Nonparametric Estimation of Dynamic Panel Models

Yoonseok Lee*

Department of Economics

Yale University

This version: December 15, 2005

[JOB MARKET PAPER II]

*I am especially grateful to my advisor, Peter Phillips, for his invaluable guidance and encouragement. I would like to thank Donald Andrews and Yuichi Kitamura for their insightful comments and supervision. I also obtained valuable comments from Taisuke Otsu, Rosa Matzkin, Paramaraj Jegenathan; seminar participants at Yale University and at Greater New York Metropolitan Area Econometrics Colloquium, Columbia University. I gratefully acknowledge financial support from the Cowles Foundation under the Carl Arvid Anderson Prize. All errors are solely mine. *E-mail:* `y.lee@yale.edu`. *Address:* Department of Economics, Yale University, P.O.Box 208268, New Haven, CT 06520-8268, USA.

Abstract

This paper develops nonparametric estimation of dynamic panel models using series approximations. We extend the standard linear dynamic panel model to a nonparametric form that maintains additive fixed effects, where the fixed effects are eliminated by the within transformation. This approach generalizes earlier work on cross sectional series estimation by Newey (1997). Nonlinear homogeneous Markov process is properly conditioned to be stationary β-mixing. Convergence rates and the asymptotic distribution of the series estimator are derived when both the cross section sample size and the length of the time series are large and of comparable sizes. Just as for pooled estimation in linear dynamic panels, an asymptotic bias is present, which reduces the mean square convergence rate compared with the cross section case. To tackle this problem, bias correction is developed using a heteroskedasticity and autocorrelation consistent (HAC) type estimator. Some extensions of this framework are also considered under exogenous variables and partial linear models. The limit theory and bias correction formulae follow by extending the main results. Finally, an empirical study on nonlinearities in cross-country growth regressions is presented to illustrate the use of the nonparametric dynamic panel models with fixed effects. After bias correction, the convergence hypothesis is true only for countries in the upper income range and for OECD countries.

Key words and phrases: Nonparametric estimation, series estimation, dynamic panel, fixed effects, within transformation, convergence rates, asymptotic normality, bias correction, partial linear model, β-mixing, growth convergence.

JEL classifications: C14, C23, O40

1 Introduction

In spite of the large and growing literature on nonparametric modelling in econometrics, little attention has been given to nonparametric estimation in dynamic panels. One explanation is the difficulty of treating individual effects and the autoregressive structure simultaneously in the context of nonparametric estimation, especially when the unobserved individual effects are specified as fixed effects. This paper seeks to overcome this problem by developing series approximations for nonlinear dynamics in a panel system and by extending the standard linear dynamic panel model to a nonparametric form that maintains additive fixed effects.

There are several studies on nonparametric or semiparametric models for panel systems. For pure nonparametric models, Porter (1996) derives a limit distribution of the nonparametric estimator in static (i.e., non-dynamic) independent panel models with fixed effects, when the cross section sample size, N, is large but the length of the time series, T, is fixed. Both series estimations and kernel techniques are explored. Under similar conditions, Ullah and Roy (1998) consider kernel estimation for panels when both N and T are large. In a recent study by Mundra (2005), the local polynomial estimation technique is used to estimate the slope of the unknown function. Instead of considering fixed effects, Henderson and Ullah (2005) look at nonparametric estimation of random effect models. All of these studies look at static panel systems and show that the conventional nonparametric analysis can be extended to panel models. For semiparametric models, Baltagi and Li (2002) extend the partial linear model of Robinson (1988) to panel systems with fixed effects, but they consider static and independent panels. Li and Stengos (1996), and Li and Kniesner (2002) investigate partial linear models in the context of dynamic panel models, but only consider random effects. In a similar vein, Hahn and Kuersteiner (2004) examine nonlinear (parametric) dynamic panel models with fixed effects.

Though several studies have analyzed nonparametric or semiparametric panel models with individual effects, there appear to be no theoretical studies tackling both dynamics and fixed effects at the same time in the context of nonparametric panel estimation. Therefore, the main contribution of this study is that it develops nonparametric estimation techniques suitable for dynamic panel models with fixed effects, in which the fixed effects are eliminated by the within transformation (i.e., deviations from the individual sample average over time). Moreover, the limit properties of the within-transformation-based nonparametric estimator are explored under large N and T asymptotics when N and T are of comparable sizes. Such asymptotic results are expected to be of practical relevance when T is not too small compared to N as is in the cases of cross-country studies (e.g., the Penn World Tables) and cross-firm financial studies.

This paper mainly looks at the within transformation instead of the first-differencing transformation. First-differenced dynamic panel models are, unlike static panel models, estimated using instrumental variables (IV) because the first-differencing transformation provokes nonzero correlation between the error and regressors. For the cross section case, nonparametric IV estimation is examined in several recent studies. See Blundell and Powell (2003), Newey and Powell (2003), Darolles, Florens and Renault (2003), Ai and Chen

(2003), and Hall and Horowitz (2003) among others. Though these studies are mainly for cross section data, the extension to dynamic panels can be done as long as T is small and fixed. We will discuss some ideas and limitations in this paper. Meanwhile, there seems to be no attempt to develop nonparametric estimation for the within-transformed model in the context of dynamic panels.

Taking it into account, this paper develops nonparametric estimation for the within-transformed dynamic panel models using series approximation. Series estimation is convenient in this context because it approximates an unknown function with a linear combination of known functions; therefore, the within transformation of the unknown function is simply the same linear combination of the within-transformed series functions. Moreover, as in the conventional within-group (WG; or the least squares dummy variable, LSDV) estimation, the new estimation procedure is based on least squares estimation, and thus it is much easier to implement in practice than IV-based estimation.

Specifically, this approach follows earlier works on cross sectional series estimation by Andrews (1991) and Newey (1997), and generalizes their asymptotic results to dynamic panels using nonlinear time series techniques. Under proper conditions, a panel homogeneous Markov process is shown to satisfy stationary β-mixing condition, which will be the basic building block to control temporal dependence. We derive the mean square convergence rate and the asymptotic distribution of the series estimator when both N and T are large. Just as for pooled estimation in linear dynamic panels (e.g., Hahn and Kuersteiner, 2002; Alvarez and Arellano, 2003; Lee, 2005), an asymptotic bias is present, which reduces the mean square convergence rate compared with the cross section case. To tackle this problem, bias correction is developed using a heteroskedasticity and autocorrelation consistent (HAC) type estimator.

Some extensions of this framework are also considered under exogenous variables and partial linear models, which are more relevant in applications. The limit theory and bias correction for these cases follow by extending the main results. Finally, an empirical study on nonlinearities in cross-country growth regressions is presented to illustrate the use of the nonparametric estimation techniques for dynamic panels with fixed effects. Including fixed effects in the growth regression allows heterogeneous production functions across countries. In addition, several recent studies question the assumption of linearity in growth equations and propose nonlinear alternatives that allow for multiple regimes of growth patterns among different countries. The findings of this paper support this idea, similarly as in Liu and Stengos (1999), where we analyze a semiparametric dynamic panel growth equation with fixed effects using the Penn World Table.

This paper is organized as follows. Section 2 introduces the basic model and discusses the stability condition for the nonlinear autoregressive panel systems. In Section 3, WG series estimation is developed and its limit properties are examined under large N and T asymptotics. A pointwise bias correction method is also discussed. In Section 4, the main results are generalized to include exogenous variables and partial linear models. Some ideas on two stage nonparametric IV estimation, based on the first-differenced model, are also briefly discussed. In Section 5, Monte Carlo experiments are conducted to examine the performance of the WG series estimator and bias correction. In Section 6, an empirical study on nonlinear cross-country growth regressions is presented. Section 7 concludes the study with some remarks. All the mathematical proofs are provided in the Appendix.

A word on notation: for a vector x and a matrix B, we let $\|x\| = (x'x)^{1/2}$ and $\|B\| = (tr(B'B))^{1/2}$, where $tr(\cdot)$ is the trace operator.

2 Model

We consider a panel process $\{y_{i,t}\}$ generated from a nonlinear autoregressive model given by[1]

$$y_{i,t} = m(y_{i,t-1}) + \mu_i + u_{i,t} \tag{1}$$

for $i = 1, 2, \cdots, N$ and $t = 1, 2, \cdots, T$, where $m : \mathbb{R} \to \mathbb{R}$ is an unknown Borel measurable function. We assume that the realization of the initial values, $y_{i,0}$, are observed for all i. The random variable μ_i is an individual effect, assumed to be independent of $u_{i,t}$ for all i and t, but possibly correlated with $y_{i,t-1}$. In other words, μ_i is assumed to be a fixed effect. Unlike a random effect, the fixed effect captures the omitted−and thus unobserved−cross sectional heterogeneity;[2] therefore, they are allowed to be correlated with the explanatory variables, $y_{i,t-1}$. On the other hand, $u_{i,t}$ is independent of $\{y_{i,s}\}_{s\leq t-1}$ so that $\mathbb{E}(u_{i,t}|y_{i,t-1}, \cdots, y_{i,0}) = 0$. This specification assumes a common shape for the conditional mean function $m(\cdot)$ for all i but different level shifts according to μ_i. We can consider a more general specification given by

$$y_{i,t} = m(y_{i,t-1}, \cdots, y_{i,t-p}; x_{i,t}, \cdots, x_{i,t-q+1}) + \mu_i + u_{i,t},$$

which allows for higher order lag terms of $y_{i,t}$ and lags of exogenous variables $x_{i,t} \in \mathbb{R}^r$ in the unknown function m. Including exogenous variables in the regression is relevant in empirical studies. We will discuss such an extension in Section 4. The main analysis of this paper, however, focuses on the simple nonlinear model given in (1).

The independence assumption between μ_i and $u_{i,t}$ is important to avoid an endogeneity problem. The data generating process in (1) implies that $y_{i,t}$ is a function of μ_i and $\{u_{i,s}\}_{s\leq t}$. Therefore, the conditional expectation, $\mathbb{E}(u_{i,t}|y_{i,t-1}, \cdots, y_{i,0}) = \mathbb{E}(u_{i,t}|\mu_i, u_{i,t-1}, u_{i,t-2}, \cdots)$, is zero only when $u_{i,t}$ is independent of μ_i, and $u_{i,t}$ is uncorrelated with its own lags for each i. For example, $\{u_{i,t}\}$ needs to be a martingale difference sequence on the filtration $\{\mathcal{F}_{i,t}\}$, where $\mathcal{F}_{i,t} = \sigma(u_{i,s} : s \leq t)$. These conditions imply that $u_{i,t}$ is independent of $\{y_{i,s}\}_{s\leq t-1}$ for each i. The independence between μ_i and $u_{i,t}$, on the other hand, does

[1] One could consider a model $y_{i,t} = m_\mu(y_{i,t-1}; \mu_i) + u_{i,t}$, but μ_i and m_μ are not seperately identifiable without further restrictions on m_μ.

