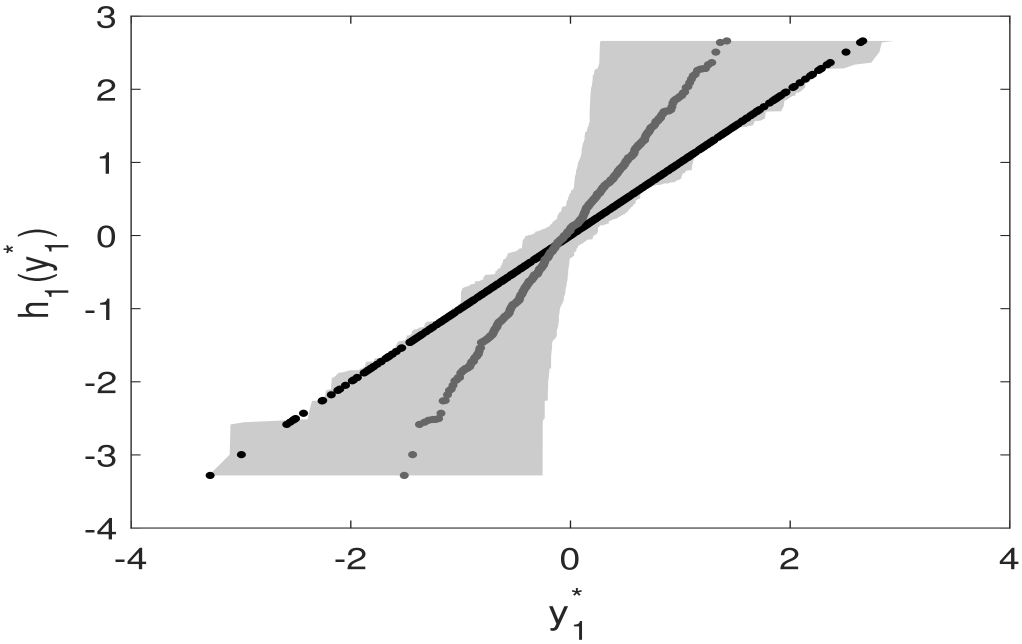

Figure 5: **Simulation result with** 500 **draws for** h_1**, monotonicity imposed by rearrangement**, $n = 500$, DGP2

$$Y_{it} = h_t(Y_{it}^*) = \frac{\exp(Y_{it}^*)}{1 + \exp(Y_{it}^*)} \quad for \quad t = 3.$$

As in the first DGP, we generate samples of sizes 100, 200, 500 and 1000 and we replicate the simulation for 500 times. We estimate h_2, h_3 and β.

Figures 5 and 6 show the results for $\hat{h}_2^{-1}$ and $\hat{h}_3^{-1}$ for a sample size of 500. As before, Table 3 shows the the mean, standard error, and root mean square error of $\hat{\beta}$ and Table 4 shows the mean, standard error, and RMSE of the estimated APE $\hat{\delta}_{k3}$ as well as the true APE at $t = 3$.

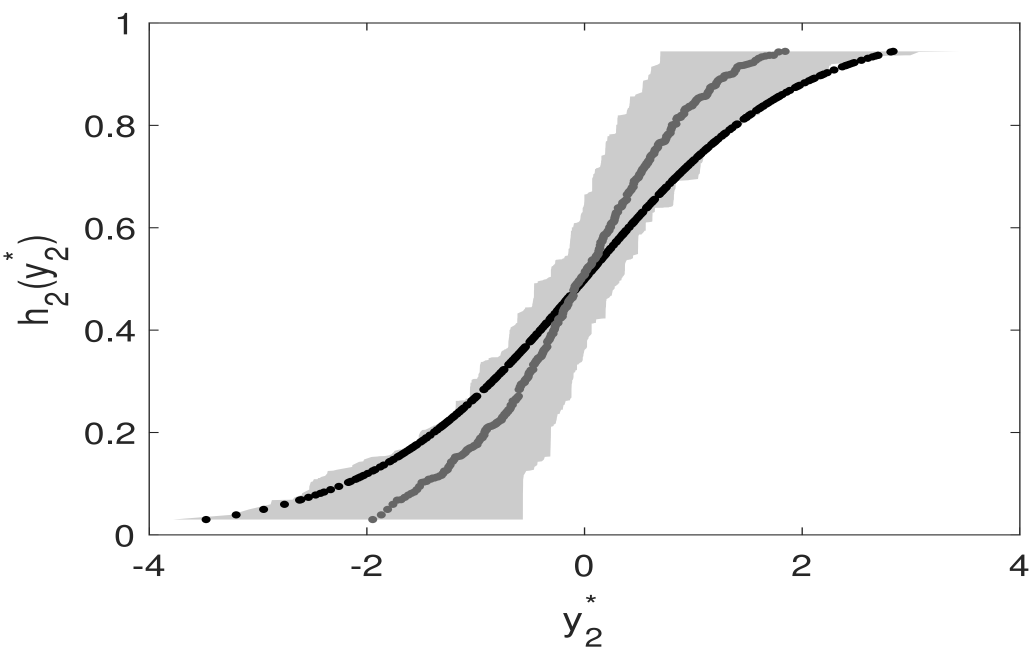

Figure 6: **Simulation result with** 500 **draws for** h_2**, monotonicity imposed by rearrangement**, $n = 500$, DGP2