[2] It is an important issue especially when the equation includes exogenous variables.

not implies $\mathbb{E}(\mu_i|y_{i,t-1},\cdots,y_{i,0}) = 0$ since $\{y_{i,s}\}_{s\leq t-1}$ are still functions of μ_i. The potential correlation between the individual effects and the regressors thus remains. To make the notation as simple as possible, however, we simply let $\{u_{i,t}\}$ be an independent and identically distributed (*i.i.d.*) process with mean zero. Therefore, across i, $\{y_{i,t}\}$ is also independent with heterogeneous means. The generalization to serially (weak) dependent $u_{i,t}$ such as a martingale difference sequence can be easily done, but at the cost of notational complexity. On the other hand, the generalization to cross sectional dependence as in Phillips and Sul (2004) is not straightforward, and we do not pursue it in this paper.

The stability of the linear autoregressive process is determined by the support of the roots of the polynomial characteristic function. In the nonlinear case, however, such techniques are infeasible, and proper conditions are required to satisfy ergodicity and mixing property. To derive such conditions, we suppose $\{y_{i,t}\}$ is a Markov process given in (1), with homogeneous transition probability F_i and initial distribution as its invariant measure π_i for each i. Then the process $\{y_{i,t}\}$ is stationary over t and its marginal distribution is given by π_i. We define the β-mixing coefficient $\beta_i(\tau)$ as (e.g., Davydov, 1973; Doukhan, 1994)

$$\beta_i(\tau) = \sup_t \mathbb{E}\left[\sup_{A\in\mathcal{G}_{i,t+\tau}^{\infty}} \left\|\mathbb{P}\left(A|\mathcal{G}_{i,-\infty}^{t}\right) - \mathbb{P}(A)\right\|_{TV}\right]$$

for $\tau > 0$, where $\mathcal{G}_{i,t_1}^{t_2}$ is the σ-field generated by $\{y_{i,t} : t_1 \leq t \leq t_2\}$. $\|\cdot\|_{TV}$ is the total variation[3] of a signed measure. If $\beta_i(\tau) \to 0$ as $\tau \to \infty$, then $\{y_{i,t}\}$ is β-mixing for a fixed i. Davydov (1973) gives the following equivalent definition of $\beta_i(\tau)$ for a homogeneous stationary Markov chain $\{y_{i,t}\}$:

$$\beta_i(\tau) = \int \pi_i(dy) \|F_i^{\tau}(y,\cdot) - \pi_i(\cdot)\|_{TV},$$

where $F_i^{\tau}(y,\cdot)$ is the τ-th step transition probability. We define $\beta(\tau) = \sup_{1\leq i\leq N} |\beta_i(\tau)|$ for all $\tau > 0$; and we will say $\{y_{i,t}\}$ is *β-mixing* (i.e., absolutely regular) if $\beta(\tau) \to 0$ as $\tau \to \infty$.

In the nonlinear time series literature, it is well established that a homogeneous Markov chain is β-mixing with mixing coefficients tending to zero at an exponential rate if it is geometrically (Harris) ergodic. (e.g., Doukhan, 1994) Moreover, it is straightforward that geometric ergodicity implies stationarity of the process $\{y_{i,t}\}$ if the distribution of the initial values $y_{i,0}$ are defined by the stationary probability measure π_i. When individual effects are present in the dynamics as in (1), however, $\{y_{i,t}\}$ cannot be (geometrically) ergodic because a common random constant μ_i will affect the temporal dependence structure. But when the whole process is conditional on μ_i, the random constant[4] becomes simply a common constant shift in the distribution of the process; therefore, μ_i no longer affects the statistical temporal dependence of $\{y_{i,t}\}$. In

[3] We denote the total variation norm of the signed measure σ on a σ-field $\mathcal{B}$ by $\|\sigma\|_{TV}$ such that $\|\sigma\|_{TV} \doteq \sup_{B\in\mathcal{B}}\sigma(B) - \inf_{B\in\mathcal{B}}\sigma(B)$. If σ_1 and σ_2 are two probability measures and $\sigma = \sigma_1 - \sigma_2$, then we have $\|\sigma\|_{TV} = 2\sup_{B\in\mathcal{B}}|\sigma_1(B) - \sigma_2(B)|$ in viwe of Scheffe's theorem. (cf. Liebscher, 2005, p.671)

[4] Considering μ_i as random is essential to allow correlation between μ_i and $y_{i,t-1}$. Otherwise, there remains no correlation

what follows, even though we do not explicitly indicate "conditional on μ_i," all the arguments presume it. The following two assumptions summarize the conditions for the homogeneous Markov process $\{y_{i,t}\}$ to be geometrically ergodic.

Assumption E1 *(i) $\{u_{i,t}\}$ is i.i.d. with mean zero, variance σ^2 and $\mathbb{E}|u_{i,t}|^{\nu} < \infty$ for some $\nu > 4$. (ii) $\{u_{i,t}\}$ has a density, which is positive almost everywhere, and a marginal distribution, which is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}$. (iii) $u_{i,t}$ is independent of μ_i for all i and t.*

As discussed, the condition E1 implies that $u_{i,t}$ is independent of $\{y_{i,s}\}_{s \leq t-1}$, which is natural when the process $\{y_{i,t}\}$ is generated from a model in the natural time order. We need to assume that the distribution of $u_{i,t}$ has positive density almost everywhere in order to reduce restriction on the nonlinear function $m(\cdot)$. In a linear autoregressive model, however, this assumption is not necessary. The next condition controls the nonlinear function $m(\cdot)$ to ensure the stability of the process $\{y_{i,t}\}$. We let $\phi_i(y) = \mu_i + m(y)$ for $y \in \mathcal{Y} \subseteq \mathbb{R}$ and for each i.

Assumption E2 *(i) For each i and for the Borel measurable functions $\phi_i : \mathbb{R} \to \mathbb{R}$, there exist positive constants $\overline{y}$, $c_1 < 1$ and c_{i0} satisfying $|\phi_i(y)| \leq c_1|y| + c_{i0}$ if $|y| > \overline{y}$; and $\sup_{y:|y|\leq\overline{y}} |\phi_i(y)| < \infty$. (ii) For each i, the Markov process $\{y_{i,t}\}$ has a homogeneous transition probability F_i, and the initial value $y_{i,0}$ is drawn from the invariant distribution π_i.*

The assumption E2-(i) implies that for large $|y|$ the behavior of the function ϕ_i is dominated by a stable linear function. A wide class of nonlinear autoregressive functions, such as (bounded) autoregressive processes, semi-parametric autoregressive processes and threshold autoregressive processes, satisfy this assumption. For more examples and discussions, readers may refer to Tong (1990), Doukhan (1994), An and Huang (1996) and references therein. The condition E2-(ii) is necessary for stationarity. The following propositions establish that the homogeneous Markov process $\{y_{i,t}\}$ is geometrically ergodic and thus β-mixing with mixing coefficients $\beta(\tau)$ tending to zero as $\tau \to \infty$ at an exponential rate. Since $\{y_{i,t}\}$ is simply an autoregressive time series for each i and conditional on μ_i, the proofs of Proposition 2.1 and 2.2 directly follow from Doukhan (1994), An and Huang (1996), or Liebscher (2005).

Proposition 2.1 *Suppose that the process $\{y_{i,t}\}$ is generated by (1). Then, for each i, the process $\{y_{i,t}\}$ is geometrically ergodic conditioning on μ_i, provided that Assumptions E1 and E2 hold.*

Proposition 2.2 *For each i and conditioning on μ_i, the homogeneous Markov process $\{y_{i,t}\}$ is stationary and geometrically ergodic if and only if $\{y_{i,t}\}$ is stationary β-mixing with exponential decay.*

---

between μ_i and $y_{i,t-1}$, and μ_i is no longer a *fixed effect* in the sense of Wooldridge (2002, Chapter 10).

Note that β-mixing implies α-mixing (i.e., strong mixing; Doukhan, 1994). Therefore, Assumptions E1 and E2 imply that $\{y_{i,t}\}$ is α-mixing and we can use well-established results for α-mixing processes. The strong mixing condition has been frequently employed in the nonparametric time series literature, following Robinson (1983). In fact, we only require a mixing condition in order to control the temporal dependence in the proof, and using more general mixing condition (i.e., α-mixing condition) does not alter any implication of the study. To make this section complete, we define α-mixing coefficients of $\{y_{i,t}\}$ as

$$\alpha_i(\tau) = \sup_t \sup_{A \in \mathcal{G}_{i,-\infty}^t, B \in \mathcal{G}_{i,t+\tau}^\infty} |\mathbb{P}(A \cap B) - \mathbb{P}(A)\mathbb{P}(B)| \tag{2}$$

for $\tau > 0$ and for each i. We let $\alpha(\tau) = \sup_{1 \leq i \leq N} |\alpha_i(\tau)|$. Since $\alpha(\tau) \leq \beta(\tau)$ for each τ, $\alpha(\tau)$ also tends to zero at an exponential rate. That is, $\alpha(\tau) \leq C_\alpha a^\tau$ for some a such that $0 < a < 1$ and some constant $0 < C_\alpha < \infty$. The following proposition gives that an α-mixing process is invariant under arbitrary Borel measurable transformations. The details can be found in White and Domowitz (1984).

Proposition 2.3 *Let $\psi : \mathbb{R}^{k_1} \to \mathbb{R}^{k_2}$ be measurable with finite k_1 and k_2. If $\{z_t\}$ is α-mixing such that its mixing coefficient is $O(\tau^{-\epsilon})$ for some $\epsilon > 0$, then $\{\psi(z_t, \cdots, z_{t-k})\}$ is also α-mixing such that its mixing coefficient is $O(\tau^{-\epsilon})$.*

The following mixing inequalities will be used frequently. The proof can be found in, for example, Billingsley (1968), Bierens (1994), or Fan and Yao (2003).

Proposition 2.4 *Let $\alpha = \sup_{A \in \sigma(X), B \in \sigma(Y)} |\mathbb{P}(A \cap B) - \mathbb{P}(A)\mathbb{P}(B)|$, then*

(1) *$|cov(X,Y)| \leq 4\alpha C_1 C_2$ if $\mathbb{P}(|X| < C_1) = 1$ and $\mathbb{P}(|Y| < C_2) = 1$ for some C_1 and C_2;*

(2) *$|cov(X,Y)| \leq 8\alpha^{1-1/p-1/q} (\mathbb{E}|X|^p)^{1/p} (\mathbb{E}|Y|^q)^{1/q}$ if $\mathbb{E}|X|^p + \mathbb{E}|Y|^q < \infty$ for some $p, q \geq 1$ and $1/p + 1/q < 1$.*

3 Within-Group Series Estimation

3.1 Within-group estimator

The primary effort in estimating (1) should be on eliminating individual effects, μ_i, in order to avoid incidental parameter problem as N increases. We can employ one of two conventional methods in linear panel models: the within transformation (i.e., deviations from the individual sample average over time) and the first-differencing transformation. Pooled least squares estimation based on the within transformation is known as within-group (WG) estimation or least squares dummy variable (LSDV) estimation. Specifically, the within transformation of (1) yields

$$y_{i,t}^{0}=\left\{m\left(y_{i,t-1}\right)-\frac{1}{T}\sum_{s=1}^{T}m\left(y_{i,s-1}\right)\right\}+u_{i,t}^{0}, \tag{3}$$

and the first-differencing transformation of (1) yields

$$\Delta y_{i,t}=\left\{m\left(y_{i,t-1}\right)-m\left(y_{i,t-2}\right)\right\}+\Delta u_{i,t}, \tag{4}$$

where for any variable $w_{i,t}$ we define $\Delta w_{i,t}=w_{i,t}-w_{i,t-1}$ and $w_{i,t}^{0}=w_{i,t}-(1/T)\sum_{s=1}^{T}w_{i,s}$.

The equations (3) and (4) show that we cannot directly estimate the unknown function $m(\cdot)$ by simple kernel regressions. The main reason is an endogeneity problem incurred by the within or first-differencing transformations. To explain this, we rewrite the within-transformed model (3) as $y_{i,t}^{0}=\ell_{WT}\left(y_{i,t-1},\cdots,y_{i,0}\right)+u_{i,t}^{0}$, where $\ell_{WT}\left(y_{1},\cdots,y_{T}\right)=m\left(y_{1}\right)-(1/T)\sum_{s=1}^{T}m\left(y_{s}\right)$. Then estimating ℓ_{WT} by kernel regression is infeasible since its dimension increases as $T\rightarrow\infty$; moreover, the regression involves an endogeneity problem. We can also rewrite the first-differenced model (4) as $\Delta y_{i,t}=\ell_{FD}\left(y_{i,t-1},y_{i,t-2}\right)+\Delta u_{i,t}$, where $\ell_{FD}\left(y_{1},y_{2}\right)=m\left(y_{1}\right)-m\left(y_{2}\right)$. Though this regression model does not incur the curse of dimensionality as in the within transformation case, it still has an endogeneity problem and therefore, we need nonparametric instrumental variables (IV) estimation for (4).

Recently, a large and growing literature has been devoted to studying endogeneity in nonparametric and semiparametric regression models in the cross section case. Blundell and Powell (2003) provide a good survey of the recent development. For example, there are well established limit theories for a two stage nonparametric IV estimator as in Newey and Powell (2003), Darolles, Florens and Renault (2003), Ai and Chen (2003), and Hall and Horowitz (2003) among others. Though these results are based on the independent cross section case, the extension to the dynamic panel in (4) can be done when T is small and fixed. We will briefly discuss such extension in Section 4. One drawback of this approach is, however, slow rate of convergence; so we take a different approach in this paper, the within-transform-based method.

To the author's knowledge, no study has developed nonparametric estimation for the within-transformed model (3); the existing panel literature only considers static (i.e., non-dynamic) models. For example, see Porter (1996), Ullah and Roy (1998), and Baltagi and Li (2002). One explanation is the difficulty of treating fixed effects and the autoregressive structure simultaneously in the context of nonparametric estimation. We tackle this problem by extending series estimation to dynamic panels, where the temporal dependence of nonlinear dynamics is controlled by the absolute regularity conditions introduced in Section 2. Series estimation is convenient in this context because it approximates an unknown function with a linear combination of known functions; therefore, the within transformation of the unknown function is simply the same linear combination of the within-transformed series functions. As in the conventional WG estimation, the new estimation procedure is based on least squares estimation, and thus it is much easier to implement in practice than IV-based estimation. Moreover, as is well known in linear dynamic panel models, the IV-based estimators are inferior to the WG estimators in their efficiencies in finite samples. As long as the asymptotic bias of the WG estimators are properly corrected, their statistical properties are expected to be superior to those of the IV-based estimators. We expect that the within-transformation-based nonparametric estimator performs better than the IV-based estimators after a proper bias correction. Such expectation is closely related to the slow rate of convergence of the IV-based nonparametric estimators.

Before we develop a new estimation method, we need to consider an additional condition for identifying μ_i and m separately. By applying either the within transformation or the first-differencing transformation, we successfully eliminate fixed effects, μ_i. The elimination, however, gets rid of both fixed effects and a constant term imbedded in the unknown function $m(\cdot)$ together. We thus need more condition to correctly identify the heterogeneous constants μ_i from the homogeneous unknown function $m(\cdot)$. The following normalization condition is sufficient for the identification.

Assumption ID (normalization and identification) $m(0) = 0$.

In Porter (1996), it is instead assumed that[5]

$$\mathbb{E}\mu_i = 0 \tag{5}$$

or $\sum_{i=1}^{N} \mu_i = 0$ if μ_i's are regarded as fixed parameters. The condition (5) allows the level of $m(0)$ unrestricted, but normalizes the sum of individual effects μ_i to zero. $m(0)$ could be nonzero so that $m(\cdot)$ is allowed to contain a constant term. On the other hand, the normalization condition ID allows μ_i to be unrestricted, but requires that $m(\cdot)$ passes through the origin so that it excludes a constant term from $m(\cdot)$. This condition implies that μ_i absorbs all level shifts from both homogeneous and heterogeneous constant terms, and thus it merely shifts a common graph $m(\cdot)$ vertically for each i. Therefore, if $m(0) \neq 0$, we can simply reparametrize $\mu_i + m(y) = (\mu_i + m(0)) + (m(y) - m(0))$ and consider $\mu_i + m(0)$ and $m(y) - m(0)$ as fixed effects and the unknown function, respectively, to restore this condition. The distinction between the condition ID and (5) explains why Porter (1996) is only able to identify $m(\cdot)$ up to a constant addition.[6] For example, when we consider the first-differenced model given in (4), we need to restore $\widehat{m}(y)$ from the estimator $\widehat{\ell}_{FD}(y_1, y_2) = \widehat{m}(y_1) - \widehat{m}(y_2)$. The constant term in $m(y)$ is, however, already eliminated by the

[5] As noted in Porter (1996), the condition (5) is weaker than $\mathbb{E}(\mu_i|\mathcal{F}_{i,t-1}) = 0$ which assumes away any potential correlation between individual effects and regressors. Thus, under $\mathbb{E}(\mu_i|\mathcal{F}_{i,t-1}) = 0$, heterogeneity bias is no longer an issue. This is the situation of random effects models.

[6] Another merit of the condition ID is that it enables us to readily restore $\widehat{m}(y)$ from the estimator $\widehat{\ell}_{FD}(y_1, y_2)$ or $\widehat{\ell}_{WT}(y_1, y_2, \cdots, y_T)$ because $\ell_{FD}(y_1, 0) = m(y_1) - m(0) = m(y_1)$ and $(T/(T-1))\,\ell_{WT}(y_1, 0, \cdots, 0) = m(y_1)$. Porter (1996), on the other hand, needs to use the partial integration method of Newey (1994) to restore the original unknown function (up to a constant addition).

first-differencing transformation and thus we cannot restore it unless it is zero. The same argument can be made for the within-transformed model in (3).

We now develop the nonparametric estimator based on the within-transformed model given in (3). For notational convenience, we define $m^0(y_{i,t-1}) = m(y_{i,t-1}) - (1/T)\sum_{s=1}^{T} m(y_{i,s-1})$ and rewrite the within-transformed model (3) as

$$y_{i,t}^0 = m^0(y_{i,t-1}) + u_{i,t}^0.$$

In what follows, any variables or functions with superscript ^{0} indicate that they are within-transformed. Note, however, that $m^0(y_{i,t-1})$ does not imply that it is a function of $y_{i,t-1}$ only; instead, it is a function of a complete series of $(y_{i,0}, y_{i,1}, \cdots, y_{i,T-1})$ for each i. To estimate m, we use series approximation as in Andrews (1991) and Newey (1997), which approximates an unknown function m by some linear combination of K known series functions $\{q_{Kk}\}$:

$$m(y) \approx \sum_{k=1}^{K} \theta_{Kk} q_{Kk}(y), \tag{6}$$

where $q_{Kk} : \mathbb{R} \to \mathbb{R}$ are measurable and $\theta_{Kk} \in \mathbb{R}$ for all $k = 1, 2, \cdots, K$. "$\approx$" indicates series approximation; namely, it means "is approximately equal for large K." The choice of the sequence must be such that the approximation to m improves as K gets larger, where $K = K(N,T)$ and $K \to \infty$ as $N, T \to \infty$. Using the series approximation in (6), we can rewrite $m^0(y)$ as

$$m^0(y) \approx \sum_{k=1}^{K} \theta_{Kk} q_{Kk}^0(y),$$

where we transform the series functions as $q_{Kk}^0(y_{i,t}) = q_{Kk}(y_{i,t}) - (1/T)\sum_{s=1}^{T} q_{Kk}(y_{i,s})$. The series functions $q_{Kk}(\cdot)$ are chosen to satisfy $\sum_{k=1}^{K} \theta_{Kk} q_{Kk}(0) = 0$ in order to meet the identification condition in Assumption ID.

In examining limit theories, it is convenient to introduce a trimming function, which bounds the regressor $y_{i,t-1}$ at time t and for each i.[7] In the stability condition in Assumption E2, we presume that the unknown function $\phi_i(y) = \mu_i + m(y)$ is uniformly bounded over a compact set $\mathcal{Y}_c = \{y : |y| \leq \overline{y}\}$ for some $\overline{y} > 0$; and it is dominated by stable linear functions outside $\mathcal{Y}_c$. Therefore, the statistical properties outside $\mathcal{Y}_c$ can be controlled by the estimators for linear dynamic panel models, which is already well established in the literature. We thus only consider estimating the unknown function m over a bounded range of the regressor $y_{i,t-1}$ given by $\mathcal{Y}_c$. Note that, however, for each t, we will only restrict the range of the independent variable

[7] Note that, unlike the static panel models as in Porter (1996), assuming the entire support of y to be bounded does not seem appropriate in case of the autoregressive model (1) since it will not only restrict the support of independent variables $y_{i,t-1}$ but also the support of the dependent variable $y_{i,t}$. Since the error $u_{i,t}$ is defined over $\mathbb{R}$, restricting the support of $y_{i,t}$ bounded can be too strong an assumption.