Table 3: Estimation results for β

	Mean	Std. Err
$n = 100$	0.5069	0.1439
$n = 200$	0.5496	0.1202
$n = 500$	0.5851	0.1106
$n = 1000$	0.6122	0.1114

Table 4: Estimation results for APE

	Mean	Std. Err	RMSE	True APE
$n = 100$	0.1610	0.0800	0.0075	0.1938
$n = 200$	0.1800	0.0802	0.0066	0.1932
$n = 500$	0.1881	0.0800	0.0064	0.1931
$n = 1000$	0.1934	0.0807	0.0065	0.1927

7 Empirical Illustration

In this section we illustrate our method by estimating the model in Arellano and Bond (1991). Using their data we estimate the following model:

$$n_{it} = h_t(\alpha_i + ys_{it} + \beta_1 n_{it-1} + \beta_2 w_{it} + \beta_3 k_{it} + \epsilon_{it}) \tag{26}$$

where n_{it} is the logarithm of UK employment in company i at the end of year t, ys_{it} is log of industry output, w_{it} is the log of real product wage and k_{it} is the log of gross capital of company i in year t. Finally, ϵ_{it} is the error term.

We also estimate a linear model specified as:

$$n_{it} = \beta_0 + \beta_1 n_{it-1} + \beta_2 w_{it} + \beta_3 k_{it} + \beta_4 ys_{it} + \alpha_i + \lambda_t + u_{it} \tag{27}$$

For the nonlinear regression we normalized the coefficient of ys_{it} to 1. We estimate the above models for rolling years of 1979-1980, 1980-1981 and 1981-1982. For each year pair we have 140 firms, hence $n = 140$ and $T = 2$. The estimates for the linear specification can be found in Table 3, while the estimates for the nonlinear model can be found in Table 4 and Figure 7.

Table 6: Estimation results for Linear Model

	$\hat{\boldsymbol{\beta}}_0$	$\hat{\boldsymbol{\beta}}_1$	$\hat{\boldsymbol{\beta}}_2$	$\hat{\boldsymbol{\beta}}_3$	$\hat{\boldsymbol{\beta}}_4$
79-80	0.0406	-3.1484	-0.1599	-0.0864	1.0848
	[-1.47 1.55]	[-167.41 161.13]	[-2.18 1.86]	[-18.21 18.05]	[-23.6 25.76]
80-81	0.0060	-4.3646	-1.6789	1.2894	1.3397
	[-0.21 0.23]	[-58.16 49.44]	[-13.23 9.85]	[-7.63 10.21]	[-9.17 11.85]
81-82	2.605	27.849	-2.018	-5.082	2.686
	[-88.37 93.59]	[-934.37 990.07]	[-49.08 45.04]	[-192.1 181.94]	[-83.41 88.79]

Table 5: Estimation results for Nonlinear Model

	$\hat{\boldsymbol{\beta}}_1$	$\hat{\boldsymbol{\beta}}_2$	$\hat{\boldsymbol{\beta}}_3$
79-80	-0.1039	-0.4404	-0.0502
	[-0.3250 0.1173]	[-0.6151 -0.2656]	[-0.1470 0.0465]
80-81	0.0543	-0.3042	-0.0454
	[-0.0902 0.1987]	[-0.4461 -0.1622]	[-0.1252 0.0345]
81-82	0.0055	0.0512	-0.0173
	[-0.0479 0.0589]	[-0.0855 0.1878]	[-0.0785 0.0437]

We calculate the individual counterfactual outcomes and the average effects using the expression given in (11). We use the estimation results from the years 1979-1980. We present our results in Figure 7 below. The increasing curves represent

$$\hat{h}_t\left(\hat{h}_t^{-1}\left(Y_{it}\right)+\hat{\beta}_k\right)-Y_{it}$$

appearing in (12), for $t = 1979, 1980$, while the dashed lines represent the cross-sectional average of the expression above, i.e. the estimate of (12). The black straight line represents the average partial effect for the linear model.

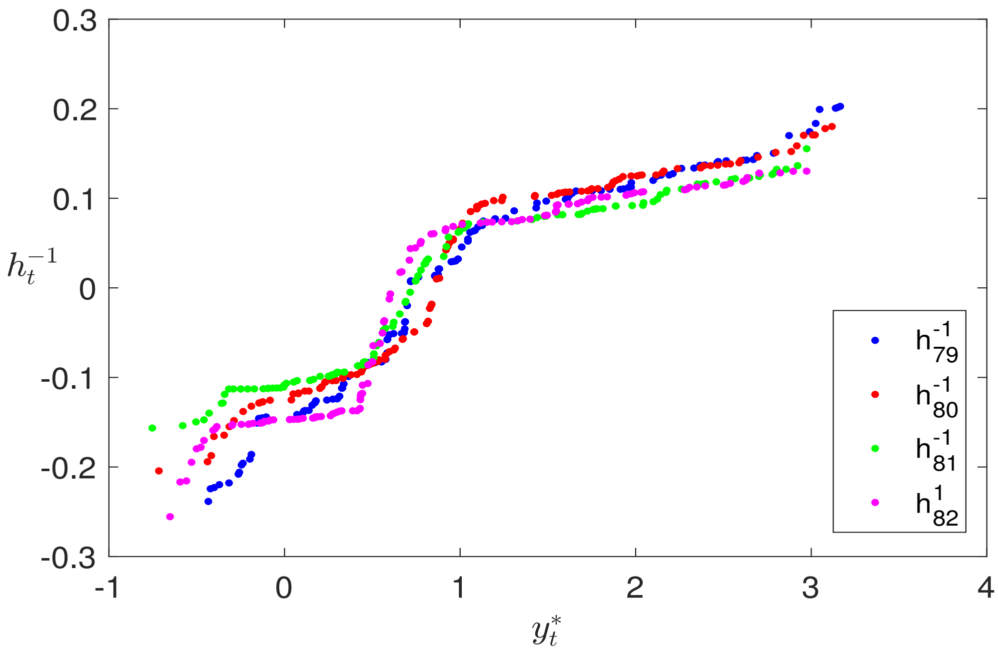

Figure 7: Estimated h_t^{-1} functions

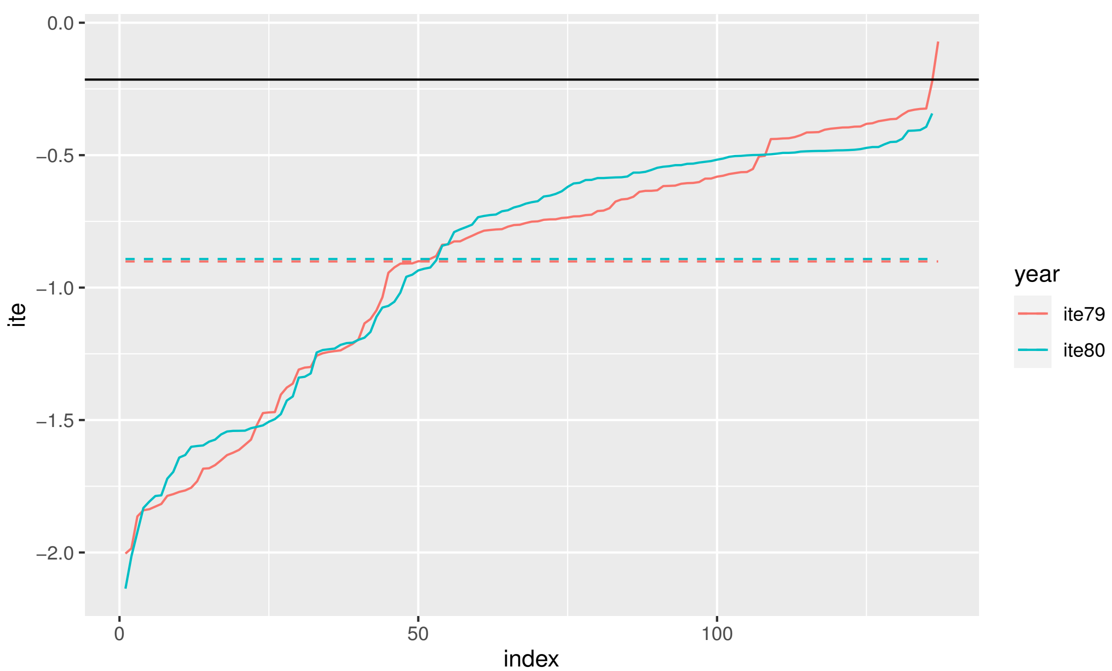

Figure 8: Average partial effects for the linear specification (black, straight line) and nonlinear specification (dashed)

References

Abrevaya, J. (1999): "Leapfrog estimation of a fixed-effects model with unknown transformation of the dependent variable," *Journal of Econometrics*, 93, 203–228.

Aguirregabiria, V. and J. Carro (2021): "Identification of Average Marginal Effects in Fixed Effects Dynamic Discrete Choice Models," Https://arxiv.org/abs/2107.06141.

Aguirregabiria, V., J. Gu, and Y. Luo (2021): "Sufficient Statistics for Unobserved Heterogeneity in Structural Dynamic Logit Models," *Journal of Econometrics*, 223, 280–311.

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

Andrews, D. W. (2017): "Examples of L2-complete and boundedly-complete distributions," *Journal of econometrics*, 199, 213–220.

Arellano, M. and S. Bond (1991): "Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations," *The Review of Economic Studies*, 58, 277–297.

Arellano, M. and S. Bonhomme (2011): "Nonlinear Panel Data Analysis," *Annual Review of Economics*, 3, 395–424.

Arellano, M. and B. Honoré (2001): "Panel Data Models: Some Recent Developments," in *Handbook of Econometrics*, ed. by J. J. Heckman and E. Leamer, Elsevier, vol. 5, 3229–3296, zSCC: NoCitationData[s0].

Athey, S. and G. W. Imbens (2006): "Identification and inference in nonlinear difference-in-differences model," *Econometrica*, 74, 431–497.

Babii, A. and J.-P. Florens (2020): "Is completeness necessary? Estimation in nonidentified linear models," Working paper.

Birke, M., S. Van Bellegem, and I. Van Keilegom (2017): "Semi-parametric estimation in a single-index model with endogenous variables," *Scandinavian Journal of Statistics*, 44, 168–191.