$y_{i,t-1}$ without restricting the support of the dependent variable $y_{i,t}$. Restricting the support of the dependent variable $y_{i,t}$ produces the truncated regression problem, which renders the least squares estimators biased. Specifically, we define a nonrandom trimming function $\lambda : \mathbb{R} \to \{0, 1\}$ as follows.

Definition TR (trimming function) *We define a sequence of trimming functions* $\{\lambda(y_{i,t})\}$ *given by* $\lambda(y_{i,t}) = \mathbf{1}\{y_{i,t} \in \mathcal{Y}_c\}$ *for some compact* $\mathcal{Y}_c \subset \mathcal{Y}$*, where* $\mathbf{1}\{\cdot\}$ *is the binary indicator function.*

Definition TR along with properly chosen series functions, such as power series or splines,[8] guarantees that $\lambda(y) q_{Kk}(y)$ are uniformly bounded over a bounded subset $\mathcal{Y}_c$.[9] Looking at the unknown function over some bounded range is reasonable and innocuous in empirical studies. Finally, we also note that the trimming is only used for defining the estimator, not for defining the data generating process of $\{y_{i,t}\}$ itself. Therefore, if we let $g_{Kk}(y) = \lambda(y) q_{Kk}(y)$ and $g_K^0(y) = \left(g_{K1}^0(y), g_{K2}^0(y), \cdots, g_{KK}^0(y)\right)'$, where $g_{Kk} : \mathcal{Y}_c \to \mathbb{R}$, then $\theta_K = (\theta_{K1}, \theta_{K2}, \cdots, \theta_{KK})'$ can be estimated by

$$\widehat{\theta}_K = \left(\sum_{i=1}^{N}\sum_{t=1}^{T} g_K^0(y_{i,t-1}) g_K^0(y_{i,t-1})'\right)^{-} \left(\sum_{i=1}^{N}\sum_{t=1}^{T} g_K^0(y_{i,t-1}) y_{i,t}^0\right), \tag{7}$$

where $(\cdot)^{-}$ denotes the generalized inverse. Under conditions given below, however, the denominator will be nonsingular with probability approaching one, and hence the generalized inverse will be the standard inverse. The WG series estimator of $m(\cdot)$ is thus given as

$$\widehat{m}(y) = \sum_{k=1}^{K} \widehat{\theta}_{Kk} g_{Kk}(y) \tag{8}$$

for $y \in \mathcal{Y}_c$. In what follows, we only consider the trimmed series functions $\{g_{Kk}\}$ and estimate the unknown function m over some bounded support $\mathcal{Y}_c$.

[8] A power series approximation corresponds to $q_{Kk}(y) = y^k$ for $k = 0, \cdots, K-1$, where it is conventionally orthogonalized using the Gram-Schmidt orthonormalization. The estimator will be numerically invariant to such transformation, but it may alleviate the multicollinearity problem for power series. An r-th degree spline with L knots $\left(\underline{y}_1, \cdots, \underline{y}_L\right)$ over the known (and empirically bounded) support of y is a linear combination of

$$q_{Kk}(y) = \begin{cases} y^k & \text{for } 0 \le k \le r; \\ \left(y - \underline{y}_{k-r}\right)_+ & \text{for } r+1 \le k \le r+L, \end{cases}$$

where $K = 1 + r + L$; $(z)_+ = z$ if $z > 0$ and zero otherwise. For example, $r = 3$ for cubic splines.

[9] Alternatively, Newey and Powell (2003) approach this problem by specifying $m(y) = m_1(y)' b + m_2(y)$, where $m_1(y)$ are vectors of known functions and b are unknown parameters. b is bounded and $m_2(y)$ and its derivatives are small in the tails. Thus the unknown function is allowed to be nonparametric over the middle of the distribution but is restricted to be almost parametric in the tail. This specification allows for unbounded y.

3.2 Regularity conditions

In this subsection, we list and discuss some regularity conditions on which we base all the following results. Note that we only consider the case where K is not data dependent, but we let it increase as the number of individual observations, N, and the length of time, T, increases, where N and T satisfy the following condition.

Assumption NT $\lim_{N,T\to\infty} N/T = \kappa$, *where* $0 < \kappa < \infty$.

The properties of dynamic panel models are usually discussed under the implicit assumption that T is small and N is large, and they are relying on fixed T and large N asymptotics. Such asymptotics seem quite natural when T is indeed very small compared to N such as the Panel Study of Income Dynamics (PSID) and the National Longitudinal Surveys (NLS). On the other hand, the alternative asymptotic approximation based on large N and T satisfying Assumption NT is expected to be of practical relevance if T is not too small compared to N as is the case, for example, in cross-country studies (e.g., the Penn World Tables) and cross-firm financial studies.

The following assumption is useful for controlling the inverse matrix of $(1/NT)\sum_{i=1}^{N}\sum_{t=1}^{T}\underline{g}_K(y_{i,t})\,\underline{g}_K(y_{i,t})'$ and its convergence in probability in the Euclidean norm as in Newey (1997), where $\underline{g}_K(y)$ denotes the demeaned process of $g_K(y)$ such that $\mathbb{E}\underline{g}_K(y) = 0$.

Assumption W1 *(i) For every K, the $K \times K$ matrix $\Gamma_K = \mathbb{E}\underline{g}_K(y)\,\underline{g}'_K(y)$ has the smallest eigenvalue bounded away from zero and the largest eigenvalue bounded above. (ii) For every K, there is a sequence of constants $\zeta_0(K)$ satisfying $\zeta_0(K) \geq \sup_{y\in\mathcal{Y}_c}\max_{1\leq k\leq K}|g_{Kk}(y)|$ and $K = K(N,T)$ such that $\zeta_0^4(K)\,K^2/NT \to 0$ as $N,T\to\infty$, where $\mathcal{Y}_c \subset \mathcal{Y} \subset \mathbb{R}$ is some bounded subset of the support of $\{y_{i,t}\}$.*

The condition W1-(ii) seems stronger than in Newey (1997), who assumes $\zeta_{0*}^2(K)\,K/NT \to 0$, where $\zeta_{0*}(K)$ is the uniform bound of the $K \times 1$ vector $g_K(y)$. Since we assume that N and T are of the same order of magnitude, however, $\zeta_0^4(K)\,K^2/NT$ is of the same order of magnitude as $\left(\zeta_0^2(K)\,K/N\right)^2$. Therefore, the condition can be read as $\zeta_0^2(K)\,K/N \to 0$, which is comparable to that of Newey (1997).

Since g_{Kk} are measurable, Proposition 2.3 in the previous section implies that $\{g_{Kk}(y_{i,t})\}$ is also α-mixing whose mixing coefficient is of the same order of magnitude as that of $\{y_{i,t}\}$ for all $k = 1, 2, \cdots, K$ by Assumptions E1 and E2. Therefore, in what follows, we will simply let the mixing coefficient of $\{g_{Kk}(y_{i,t})\}$ be $\alpha(\tau)$, which is originally the mixing coefficient of $\{y_{i,t}\}$. Since the mixing coefficient is only meaningful in its order of magnitude, such an abuse of notation does not lose generality. If we assume $g_{Kk}(y)$ are uniformly bounded over $y \in \mathcal{Y}_c$ with probability one for all k, then the process $\{g_{Kk}(y)\}$ satisfies the condition A3.1 in Robinson (1983) since $\sum_{\tau=1}^{\infty}\alpha(\tau) < \infty$. On the other hand, if we relax boundedness of $g_{Kk}(y)$ to the finite moment condition, then the process $\{g_{Kk}(y)\}$ satisfies the condition A3.2 in Robinson (1983) since

$\sum_{\tau=1}^{\infty} \alpha(\tau)^{1-2/\nu} < \infty$ is still satisfied, where $\nu > 4$ as in Assumption E1. More precisely, we can have an alternative condition to Assumption W1 as follows.

Assumption W1′ *(i) For every K, the $K \times K$ matrix $\Gamma_K = \mathbb{E}\underline{g}_K(y)\underline{g}_K'(y)$ has the smallest eigenvalue bounded away from zero and the largest eigenvalue bounded above. (ii) For every K, there is a sequence of constants $\zeta_{0\nu}(K)$* satisfying $\zeta_{0\nu}(K) \geq \max_{1\leq k\leq K} \mathbb{E}|g_{Kk}(y)|^{\nu/2}$ *and $K = K(N,T)$ such that $\zeta_{0\nu}(K)^{2/\nu} K^2/NT \to \infty$ as $N, T \to \infty$.*

Using either of the conditions does not alter the result much because the boundedness condition on $g_{Kk}(y)$ is mainly for controlling temporal dependence and for using an adequate mixing inequality in Proposition 2.4. In this study, therefore, we use the condition W1 instead of W1′. Note that, unlike Robinson (1983), Assumption W1 implies Assumption W1′ only when the entire support of $y_{i,t}$ is bounded. Finally, we need an additional condition, which specifies a rate of approximation for the series, as in Newey (1997),

Assumption W2 *There exist θ_K and $\delta > 0$ satisfying $\sup_{y\in\mathcal{Y}_c} \left|m(y) - g_K(y)'\theta_K\right| = O\left(K^{-\delta}\right)$ as $K \to \infty$.*

The uniform approximation condition is a conventional one in the series approximation literature and it is useful to specify a rate of approximation for the series. We only specify the convergence rate of the series $g_K(y)$ over a bounded support $\mathcal{Y}_c$ instead of the entire support. This is because we are only interested in estimating m over a specific bounded range $\mathcal{Y}_c$. As noted in Newey (1997), δ is related to the smoothness of $m(y)$ and the dimensionality of y. For example, for regression splines and power series, this assumption will be satisfied with $\delta = D/\dim(y)$, where D is the number of continuous derivatives of $m(y)$ that exists and $\dim(y)$ is the dimension of y. When we consider an $AR(1)$ model as in (1), therefore, δ corresponds to the smoothness of $m(y)$; the following condition can replace Assumption W4. Assumption W2′ is intuitively more appealing in that the smoother $m(y)$, the easier it is to approximate it.