Blundell, R. and S. Bond (1998): "Initial conditions and moment restrictions in dynamic panel data models," *Journal of econometrics*, 87, 115–143.

Botosaru, I. and C. Muris (2017): "Binarization for panel models with fixed effects," Cemmap working paper CWP31/17.

Botosaru, I., C. Muris, and K. Pendakur (2021): "Identification of Time-Varying Transformation Models with Fixed Effects, with an Application to Unobserved Heterogeneity in Resource Shares," Tech. rep.

Bun, M. and V. Sarafidis (2015): "Dynamic panel data models," *The Oxford handbook of panel data*, 76–110.

Carrasco, M. and J. Florens (2011): "A Spectral Method for Deconvolving a Density," *Econometric Theory*, 27, 546–581.

Centorrino, S., F. Féve, and J.-P. Florens (2017): "Additive Nonparametric Instrumental Regressions: A Guide to Implementation," *Journal of Econometric Methods*, 6, 1–25.

Chen, X., V. Chernozhukov, S. Lee, and W. Newey (2014): "Local Identification of Nonparametric and Semiparametric Models," *Econometrica*, 2, 785–809.

Chen, X., O. Linton, and I. Van Keilegom (2003): "Estimation of Semiparametric Models when the Criterion Function Is Not Smooth," *Econometrica*, 71, 1591–1608.

Chernozhukov, V., I. Fernández-Val, and A. Galichon (2010): "Quantiles and Probability Curves without Crossing," *Econometrica*, 78, 1093–1125.

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

Chiappori, P.-A., I. Komunjer, and D. Kristensen (2015): "Nonparametric identification and estimation of transformation models," *Journal of Econometrics*, 188, 22–39.

Colling, B. and I. Van Keilegom (2019): "Estimation of fully nonparametric transformation models," *Bernoulli*, 25, 3762–3795.

——— (2020): "Estimation of a semiparametric transformation model: A novel approach based on least squares minimization," *Electornic Journal of Statistics*, 14, 769–800.

Darolles, S., Y. Fan, J.-P. Florens, and E. Renault (2011): "Nonparametric instrumental regression," *Econometrica*, 79, 1541–1565.

Davezies, L., X. d'Haultfoeuille, and L. Laage (2021): "Identification and Estimation of Average Marginal Effects in Fixed Effects Logit Models," Https://arxiv.org/abs/2105.00879.

d'Haultfoeuille, X. (2010): "A new instrumental method for dealing with endogenous selection," *Journal of Econometrics*, 154, 1–15.

——— (2011): "On the completeness condition in nonparametric instrumental problems," *Econometric Theory*, 27, 460–471.

Dobronyi, C., J. Gu, and K. il Kim (2021): "Identification of Dynamic Panel Logit Models with Fixed Effects," Https://arxiv.org/abs/2104.04590.

Féve, F. and J.-P. Florens (2010): "The practice of nonparametric estimation by solving inverse problems: the example of transformation models," *The Econometrics Journal*, 13, S1–S27.

——— (2014): "Non parametric analysis of panel data models with endogenous variables," *Journal of Econometrics*, 181, 151–164.

Florens, J.-P., J. Johannes, and S. Van Bellegem (2012): "Instrumental regression in partially linear models," *The Econometrics Journal*, 15, 304–324.

Florens, J.-P. and S. Sokullu (2017): "Nonparametric estimation of semiparametric transformation models," *Econometric Theory*, 33, 839–873.

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

Horowitz, J. L. (2009): *Semiparametric and Nonparametric Methods in Econometrics*, Springer.

Hu, Y. and J.-L. Shiu (2018): "Nonparametric identification using instrumental variables: sufficient conditions for completeness," *Econometric Theory*, 34, 659–693.

Liu, L., A. Poirier, and J.-L. Shiu (2021): "Identification and Estimation of Average Partial Effects in Semiparametric Binary Response Panel Models," Https://arxiv.org/abs/2105.12891.

Newey, W. K. and J. L. Powell (2003): "Instrumental variable estimation of nonparametric models," *Econometrica*, 71, 1565–1578.

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

Pakel, C. and M. Weidner (2021): "Bounds on Average Effects in Discrete Choice Panel Data Models," .

van der Vaart, A. and J. Wellner (1996): *Weak convergence and Empirical processes*, Springer.

Vanhems, A. and I. Van Keilegom (2019): "Estimation of a semiparametric transformation model in the presence of endogeneity," *Econometric Theory*, 35, 73–110.

A Additional results

A.1 Measurable separability

Assumption 15. *The random variables* $W \equiv (Y_1, Y_2, \Delta X)$ *are such that for any* $\delta_2 - \delta_1 \in L^2_W$ *and any* $b \in \mathbb{R}^k$ *, if*

$$\delta_2(Y_2) - \delta_1(Y_1) - \Delta X b = 0 \text{ a.s. } F_{Y_1,Y_2,\Delta X},$$

then there exist constants $c_t \in \mathbb{R},\ t = 1, 2,$ *such that*

$$\delta_t(Y_t) = c_t \text{ a.s. } F_{Y_t},\ t = 1, 2.$$

Assumption 15 is a high-level assumption that rules out a linear relationship between $Y_1, Y_2,$ and ΔX. The assumption is a slightly weaker version of the measurable separability assumption made in the NPIV literature. The assumption fails if there exists an *additive* functional relationship between Y_1, Y_2, and ΔX, see, e.g., Newey et al. (1999).[9] Identification may still occur in the presence of a nonadditive functional relationship between the three random variables. The Lemma below establishes sufficient low-level assumptions for Assumption 15.

[9] For example, Newey et al. (1999) write that there exists a functional relationship between two random variables W_1 and W_2 provided that there exist functions $H(W_1, W_2)$ and a set $\mathcal{W}$ such that $P(\mathcal{W}) > 0$ and

$$P(H(W_1, W_2) = 0) = 1$$
$$P\left(H\left(W_1, \bar{W}_2\right) = 0\right) < 1$$

for all fixed $\bar{W}_2 \in \mathcal{W}$. In fact, Assumption 15 is implied by two measurable separability assumptions: one between (Y_1, Y_2) and ΔX, and another between Y_1 and Y_2.

Lemma 1. *Let the following assumptions hold: (L1) h_t is continuously differentiable for all t; (L2) the support of X_t contains an open set and is continuous on that set; (L3) U_t is continuous for all t and is serially independent. Then for any $h_1,\ \ h_2,\ \beta$ satisfying Assumptions 1 to 5 and Assumptions L1, L2, L3, for any random variables $Y_t,\ X_t,\ Z$ following the model above, and for any $\delta_2 - \delta_1 \in L_W^2$ and for any $b \in \mathbb{R}^k$, if*

$$\delta_2\left(Y_2\right) - \delta_1\left(Y_1\right) - \Delta X b = 0 \ a.s. \ F_{Y_1,Y_2,\Delta X}, \tag{28}$$

then there exist constants $c_t \in \mathbb{R}$ such that $\delta_t\left(Y_t\right) = c_t$ a.s. F_{Y_t}, $t = 1, 2$.

Proof. The conclusion of the Lemma follows by contradiction. That is, assuming both (28) and $\delta_1\left(Y_1\right) \neq c_1$ a.s. or $\delta_2\left(Y_2\right) \neq c_2$ a.s. for all $c_1, c_2 \in \mathbb{R}$, leads to a contradiction.

First, solve for α from the outcome equation for Y_1 and plug the resulting expression in the outcome equation for Y_2 to obtain:

$$Y_2 = h_2\left(h_1^{-1}\left(Y_1\right) + \left(X_2 - X_1\right)\beta + U_2 - U_1\right).$$

Consider then (28):

$$\delta_2\left(h_2\left(h_1^{-1}\left(y_1\right) + \left(x_2 - x_1\right)\beta - u_1 + u_2\right)\right) - \delta_1\left(y_1\right) \equiv \left(x_2 - x_1\right)b, \tag{29}$$

for all $x_t \in \mathcal{X}_t, y_t \in \mathcal{Y}_t, u_t \in \mathcal{U}_t$, $t = 1, 2$.

First, note that since X_2 and U_2 are correlated, we can think of X_2 as a function of U_2, e.g., $X_2 = \gamma\left(U_2\right) + \eta_2$, $\eta_2 = X_2 - \gamma\left(U_2\right)$. Second, note that h_t being differentiable guarantees that δ_t is also differentiable. Then differentiating (29) wrt u_2 obtains

$$\frac{\partial \delta_2}{\partial h_2}\left(\frac{\partial h_2}{\partial x_2}\frac{\partial \gamma_2}{\partial u_2}\beta + \frac{\partial h_2}{\partial u_2}\right) = \frac{\partial \gamma_2}{\partial u_2}b, \tag{30}$$

where we used Assumptions L1, L2, and L3, and that X_2 is correlated with U_2.

However, since $\delta_2(Y_2) \neq c_2$ a.s. it follows that

$$\frac{\partial \delta_2}{\partial h_2}\left(\frac{\partial h_2}{\partial x_2}\frac{\partial \gamma_2}{\partial u_2}\beta + \frac{\partial h_2}{\partial u_2}\right) \neq 0. \tag{31}$$

Combining (30) and (31), it must be that for all $b \in \mathbb{R}^k$,

$$\frac{\partial \gamma_2}{\partial u_2} b \neq 0.$$

Since X_2 is correlated with U_2, $\frac{\partial \gamma_2}{\partial u_2} \neq 0$. Hence it follows that $b \neq 0$, which is not true since $b \in \mathbb{R}^k$.

Similarly, we can show that assuming (28) and $\delta_1(Y_1) \neq c_1$ for all $c_1 \in \mathbb{R}$ leads to a contradiction. □

A.2 Normal equations

The normal equations using the three operators K_x, K_{y_1}, and K_{y_2} are:

$$K_{y_1}^* r = K_{y_1}^* K_{y_2} h_2^{-1} - K_{y_1}^* K_{y_1} h_1^{-1} - K_{y_1}^* K_x \beta, \tag{32}$$
$$K_{y_2}^* r = K_{y_2}^* K_{y_2} h_2^{-1} - K_{y_2}^* K_{y_1} h_1^{-1} - K_{y_2}^* K_x \beta, \tag{33}$$
$$K_x^* r = K_x^* K_{y_2} h_2^{-1} - K_x^* K_{y_1} h_1^{-1} - K_x^* K_x \beta. \tag{34}$$

Notice that (34) can be written as

$$K_x^* r = K_x^*\left(K_{y_2} h_2^{-1} - K_{y_1} h_1^{-1}\right) - K_x^* K_x \beta = K_x^* K_y \left(h_1^{-1}, h_2^{-1}\right) - K_x^* \beta,$$

where we used the definition of K_y. The expression above is (16) in the main text.