Assumption W2′ *There exists a nonnegative integer $D(=\delta)$ such that $m(y)$ is continuously differentiable to order D on $\mathcal{Y}_c$.*

Since we are only interested in estimating the unknown function over the bounded support $\mathcal{Y}_c$, $m(y)$ only needs to be smooth enough on $\mathcal{Y}_c \subset \mathcal{Y}$.

3.3 Asymptotic properties

In this subsection, we derive the main asymptotic results of the WG series estimator $\widehat{m}(y)$ defined in (8). The first theorem provides the mean square convergence rate of $\widehat{m}(y)$.

Theorem 3.1 (Convergence rate) *Under Assumptions E1, E2, W1 and W2,*

$$\int_{y\in\mathcal{Y}_c} \left[\widehat{m}(y) - m(y)\right]^2 dP(y) = O_p\left(K/NT + K^{-2\delta} + \zeta_0^2(K)\,K/NT\right) \tag{9}$$

as $N, T \to \infty$, where $P(y)$ denotes the cumulative distribution function of $y_{i,t}$.[10]

Since $K^{-2\delta} \to 0$ and $\zeta_0^2(K)\,K/NT \to 0$, Theorem 3.1 implies that $\int_{y\in\mathcal{Y}_c} \left[\widehat{m}(y) - m(y)\right]^2 dP(y) \to_p 0$ as $N, T \to \infty$, where "$\to_p$" means convergence in probability. For the mean square convergence rate (9), the first term K/NT in (9) essentially corresponds to the convergence rate of the variance, whereas the remaining terms, $K^{-2\delta}$ and $\zeta_0^2(K)\,K/NT$, correspond to the convergence rate of the bias. The third term $\zeta_0^2(K)\,K/NT$ is new and it does not appear in the conventional series estimators for the cross section case as in Newey (1997). Just as for pooled estimation in linear dynamic panels, it is from the endogeneity bias. It reduces the mean square convergence rate compared with the cross section case since $\zeta_0(K)$ is a nondecreasing function of K.

If we assume $g_K(\cdot)$ and $m(\cdot)$ are differentiable up to D-th order as in Assumption W2′, and if we introduce $\zeta_D(K) \geq \sup_{y\in\mathcal{Y}_c} \max_{1\leq k\leq K} \max_{s\leq D} \left|d^s g_{Kk}(y)/dy^s\right|$, which is assumed to be larger than $O\left(K^{-1/2}\right)$ and to exist, then we have a uniform convergence rate of $\widehat{m}(y)$ given by $\sup_{y\in\mathcal{Y}_c} \max_{s\leq D} \left|d^s\left(\widehat{m}(y) - m(y)\right)/dy^s\right| = O_p\left(K^{1/2}\zeta_D(K)\left[\zeta_0(K)\,K^{1/2}/\sqrt{NT} + K^{-\delta}\right]\right)$. Its derivation is provided in the proof of Theorem 3.1. Note that the uniform convergence rate is not optimal as discussed in Newey (1997). Recently, De Jong (2004) proposes a sharper rate of the bound under the stronger conditions for the *i.i.d.* cross section case. The first two terms of the mean square convergence rate in (9), however, attain Stone's (1982) optimal bound as noted in Newey (1997).

We now derive the asymptotic normality of the WG series estimator of the unknown function $m(\cdot)$ as follows. "$\to_d$" means convergence in distribution.

Theorem 3.2 (Asymptotic normality) *If Assumptions NT, E1, E2, W1 and W2 are satisfied and* $\sqrt{NT}K^{-\delta} \to 0$,[11] *then as $N, T \to \infty$*

$$v(y, K, N, T)^{-1/2}\left(\widehat{m}(y) - m(y) + (1/T)\,b_K(y)\right) \to_d \mathcal{N}(0, 1) \tag{10}$$

---

[10] In fact, the formulae (9) should read $\int_{y\in\mathcal{Y}_c} \left[\widehat{m}(y) - m(y)\right]^2 dP(y) = O_p\left(\zeta_0^2(K)\,K/NT + K^{-2\delta}\right)$ since $\zeta_0(K)$ is a nondecreasing function of K and thus $\zeta_0^2(K)\,K/NT$ dominates K/NT for large K. However, writing as in (9) is helpful to compare the result with the findings in Newey (1997).

[11] Since K is usually chosen not too large (mostly less than ten), the rate condition $\sqrt{NT}K^{-\delta} \to 0$ seems too strong and δ seems to be very large. However, if $K = K(N, T)$ is chosen to satisfy reasonably small rate with respect to N and T, e.g., $K = O\left((NT)^{1/6}\right)$, then δ only needs to satisfy $\delta > 3$. That is, m is continuously differentiable to order three. We will discuss more about selecting K in Remark 3.4.

for $y \in \mathcal{Y}_c$*, where* $v(y,K,N,T) = \sigma^2 g_K(y)' \Gamma_K^{-1} g_K(y)/NT$*,* $b_K(y) = g_K(y)' \Gamma_K^{-1}\Phi_K$*.* Φ_K *is given by* $\Phi_K = \sum_{j=0}^{\infty} cov\left(g_K(y_{i,t+j}), u_{i,t}\right)$ *satisfying* $\|\Phi_K\| < \infty$ *for each* K*. Asymptotic normality still holds after replacing* $v(y,K,N,T)$ *with its consistent estimator* $\widehat{v}(y,K,N,T) = \widehat{\sigma}^2 g_K(y)' \widehat{\Gamma}_K^{-1} g_K(y)/NT$*, where* $\widehat{\Gamma}_K = (1/NT)\sum_{i=1}^{N}\sum_{t=1}^{T} g_K^0(y_{i,t}) g_K^0(y_{i,t})'$ *and* $\widehat{\sigma}^2 = (1/N(T-1))\sum_{i=1}^{N}\sum_{t=1}^{T}\left(y_{i,t}^0 - \widehat{m}^0(y)\right)^2$*.*

Theorem 3.2 is similar to the *i.i.d.* cross section cases as in Andrews (1991) and Newey (1997). The only difference is that $\widehat{m}(y)$ has non-degenerating asymptotic bias incurred by the within transformation, especially when $\lim_{N,T\to\infty} N/T \neq 0$. Therefore, it requires bias correction as in (10) by adding $(1/T)\, b_K(y)$ for each $y \in \mathcal{Y}_c$. Also note that the rate of convergence in (10) is not $\sqrt{NT}$. As the usual nonparametric regression estimators, the convergence rate cannot achieve $\sqrt{NT}$ rate; it is slower than $\sqrt{NT}$ rate as the smoothing parameter shrinks. In (10), the smoothing parameter corresponds to $1/K$. Even though it is not explicitly revealed, the smoothing parameter is embedded in the $K \times K$ matrix $\sigma^2 g_K(y)' \Gamma_K^{-1} g_K(y)$. So the convergence rate is determined by the entire term of $v(y,K,N,T)^{-1/2} = \sqrt{NT}\left(\sigma^2 g_K(y)' \Gamma_K^{-1} g_K(y)\right)^{-1/2}$. Since we assume the smallest eigenvalue of Γ_K is bounded away from zero and its largest eigenvalue is bounded above for every K, if we simply let Γ_K be the identity matrix I_K, then the rate of convergence is $\sqrt{NT/K}$.

Finally, the following theorem suggests a bias corrected estimator for $m(\cdot)$.

Theorem 3.3 (Bias correction) *Under the same conditions as in Theorem 3.2, as* $N, T \to \infty$

$$v(y,K,N,T)^{-1/2}\left(\widetilde{m}(y) - m(y)\right) \to_d \mathcal{N}(0,1)$$

for $y \in \mathcal{Y}_c$*, where* $\widetilde{m}(y) = \widehat{m}(y) + (1/T)\,\widehat{b}_K(y)$ *and*

$$\widehat{b}_K(y) = g_K(y)'\left(\sum_{i=1}^{N}\sum_{t=1}^{T} g_K^0(y_{i,t})\, g_K^0(y_{i,t})'\right)^{-1}\sum_{i=1}^{N}\sum_{j=0}^{J}\sum_{t=1}^{T-j}\left(1 - \frac{j}{J+1}\right) g_K(y_{i,t+j})\,\widehat{u}_{i,t}^0$$

with $J = J(T) \leq O\left(T^{1/3}\right)$ *and* $\widehat{u}_{i,t}^0 = y_{i,t}^0 - \widehat{m}^0(y_{i,t-1})$*. Asymptotic normality still holds after replacing* $v(y,K,N,T)$ *with its consistent estimator* $\widehat{v}(y,K,N,T)$ *defined as in Theorem 3.2.*

Since the bias is $b_K(y) = g_K(y)' \Gamma_K^{-1}\Phi_K$ from Theorem 3.2, Theorem 3.3 follows by estimating $b_K(y)$ with $\widehat{b}_K(y) = g_K(y)' \widehat{\Gamma}_K^{-1}\widehat{\Phi}_K$, where $\left\|\widehat{\Gamma}_K^{-1} - \Gamma_K^{-1}\right\| = o_p(1)$ and $\left\|\widehat{\Phi}_K - \Phi_K\right\| = o_p(1)$ for large N and T. Recall that for each K, Φ_K is defined as $\Phi_K = \sum_{j=0}^{\infty} cov\left(g_K(y_{i,t+j}), u_{i,t}\right)$. In Appendix A1.10, it is shown that

$$\widehat{\Phi}_K = \frac{1}{NT}\sum_{i=1}^{N}\sum_{j=0}^{J} w(j,J)\sum_{t=1}^{T-j} g_K(y_{i,t+j})\,\widehat{u}_{i,t}^0$$

is a consistent estimator for Φ_K satisfying $\left\|\widehat{\Phi}_K - \Phi_K\right\| = o_p(1)$ for large N and T, provided that the truncation parameter, J, satisfies $J = J(T) \leq O\left(T^{1/3}\right)$ and that the weight function, $w(j, J)$, is uniformly bounded. Note that the truncation is necessary since there remain a smaller number of summands as j gets larger. This idea follows the studies on heteroskedasticity and autocorrelation consistent (HAC) estimation of covariance matrices such as White and Domowitz (1984), Newey and West (1987), and Andrews (1991) to name a few. The simple weights $(1 - j/(J+1))$ are borrowed from Newey and West (1987), in which J is required to be smaller than $O\left(T^{1/4}\right)$ for the consistency of the long-run autocovariance matrix estimator. Notice that we need a weaker condition that J to grow slower than $T^{1/3}$. We can modify the simple weight using kernel functions as in Andrews (1991).

Remark 3.4 (Determining the order of K**)** The smoothing parameters are chosen by minimizing the (integrated) mean square error. Similarly, we can determine the optimal order of K in terms of N and T by minimizing the mean square convergence rate (9). This result does not provide the exact value of K, but gives a guideline as to how to select it as a function of N and T. The basic idea is that K is chosen so that the two terms, $K^{-2\delta}$ and $\zeta_0^2(K) K/NT$ in (9), go to zero at the same rate.