Consider now (32) and (33), and rewrite them as

$$K_{y_1}^* K_y \left(h_1^{-1}, h_2^{-1}\right) = K_{y_1}^* r + K_{y_1}^* K_x \beta,$$
$$K_{y_2}^* K_y \left(h_1^{-1}, h_2^{-1}\right) = K_{y_2}^* r + K_{y_2}^* K_x \beta.$$

Imposing Assumption 2(ii), multiplying the second equation above by -1, and using the definition of K_y^*, obtains equation (15) in the main text.

B Proofs

B.1 Proof of Theorem 1

Let (h_1, h_2, β) be the true value of the model parameters, and let $(g_1, g_2, B) \in L^2_{\mathcal{Y}_1} \otimes L^2_{\mathcal{Y}_2} \otimes \mathbb{R}^k$ be alternative values such that

$$(g_1, g_2, B) \neq (h_1, h_2, \beta)$$

and such that they satisfy the same assumptions as (h_1, h_2, β), i.e. assumptions 1, 2, 3, and 4. In particular, for any $z \in \mathcal{Z}$:

$$E\left(g_2^{-1}(Y_2) - g_1^{-1}(Y_1) - \Delta X B \big| Z = z \right) = E\left(\Delta X_0 | Z = z\right). \tag{35}$$

Equating (6) and (35), and re-arranging yields

$$E\left(\delta_2(Y_2) - \delta_1(Y_1) - \Delta X b | Z = z\right) = 0,$$

where

$$\delta_t(Y_t) \equiv h_t^{-1}(Y_t) - g_t^{-1}(Y_t), t = 1, 2, \tag{36}$$

$$b \equiv \beta - B. \tag{37}$$

Assumption 5 obtains that

$$\delta_2(Y_2) = 0,\ \delta_1(Y_1) = 0,\ \Delta X b = 0 \text{ a.s. } F_{Y_1, Y_2, \Delta X}. \tag{38}$$

We show now that Assumption 5 is equivalent to Assumptions 6(i) and 6(ii).

First, we show by contradiction that Assumption 5 implies Assumptions 6(i) and 6(ii). Suppose first that Assumption 5 holds and that 6(i) does not. Since, K is injective,

$$K(\delta_1, \delta_2, b) = K_{y_2}\delta_2 - K_{y_1}\delta_1 - K_x b = 0 \text{ a.s.} \implies (\delta_1, \delta_2, b) = (0, 0, 0) \text{ a.s.}$$

Additionally, since $\{0\} \in \cap_{t=1}^{2} \mathcal{R}(K_{y_t}) \cap \mathcal{R}(K_x)$, it follows that

$$K_{y_t}\delta_t = 0 = K_x b \text{ a.s.}$$

However, since K_{y_t} and K_x are not injective, it follows that $\delta_t \neq 0$, $t = 1, 2$, and $b \neq 0$, obtaining a contradiction. Suppose now that Assumption 5 holds and that 6(ii) does not. Given the latter, there exists a non-zero function ξ such that $\xi \in \cap_{t=1}^{2} \mathcal{R}(K_{y_t}) \cap \mathcal{R}(K_x)$. Then there exist non-zero functions $\delta_{1\xi} \in L_{Y_1}^2, \delta_{2\xi} \in L_{Y_2}^2, b_\xi \in \mathbb{R}^k$ such that

$$K_{y_2}\delta_{2\xi} = \xi, \; K_{y_1}\delta_{1\xi} = \xi, \; K_x b_\xi = \xi.$$

In addition, since each operator above is linear, it holds that:

$$K_{y_1}(2\delta_{1\xi}) = 2\xi, \; K_x(-b_\xi) = -\xi.$$

Then

$$K(2\delta_{1\xi}, \delta_{2\xi}, -b_\xi) = \xi - 2\xi + \xi = 0,$$

but since K is injective, we obtain a contradiction, e.g., from K being injective $\delta_{2\xi} = 0$, but this is a contradiction with our assumption that $\delta_{2\xi}$ is a non-zero function.

Second, we show by contradiction that Assumptions 6(i) and 6(ii) imply Assumption 5. Suppose that Assumptions 6(i) and 6(ii) hold and that Assumption 5 does not. Let δ_1, δ_2, b be such that

$$K_{y_1}\delta_1 = 0, \; K_{y_2}\delta_2 = 0, \text{ and } K_x b = 0$$

so that, by injectivity of the operators, $\delta_1 = \delta_2 = b = 0$ a.s. Then

$$K(\delta_1, \delta_2, b) = K_{y_2}\delta_2 - K_{y_1}\delta_1 - K_x b = 0 \text{ a.s.}$$

and $(\delta_1, \delta_2, b) \neq (0,0,0)$ a.s. since K is not injective. This leads to a contradiction.

Since $h_t^{-1}, t = 1, 2$ have been identified, the pre-images of $h_t, t = 1, 2$, and, hence $h_t, t = 1, 2$, are identified.