For example, we have the explicit bound $\zeta_0(K) = O(K)$ for orthogonal polynomials over the compact support $\mathcal{Y}_c$, as noted in Newey (1997). Therefore, the mean square convergence rate is given by $O_p\left(K^3/NT + K^{-2\delta}\right)$ from Theorem 3.1, which is minimized with K such that $K^3/NT = K^{-2\delta}$. In other words, K needs to satisfy $K = O\left((NT)^{1/(3+2\delta)}\right)$. Meanwhile, K should also obey the rate condition given by $\zeta_0^4(K) K^2/NT \to 0$ in Assumptions W1, which implies $K < O\left((NT)^{1/6}\right)$ for orthogonal polynomials. Therefore, if $\delta > 3/2$, then both rate conditions are satisfied and we can simply let $K = C_1 (NT)^{1/7}$ for some constant $0 < C_1 < \infty$. This is identical to the finding $K = O\left(N^{1/7}\right)$ in Ai and Chen (2003) for the cross section case. Similarly, for B-splines over the bounded support $[-1, 1]$, we have $\zeta_0(K) = O\left(K^{1/2}\right)$ as noted in Newey (1997). In this case, the optimal order of K should satisfy $K = O\left((NT)^{1/(2+2\delta)}\right)$ and $K < O\left((NT)^{1/4}\right)$. Therefore, we need $\delta > 1$ and we can let, for example, $K = C_2 (NT)^{1/5}$ for some constant $0 < C_2 < \infty$.

However, if the series estimator, $\widehat{m}(y)$, needs to satisfy the asymptotic normality, an additional rate condition, $\sqrt{NT} K^{-\delta} \to 0$ from Theorem 3.2, is also required. This condition changes the range of δ. For example, orthogonal polynomials[12] require $\delta > 3$, and B-splines require $\delta > 2$. This implies that, loosely speaking, we need twice as much smoothness of m for the asymptotic normality. Moreover, the optimal choices of K are also changed to satisfy $K < O\left((NT)^{1/9}\right)$ for orthogonal polynomials, and $K <$

[12] For orthogonal polynomials, K needs to satisfy $K^6/NT + \sqrt{NT}/K^\delta \to 0$ in this case. The first term implies that $K = C_3 (nT)^{(1/6)-\kappa_1}$, whereas the second term implies that $K = C_4 (nT)^{(1/2\delta)+\kappa_2}$ for $\kappa_1, \kappa_2, \delta > 0$ and $0 < C_3, C_4 < \infty$. If we set these two terms same, we have $(1/6) - \kappa_1 = (1/2\delta) + \kappa_2$ and thus $(1/6) = (1/2\delta) + \kappa_3$ for $\kappa_3 > 0$. Therefore, $\delta > 3$. For the B-splines case, we can find the range of δ similarly if we use the condition $K^4/NT + \sqrt{NT}/K^\delta \to 0$.

$O\left((NT)^{1/6}\right)$ for B-splines.

4 Extensions

4.1 Partial linear model

Direct applications of the pure autoregressive panel model (1) are limited in empirical studies. In this subsection, we thus allow exogenous variables $x_{i,t} \in \mathbb{R}^r$ in the regression. For example, we consider a partial linear model given by

$$y_{i,t} = \mu_i + m\left(y_{i,t-1}\right) + \gamma' x_{i,t} + u_{i,t}, \tag{11}$$

where γ is an $r \times 1$ parameter vector. In the time series literature, the conventional partial linear model assumes that the lagged values are of linear form, whereas the exogenous variables are of nonparametric form: $y_t = \rho y_{t-1} + m\left(x_t\right) + u_t$. The purpose of such model is to control out the effects from x_t. In (11), on the other hand, we are interested in the partial effects of exogenous variable $x_{i,t}$ to $y_{i,t}$, where the dynamics of $y_{i,t}$ on its own lag is controlled by an unknown function m. It is a clear extension of the existing dynamic panel literature with $m\left(y_{i,t-1}\right) = \rho y_{i,t-1}$. In some cases, moreover, we are more interested in uncovering the unknown shape of dynamics in $y_{i,t}$ (i.e., $m(\cdot)$), where other characteristics $x_{i,t}$ are controlled linearly. Such examples can be found in semiparametric cross-country growth regressions as in Liu and Stengos (1999).

We could relax the linear part so that it is also fully nonparametric as

$$y_{i,t} = \mu_i + m\left(y_{i,t-1}, x_{i,t}\right) + u_{i,t}. \tag{12}$$

However, such generalization is limited in empirical studies because of the curse of dimensionality$-$as the dimension increases it is more difficult to estimate $m(\cdot)$ nonparametrically. Therefore, we need more restriction on $m(\cdot,\cdot)$, such as additivity, to reduce the dimension. That is, we need to assume $m\left(y_{i,t-1}, x_{i,t}\right) = m_y\left(y_{i,t-1}\right) + m_x\left(x_{i,t}\right)$ with $m_y : \mathbb{R}^1 \to \mathbb{R}^1$ and $m_x : \mathbb{R}^r \to \mathbb{R}^1$. A number of studies on semiparametric estimation use such approaches and use marginal integration to estimate the additive nonparametric component. For identification convenience, we assume $m_y(0) = m_x(0) = 0$ and we exclude the constant term in $x_{i,t}$, in this case. If we assume similar conditions on the series approximation for m_x and m_y, we can derive the asymptotic distribution of the series estimator for (12). Since $x_{i,t}$ is strictly exogenous, the conditions for m_x should correspond to those in Porter (1996). The following condition guarantees that an autoregressive process with exogenous variables, $\{y_{i,t}\}$ satisfying (12), is stationary and mixing as in Section 2. This condition also ensures the stationarity and mixing for $\{y_{i,t}\}$ in (11) since the partial linear model (11) is a special case of (12). Though we provide general conditions for (12), we will mainly examine the partial linear specification (11), which is a special case of (12) but more relevant in empirical studies.

Assumption E3 (Stability condition) *We let* $\phi_i(z) = \mu_i + m(z)$, *where* $z = (y_1, ..., y_p; x_1, ..., x_q) \in \mathbb{R}^p \times \mathbb{R}^{qr}$. *Then, (i)* $\{x_{i,t}\}$ *and* $\{u_{i,t}\}$ *are mutually independent and i.i.d;* $\{u_{i,t}\}$ *is independent of* μ_i *for all* i *and* t. *(ii)* $\{u_{i,t}\}$ *has a density, which is positive almost everywhere; and a marginal distribution, which is absolutely continuous with respect to the Lebesgue measure on* $\mathbb{R}$. *(iii) For each* i, *there exist constants* $c_i > 0$; $\overline{z} > 0$; $a_i, \cdots, a_p \geq 0$; *and a locally bounded measurable function* $h : \mathbb{R}^r \to [0, \infty)$ *such that* $|\phi_i(z)| \leq \sum_{j=1}^{p} a_j |y_j| + \sum_{\ell=1}^{q} h|x_\ell| - c_i$ *if* $\|z\| > \overline{z}$ *and* $\sup_{z:\|z\| \leq \overline{z}} |\phi_i(z)| < \infty$, *where* $w^p - a_1 w^{p-1} - \cdots - a_p \neq 0$ *for* $|w| \geq 1$ *and* $\|z\| = \max\{|y_1|, ..., |y_p|, |x_1|, ..., |x_q|\}$. *(iv)* $\mathbb{E}h(x_{i,1}) + \mathbb{E}|u_{i,1}|^{\nu} < \infty$ *for some* $\nu > 4$ *and for all* i. *(v) The Markov process* $\{y_{i,t}\}$ *has a homogeneous transition probability; and the initial values of* $y_{i,t}$ *is drawn from the invariant distribution.*

Assumption E3 is an extension of pure time series models discussed in Doukhan (1994: Section 2.4.2, Theorem 7). Conditional on μ_i, this condition ensures that $\{y_{i,t}\}$ is geometrically ergodic over t and thus β-mixing with exponentially decaying mixing coefficients for each i. One remark is that the condition assumes the exogenous variables $x_{i,t}$ are *i.i.d.* for all i and t, which is rather strong. However, extending to weakly dependent process $x_{i,t}$ over i, but with keeping independence across i, should not be complicated. For example, Doukhan (1994) considers stationary Markov $\{x_{i,t}\}$, in fact. We conjecture that conditional on μ_i for each i, provided $\{x_{i,t}\}$ is mixing with the same mixing coefficients as $\{y_{i,t}\}$, then the stability condition should also hold. In this case, we are implicitly assuming that $x_{i,t} = \xi\left(x_{i,t}^*, \mu_i\right)$, where $\xi : \mathbb{R}^2 \to \mathbb{R}^1$ and $x_{i,t}^*$ is a stable stochastic process independent of μ_i.

The partial linear model in (11) cannot be directly estimated by Robinson (1988)'s two step estimation. It is because individual effects cannot be eliminated once the conditional expectation on $y_{i,t-1}$ is subtracted from the equation (11). To show this, we take conditional expectations on (11) to have

$$\mathbb{E}(y_{i,t}|y_{i,t-1}) = \mathbb{E}(\mu_i|y_{i,t-1}) + m(y_{i,t}) + \gamma'\mathbb{E}(x_{i,t}|y_{i,t-1}) \tag{13}$$

since $\mathbb{E}(u_{i,t}|y_{i,t-1}) = 0$ by assumption. We subtract (13) from (11), and obtain

$$[y_{i,t} - \mathbb{E}(y_{i,t}|y_{i,t-1})] = [\mu_i - \mathbb{E}(\mu_i|y_{i,t-1})] + \gamma'[x_{i,t} - \mathbb{E}(x_{i,t}|y_{i,t-1})] + u_{i,t},$$

in which we cannot eliminate $[\mu_i - \mathbb{E}(\mu_i|y_{i,t-1})]$ either by the within transformation or the first-differencing transformation. This is because $[\mu_i - \mathbb{E}(\mu_i|y_{i,t-1})]$ is a function of not only μ_i but also $y_{i,t-1}$, which depends on the time index t. This illustration suggests that we need to eliminate fixed effects before estimation. Porter (1996) proposes two step estimation, in which $m(\cdot)$ is estimated using regression residuals from projecting $y_{i,t}$ on the individual dummy variables and $x_{i,t}$. We, on the other hand, suggest one step estimation using the within-transformed series functions.

In the partial linear model (11), we first eliminate fixed effects, μ_i, by the within transformation:

$$y_{i,t}^0 = m^0(y_{i,t-1}) + \gamma' x_{i,t}^0 + u_{i,t}^0.$$

For convenience, we define several matrix notations. We let $\mathbf{g}_K^0 = \left(g_K^0(y_{1,0}), \cdots, g_K^0(y_{N,T-1})\right)'$, $\mathbf{y}^0 = (y_{1,1}, \cdots, y_{N,T})'$, $\mathbf{x}^0 = \left(x_{1,1}^0, \cdots, x_{N,T}^0\right)'$, $\widehat{\mathbf{m}}^0 = \left(\widehat{m}^0(y_{1,0}), \cdots, \widehat{m}^0(y_{N,T-1})\right)'$, $M_x = I_{NT} - \mathbf{x}^0\left(\mathbf{x}^{0\prime}\mathbf{x}^0\right)^{-1}\mathbf{x}^{0\prime}$ and $M_g = I_{NT} - \mathbf{g}_K^0\left(\mathbf{g}_K^{0\prime}\mathbf{g}_K^0\right)^{-1}\mathbf{g}_K^{0\prime}$ with assuming both $\mathbf{x}^{0\prime}\mathbf{x}^0$ and $\mathbf{g}_K^{0\prime}\mathbf{g}_K^0$ are nonsingular. Then, the WG series estimator for $m(\cdot)$ is given by $\widehat{m}(y) = g_K(y)'\widehat{\theta}_K$ for $y \in \mathcal{Y}_c$, where $\widehat{\theta}_K = \left(\mathbf{g}_K^{0\prime}M_x\mathbf{g}_K^0\right)^{-1}\mathbf{g}_K^{0\prime}M_x\mathbf{y}^0$. The parameter of the linear part, γ, can be estimated either by $\widehat{\gamma} = \left(\mathbf{x}^{0\prime}\mathbf{x}^0\right)^{-1}\mathbf{x}^{0\prime}\left(\mathbf{y}^0 - \widehat{\mathbf{m}}^0\right)$ or $\widehat{\gamma} = \left(\mathbf{x}^{0\prime}M_g\mathbf{x}^0\right)^{-1}\mathbf{x}^{0\prime}M_g\mathbf{y}^0$. Both estimation procedures yield the same result using the standard argument of partitioned regressions. We also let Σ be the $(K+r)\times(K+r)$ variance-covariance matrix of $\left(\underline{g}_K(y_{i,t-1})', x_{i,t}'\right)'$, whose smallest eigenvalue is bounded above zero and the largest eigenvalue is bounded for every K. We decompose it into

$$\Sigma = \begin{pmatrix} \Sigma_{gg} & \Sigma_{gx} \\ \Sigma_{xg} & \Sigma_{xx} \end{pmatrix} \begin{matrix} K \\ r \end{matrix}$$
$$\begin{matrix} K & \quad r \end{matrix}$$

conformably as $\left(\underline{g}_K(y_{i,t-1})', x_{i,t}'\right)'$. Recall that the conditional variance of $\underline{g}_K(y_{i,t-1})$ given $x_{i,t}$ is defined as $\Sigma_{gg\cdot x} = \Sigma_{gg} - \Sigma_{gx}\Sigma_{xx}^{-1}\Sigma_{xg}$ and the conditional variance of $x_{i,t}$ given $\underline{g}_K(y_{i,t-1})$ is defined as $\Sigma_{xx\cdot g} = \Sigma_{xx} - \Sigma_{xg}\Sigma_{gg}^{-1}\Sigma_{gx}$. We now derive the asymptotic distribution of the partial linear model estimators in (11).

Theorem 4.1 (Partial linear model) *Under Assumptions NT, E3, W1 and W2, as* $N, T \to \infty$

$$v_x(y, K, N, T)^{-1/2}\left(\widehat{m}(y) - m(y) + (1/T)\, g_K(y)'\,\Sigma_{gg\cdot x}^{-1}\Phi_K\right) \to_d \mathcal{N}(0,1)$$

for $y \in \mathcal{Y}_c$, *and*

$$\sqrt{NT}V_x^{-1/2}\left(\widehat{\gamma} - \gamma - (1/T)\,\Sigma_{xx\cdot g}^{-1}\Sigma_{xg}\Sigma_{gg}^{-1}\Phi_K\right) \to_d \mathcal{N}(0,1),$$

where $v_x(y, K, N, T) = \sigma^2 g(y)'\,\Sigma_{gg\cdot x}^{-1} g(y)/NT$ *and* $V_x = \sigma^2\Sigma_{xx\cdot g}^{-1}$. *Asymptotic results still holds after replacing* $v_x(y, K, N, T)$ *and* V_x *with their consistent estimators:* $\widehat{v}_x(y, K, N, T) = \widehat{\sigma}^2 g(y)'\,\widehat{\Sigma}_{gg\cdot x}^{-1} g(y)$ *and* $\widehat{V}_x = \widehat{\sigma}^2\widehat{\Sigma}_{xx\cdot g}^{-1}$, *where the conditional variance matrices are constructed using a consistent estimator* $\widehat{\Sigma} = (1/NT)\sum_{i=1}^N\sum_{t=1}^T\left[g_K^0(y_{i,t-1})', x_{i,t}^{0\prime}\right]'\left[g_K^0(y_{i,t-1})', x_{i,t}^{0\prime}\right]$.

Unlike the conventional partial linear models, the estimator for the linear part, $\widehat{\gamma}$, exhibits asymptotic bias. But the direction of the bias is opposite to that of the nonparametric component $\widehat{m}(y)$. As in Theorem 3.3, bias correction can be conduct as follows.

Corollary 4.2 (Bias correction) *Under the same conditions as in Theorem 4.1, as $N, T \to \infty$*

$$v_x(y, K, N, T)^{-1/2}\left(\widetilde{m}(y) - m(y)\right) \to_d \mathcal{N}(0,1)$$

for $y \in \mathcal{Y}_c$, and

$$\sqrt{NT} V_x^{-1/2}\left(\widetilde{\gamma} - \gamma\right) \to_d \mathcal{N}(0,1),$$

where

$$\begin{aligned}
\widetilde{m}(y) &= \widehat{m}(y) + (1/T)\, g_K(y)' \widehat{\Sigma}_{gg\cdot x}^{-1} \widehat{\Phi}_K = g_K(y)' \left(\mathbf{g}_K^{0\prime} M_x \mathbf{g}_K^0\right)^{-1} \left\{\mathbf{g}_K^{0\prime} M_x \mathbf{y}^0 + (1/T)\, \widehat{\Phi}_K\right\} \quad \textit{and} \\
\widetilde{\gamma} &= \widehat{\gamma} - (1/T)\, \widehat{\Sigma}_{xx\cdot g}^{-1} \widehat{\Sigma}_{xg} \widehat{\Sigma}_{gg}^{-1} \widehat{\Phi}_K = \left(\mathbf{x}^{0\prime} M_g \mathbf{x}^0\right)^{-1} \mathbf{x}^{0\prime} \left\{M_g \mathbf{y}^0 - (1/T)\, \mathbf{g}_K^0 \left(\mathbf{g}_K^{0\prime} \mathbf{g}_K^0\right)^{-1} \widehat{\Phi}_K\right\}.
\end{aligned}$$

$\widehat{\Phi}_K$ *is defined as in Section 3.3.*

4.2 Two stage instrumental variables estimator

The main results of this paper are all based on the within-transformed model in (3). In this subsection, we instead consider nonparametric estimation for the first-differenced model in (4) given by

$$\Delta y_{i,t} = \ell\left(y_{i,t-1}, y_{i,y-2}\right) + \Delta u_{i,t}$$

where $\ell(y_1, y_2) = m(y_1) - m(y_2)$. Notice that $\ell(y_1, y_2) \neq m(y_1 - y_2)$. As we discussed in Section 3.1, we cannot simply regress $\Delta y_{i,t}$ on a pair of regressors $x_{i,t} = (y_{i,t-1}, y_{i,t-2})'$ because of the following two problems. The first one is an endogeneity problem since $\mathbb{E}(\Delta u_{i,t}|x_{i,t}) \neq 0$. We thus need to introduce $v \times 1$ vector of instruments $z_{i,t}$ satisfying $\mathbb{E}(\Delta u_{i,t}|z_{i,t}) = 0$ and $\mathbb{E}(x_{i,t}|z_{i,t}) \neq 0$. In dynamic panel regressions, instruments conventionally consist of the lagged observations of $y_{i,t-s}$ for $s \geq 2$. Using $z_{i,t}$, we conduct two stage estimation, such as the kernel IV regression as in Darolles et al. (2003), or sieve minimum distance estimation as in Newey and Powell (2003) and Ai and Chen (2003). The second problem is restoring the estimator of the original function $\widehat{m}(\cdot)$ from $\widehat{\ell}(\cdot)$. This identification problem is closely discussed in Porter (1996), where he uses the partial integration method as in Newey (1994). As noted in Porter (1996), however, the integration method has a drawback in that the unknown function $m(\cdot)$ can be estimated only up to a constant addition by $\mathbb{E}m(\cdot)$. On the other hand, if we preserve the additive form of $\ell(y_1, y_2) = m(y_1) - m(y_2)$, we can easily restore $\widehat{m}(\cdot)$ from $\widehat{\ell}(\cdot)$ by using the normalization condition ID in Section 3.1 (i.e., $m(0) = 0$).

The most appealing property of the IV-based method is that it does not need large T because the consistency result can be derived under fixed T and large N asymptotics. Therefore, the IV-based method has been worked out for conventional microeconomic data, in particular. When the length of time T is large, however, the total number of available instruments increases and it generates an inefficiency problem.[13] In this case, the within-transformation-based method seems to be more appropriate.