B.2 Proof of Theorem 3

Proof. The proof follows from Florens and Sokullu (2017). Here we provide a sketch. First, note that

$$\hat{H}^{\gamma_n} - H = A + B + C,$$

where

$$A \equiv (\gamma_n I + \hat{K}_y^*(I - \hat{P}_x)\hat{K}_y)^{-1}\hat{K}_y^*(I - \hat{P}_x)\hat{r} - (\gamma_n I + \hat{K}_y^*(I - \hat{P}_x)\hat{K}_y)^{-1}\hat{K}_y^*(I - \hat{P}_x)\hat{K}_y H, \tag{39}$$

$$B \equiv (\gamma_n I + \hat{K}_y^*(I - \hat{P}_x)\hat{K}_y)^{-1}\hat{K}_y^*(I - \hat{P}_x)\hat{K}_y H - (\gamma_n I - K_y^*(I - P_x)K_y)^{-1}K_y^*(I - P_x)K_y H, \tag{40}$$

$$C \equiv (\gamma_n I - K_y^*(I - P_x)K_y)^{-1}K_y^*(I - P_x)K_y H - H, \tag{41}$$

where A captures the estimation error on the right hand side of the equation, B shows the error coming from estimation of the operators, and C captures the regularisation bias. Following Florens and Sokullu (2017), A can be shown to be $O_p\left(\frac{1}{\gamma_n^2}\left(\frac{1}{n} + b_n^{2s}\right)\right)$, while B and C are $O_p\left(\frac{1}{\gamma_n^2}\left(\frac{1}{nb_n^{q+1}} + b_n^{2s}\right)\gamma_n^{\nu}\right)$ and $O_p(\gamma_n^{\nu})$, respectively.

Second, $\sqrt{n}(\hat{\beta}-\beta)$ can be decomposed as:

$$\begin{aligned}\sqrt{n}(\hat{\beta}-\beta) &= \hat{M}_\gamma^{-1}\Bigg\{\underbrace{\sqrt{n}[K_x^*(I-P_y)\hat{E}(U_2-U_1|Z)]}_{I} \\ &\quad -\underbrace{\sqrt{n}[K_x^*(I-P_y)-\hat{K}_x^*(I-\hat{P}_y^\gamma)]\hat{E}(U_2-U_1|Z)}_{II} \\ &\quad +\underbrace{\sqrt{n}[\hat{K}_x^*(I-\hat{P}_y^\gamma)\hat{K}_y(h_1^{-1},h_2^{-1})]}_{III}\Bigg\}\end{aligned}$$

where $\hat{P}_y^\gamma = \hat{K}_y(\gamma I+\hat{K}_y^*\hat{K}_y)^{-1}\hat{K}_y^*$. The proof then proceeds showing the following which lead to the final result:

$$\|\hat{M}_\gamma^{-1}-M^{-1}\| \to o_p(1),$$

where

$$M = K_x^*K_y(K_y^*K_y)^{-1}K_y^*K_x - K_x^*K_x,$$

and

$$\|II\| \to O_p(1),$$
$$\|III\| \to O_p(1),$$
$$\hat{M}_\gamma^{-1}\left\{\sqrt{n}[K_x^*(I-P_y)\hat{E}(U_2-U_1|Z)]\right\} \to \mathcal{N}\left(0,\sigma^2M^{-1}\left(\sum_{j/\psi_j\in\mathcal{R}(K_y)^\perp} E(\Delta X\psi_j)E(\Delta X\psi_j)')M^{-1}\right)\right).$$

□

B.3 Proof of Corrolary 1

Following Darolles et al. (2011), we first show that rate of convergence of $\hat{H}^{\gamma_n}$ can be shown to be equal to $n^{-\frac{\nu}{2+\nu}}$. And then we show that $\nu = 2/3$, this rate is equal to

$n^{-1/4}$. Consider the convergence rate of $\hat{H}^{\gamma_n}$ given in Theorem 1. The proof based on making the middle term negligible. Assume that $b_n^{2s} \sim \frac{1}{n}$, together with assumption $nb_n^{q+1} \to \infty$, this implies that $s \geq \frac{q+1}{2}$ and then the middle term is $O_p\left(\frac{\gamma_n^{\nu-2}}{nb_n^{q+1}}\right)$.

If the middle term is negligible, together with $b_n^{2s} \sim 1/n$, optimal γ_n is obtained by setting equal the first and the third term:

$$\frac{1}{\gamma_n^2 n} \sim \gamma_n^{\nu},$$

which will lead to $\gamma_n \sim n^{-\frac{1}{2+\nu}}$. Going back to the middle term, one can then choose a bandwidth which satisfies:

$$\frac{1}{nb_n^{q+1}} = O_p\left(\frac{\gamma_n^{\nu}}{\gamma_n^{\nu} - 2}\right)$$

If we replace the γ_n with its optimal rate in the above equation, we obtain the first condition of the corollary. Then under $\gamma_n \sim n^{-\frac{1}{2+\nu}}$ and if $s \geq \frac{(q+1)(\nu+2)}{2\nu}$, the rate of convergence of $\hat{H}^{\gamma_n}$ is given by:

$$\|\hat{H}^{\gamma_n} - H\|^2 = O_p(n^{-\nu/\nu+2}),$$

which is equal to $O_p(n^{-1/4})$ for $\nu = 2/3$.

B.4 Proof of Theorem 4

The proof consists in verifying the conditions in Theorem 2 in Chen et al. (2003). Conditions (2.1), (2.2), (2.4) are standard and hold. Condition (2.5) holds since the conditions of Lemma 1 in Chen et al. (2003) hold, which is sufficient for Condition (2.5), see Remark 2 in Chen et al. (2003). In particular, the class of functions $\{m(\delta_{k,t}, h_t, \beta_k) : h_t \in L^2(\mathbb{R}), \beta_k \in \mathbb{R}, \delta_{k,t} \in \mathbb{R}\}$ is $P-$Donsker, where P is the probability measure of Y_t given that h_t is strictly increasing and Lipshitz continuous, and given Donsker preservation results in van der Vaart and Wellner (1996). Conditions (2.3) and (2.6) are directly assumed at the time of writing.

C Extension

It is possible to analyze the more general model below. For any $z_t \in \mathcal{Z}_t$,

$$Y_{it} = h_t\left(\rho\left(X_{it}\right) + \alpha_i + U_{it}\right), \ E\left(U_{it} - U_{it-1} \middle| Z_{it} = z_t\right) = 0. \tag{42}$$

This model nests that of Féve and Florens (2014) when $h_t(s) = s$.

Assuming that the instrumental variable is time-invariant obtains for $t = 2$:

$$E\left(h_2^{-1}\left(Y_2\right) - h_1^{-1}\left(Y_1\right) + \rho\left(X_2\right) - \rho\left(X_1\right) \middle| Z = z\right) = 0. \tag{43}$$

Via an observational equivalence argument as above with (g_1, g_2, R) that are observationally equivalent to (h_1, h_2, ρ) and, in particular, that satisfy

$$E\left(g_2^{-1}\left(Y_2\right) - g_1^{-1}\left(Y_1\right) + R\left(X_2\right) - R\left(X_1\right) \middle| Z = z\right) = 0, \tag{44}$$

subtracting (44) from (43) obtains

$$E\left(\tilde{\delta}_2\left(Y_2\right) - \tilde{\delta}_1\left(Y_1\right) + r\left(X_2\right) - r\left(X_1\right) \middle| Z = z\right) = 0, \tag{45}$$

where

$$\tilde{\delta}_t\left(Y_t\right) \equiv h_t^{-1}\left(Y_t\right) - g_t^{-1}\left(Y_t\right), \ t = 1, 2, \tag{46}$$

and

$$r\left(X_t\right) = \rho\left(X_t\right) - R\left(X_t\right), \ t = 1, 2. \tag{47}$$

As before, the identification argument involves completeness and measurable separability assumptions.

Assumption 16. *(i)* $E\left(h_t^{-1}\left(Y_t\right) \middle| Z\right) \in L_Z^2$, $E\left(\rho\left(X_t\right) \middle| Z\right) \in L_Z^2$, $t = 1, 2$; *(ii) The random variables* (Y_1, Y_2, X_1, X_2) *are strongly identified by* Z, *i.e. for* $\tilde{\delta}_t \in L_2\left(Y_t\right), r \in L_2\left(X_t\right), t = 1, 2,$ *defined in (46) and (47), respectively, if*

$$E\left(\tilde{\delta}_2\left(Y_2\right) - \tilde{\delta}_1\left(Y_1\right) + r\left(X_2\right) - r\left(X_1\right) \middle| Z\right) = 0 \ a.s. \ F_z$$

then

$$\tilde{\delta}_2(Y_2) - \tilde{\delta}_1(Y_1) + r(X_2) - r(X_1) = 0 \ \ a.s. \ \ F_{Y_1,Y_2,X_1,X_2};$$

(iii) The random variables Y_1, Y_2, X_1, X_2 *are measurably separable in the sense that for* $\tilde{\delta}_t(Y_t), \ r(X_t), t = 1, 2,$ *defined in (46) and (47), respectively, if*

$$\tilde{\delta}_2(Y_2) - \tilde{\delta}_1(Y_1) + r(X_2) - r(X_1) = 0 \ \ a.s. \ \ F_{Y_1,Y_2,X_1,X_2},$$

then there exist constants $\tilde{c}_t, d_t \in \mathbb{R}$, $t = 1, 2$, *such that*

$$\tilde{\delta}_t(Y_t) = \tilde{c}_t \ \ a.s. \ \ F_{Y_t},$$
$$r(X_t) = d_t \ \ a.s. \ \ F_{X_t}.$$

(iv) $E\left(h_t^{-1}(Y_t)\right) = 0$, $E(\rho(X_t)) = 0$, $t = 1, 2$.

Theorem 5. *Let Assumptions 1 and 16 hold, and let* Y_t, X_t, Z_t *satisfy (42). Then* h_t *and* ρ *are identified.*

Proof. Consider (45). By Assumption 16(ii), it follows that:

$$\tilde{\delta}_2(Y_2) - \tilde{\delta}_1(Y_1) + r(X_2) - r(X_1) = 0 \text{ a.s.}$$

By Assumption 16(iii) there exist constants $\tilde{c}_t$, $d_t \in \mathbb{R}$ such that $\tilde{\delta}_t(Y_t) = \tilde{c}_t$ a.s., $r(X_t) = d_t$ a.s. By Assumption 16(iv) these constants are all equal to zero. □