We now present how to extend two stage nonparametric IV estimation of Newey and Powell (2003) to the context of dynamic panels. We only look at large N and fixed T cases, and argue that the consistency result of Newey and Powell (2003) still holds in dynamic panel models. We do not consider the both large N and T case and leave it as a topic for future research. If we use the series approximation as (6), we have

$$m(y_1) - m(y_2) \approx \sum_{k=1}^{K} \theta_{Kk} \left[g_{Kk}(y_1) - g_{Kk}(y_2) \right], \tag{14}$$

and

$$\mathbb{E}(\Delta y_{i,t} | z_{i,t}) = \sum_{k=1}^{K} \theta_{Kk} \mathbb{E}\left[g_{Kk}(y_{i,t-1}) - g_{Kk}(y_{i,t-2}) | z_{i,t} \right].$$

In the first stage, we estimate the conditional expectation by any nonparametric estimation method to have $\widehat{\mathbb{E}}\left[g_{Kk}(y_{i,t-1}) - g_{Kk}(y_{i,t-2}) | z_{i,t} \right]$. In the second stage, if we let $\Delta \widehat{g}_{Kk}(z_{i,t}) = \widehat{\mathbb{E}}\left[g_{Kk}(y_{i,t-1}) - g_{Kk}(y_{i,t-2}) | z_{i,t} \right]$ and $\Delta \widehat{g}_K(z_{i,t}) = (\Delta \widehat{g}_{K1}(z_{i,t}), \cdots, \Delta \widehat{g}_{KK}(z_{i,t}))'$, we can estimate $\theta_K = (\theta_{K1}, \cdots, \theta_{KK})'$ by solving the minimization problem:[14]

$$\widehat{\theta}_K = \arg\min_{\theta_K} \sum_{i=1}^{N} \left(\Delta y_i - \Delta \widehat{g}_K(z_i)' \theta_K \right)' R \left(\Delta y_i - \Delta \widehat{g}_K(z_i)' \theta_K \right), \tag{15}$$

where Δy_i and $\Delta \widehat{g}_K(z_i)$ are $T \times 1$ stacks of $\Delta y_{i,t}$ and $\Delta \widehat{g}_K(z_{i,t})$; and R is a $T \times T$ matrix given by

$$R = \begin{pmatrix} 2 & -1 & 0 & \cdots & 0 \\ -1 & 2 & -1 & \ddots & \\ 0 & -1 & \ddots & \ddots & 0 \\ & \ddots & \ddots & 2 & -1 \\ 0 & \cdots & 0 & -1 & 2 \end{pmatrix}^{-1}.$$

The nonparametric estimate is then obtained by $\widehat{m}(y) = \sum_{k=1}^{K} \widehat{\theta}_{Kk} g_{Kk}(y)$ for any $y \in \mathcal{Y}_c$. Notice that

[13] The IV estimator using a fixed number of instrumental variables will remain well-defined, and will be consistent regardless of whether T or N or both tend to infinity. However, the total number of available instruments increases as $T \to \infty$ since they consist of lagged $y_{i,t}$. It thus generates the many instruments problem. In this case, we need to let the number of instruments be fixed to avoid any potential problem. As noted in Alvarez and Arellano (2002), even if we allow the number of instruments to increase as T grows, the GMM estimator is still consistent as long as T grows much slower than N, i.e., $(\log T)^2 / N \to 0$.

[14] Newey and Powell (2003) use penalized least squares, where the penalty term is added by imposing the compactness conditions. But if we let the unknown function m to be bounded over some bounded support $\mathcal{Y}_c$, we do not have such addition restrictions.

R is derived from the variance-covariance matrix of $\Delta u_t = (\Delta u_{1,t}, , \cdots \Delta u_{i,T})'$, which is not spherical. The minimization problem (15) is therefore a simple GLS problem with a known covariance matrix. The pointwise consistency of $\widehat{m}(\cdot)$ for large N but fixed T can be derived similarly as Newey and Powell (2003), or Ai and Chen (2003) under the proper regularity conditions and the metric. The regularity conditions need to be modified in the context of dynamic panels, but once we fix T the extension is closely related to the multivariate regression. Detailed conditions are discussed in Appendix B.

As in Porter (1996), we can alternatively approximate ℓ using series functions $h_{Kk} : \mathbb{R}^2 \to \mathbb{R}^1$ for all k, given by

$$\ell(y_1, y_2) \approx \sum_{k=1}^{K} h_{Kk}(y_1, y_2)\, \theta_{Kk}. \tag{16}$$

We estimate $\widehat{h}_{Kk}(z) = \widehat{\mathbb{E}}[h_{Kk}(y_{1,}y_2)|z]$ using any nonparametric method and conduct series estimation given as

$$\widehat{\theta}_K = \arg\min_{\theta_K} \sum_{i=1}^{N} \left(\Delta y_i - \Delta\widehat{h}_K(z_i)'\theta_K\right)' R \left(\Delta y_i - \Delta\widehat{h}_K(z_i)'\theta_K\right),$$

which produces $\widehat{\ell}(y_1, y_2) = \sum_{k=1}^{K} \widehat{\theta}_{Kk} h_{Kk}(y_1, y_2)$ for any $y_1, y_2 \in \mathcal{Y}_c$. However, this approach suffers an identification problem of restoring $\widehat{m}$ from $\widehat{\ell}$.

Since we rely on conditional expectations given instruments z, we are facing an identification issue of ℓ from

$$\pi(z) = \mathbb{E}(\Delta y|z) = \mathbb{E}(\ell(y_1, y_2)|z) = \int \ell(y_1, y_2) F(dy_1|z),$$

where $y_2 \in z$. $F(y_1|z)$ is the conditional distribution of y_1 given z. As noted in Newey and Powell (2003), estimating $\pi(z)$ and $F(dy_1|z)$ are not difficult using the observations. Restoring $\ell(y_1, y_2)$ from these information, however, is not straightforward. This is the first identification issue and we need the following condition to solve this problem:

$$\int \ell(y_1, y_2) F(dy_1|z_{i,t}) = \int \ell^*(y_1, y_2) F(dy_1|z_{i,t}) \quad \text{implies} \quad \ell(y_1, y_2) = \ell^*(y_1, y_2).$$

This identification condition, however, only guarantees the uniqueness of $\ell(y_1, y_2)$ if its existence is presumed. One remaining important issue is the continuity assumption to avoid the ill-posed inverse problem. That is, if ℓ is not continuous in π and F then the consistency of $\hat{\ell}$ does not follow from the consistency of $\hat{\pi}$ and $\hat{F}$. We therefore also assume the continuity of ℓ in π and F. Blundell and Powell (2003) intensively discuss the identification problem in nonparametric IV regressions.

Even when the identification problem of ℓ is solved, we still have another identification issue: how to restore the original function m from ℓ. Conventionally this is done by the partial integration method as in Newey (1994) and Porter (1996). For example, to restore the original function $m(y_1)$, we integrate $\ell(y_1, y_2)$ over y_2 with y_1 kept fixed. But the problem is that this method does not use the original structural information that two functions of y_1 and y_2 are the same but the sign: $\ell(y_1, y_2) = m(y_1) - m(y_2)$. So Porter (1996) imposes additional restrictions that $\ell(y_1, y_2) = -\ell(y_2, y_1)$ and $\ell(y, y) = 0$. This method, however, can only identify m up to a constant addition. If we instead assume a different identification condition that $m(0) = 0$ as in Assumption ID in Section 3.2, then the second identification problem is easily solved. This is because $\ell(y_1, 0) = m(y_1) - m(0) = m(y_1)$ if we preserve the additive form of $\ell(y_1, y_2)$. Therefore, $\widehat{m}(y)$ can be obtained from $\widehat{\ell}(y_1, y_2)$ by letting the second argument fix at zero. This second identification issue disappears, in fact, if we simply use the first approach (14) instead of (16) since $\widehat{m}(y)$ is obtained from $\sum_{k=1}^{K} \widehat{\theta}_{Kk} g_{Kk}(y)$ for any $y \in \mathcal{Y}_c$.

We can also extend two stage IV estimation under partial linear models with exogenous variables. The estimation strategy for the partial linear model, after the first-differencing transformation, is identical to WG series estimation except conducting two stage estimations. In this case, we can consider more general models such as

$$y_{i,t} = \mu_i + m(y_{i,t-1}, w_{i,t}) + \gamma' x_{i,t} + u_{i,t},$$

where $x_{i,t}$ and $w_{i,t}$ do not need to be exogenous. We let m be additive (i.e., $m(y, w) = m_y(y) + m_w(w)$) so that $\Delta m(y, w) = \Delta m_y(y) + \Delta m_w(w)$. In this case, however, we need a richer set of instrumental variables $z_{i,t}$ satisfying $\mathbb{E}(\Delta u_{i,t} | z_{i,t}) = 0$ but $\mathbb{E}((w_{i,t}, w_{i,t-1}) | z_{i,t}) \neq 0$, $\mathbb{E}((x_{i,t}, x_{i,t-1}) | z_{i,t}) \neq 0$ and $\mathbb{E}(y_{i,t-1} | z_{i,t}) \neq 0$.

5 Simulations

To illustrate the implementation of the WG series estimation developed in Section 3, and to evaluate the finite sample performance of the nonparametric estimator $\widehat{m}(\cdot)$, we conduct simulation studies. The simulation is based on nonlinear dynamic panel models with fixed effects of five different dynamic structures given by

- Model 1 : $y_{i,t} = \mu_i + \{0.6 y_{i,t-1}\} + u_{i,t}$
- Model 2 : $y_{i,t} = \mu_i + \{\exp(y_{i,t-1}) / (1 + \exp(y_{i,t-1})) - 0.5\} + u_{i,t}$
- Model 3 : $y_{i,t} = \mu_i + \{\ln(|y_{i,t-1} - 1| + 1)\, sgn(y_{i,t-1} - 1) + \ln 2\} + u_{i,t}$
- Model 4 : $y_{i,t} = \mu_i + \{0.6 y_{i,t-1} - 0.9 y_{i,t-1} / (1 + \exp(y_{i,t-1} - 2.5))\} + u_{i,t}$
- Model 5 : $y_{i,t} = \mu_i + \left\{0.3 y_{i,t-1} \exp\left(-0.1 y_{i,t-1}^2\right)\right\} + u_{i,t}$

for $i = 1, 2, \cdots, N$ and $t = 1, 2, \cdots, T$. Fixed effects μ_i are randomly drawn from $\mathcal{U}(0, 1)$ and $u_{i,t}$ from $\mathcal{N}(0, 1)$. Each nonlinear function is properly centered to satisfy $m(0) = 0$. The first model is a linear dynamic model, a benchmark structure. The second model is of the logistic function, which was also investigated in Ai and Chen (2003) in the cross section case. Whereas, the third model is adopted from Newey and Powell (2003). The fourth model is known as the smoothed threshold autoregressive (STAR) model in the time series literature. In the time series context, this nonlinear structure was used in analyzing economic business cycles as in Luukkonen and Teräsvirta (1991). Instead of using the indicator function for discrete threshold as in the conventional threshold autoregressive (TAR) models, it uses a smooth non-decreasing function. The general motivation is that we need not assume any abrupt changes over the states and we let the data tell us if the changes are abrupt or not. For the smooth indicator function, we use the logistic distribution function in this example. The fifth model is referred to as the amplitude-dependent exponential autoregressive model, which is discussed in Tong (1990).

A sample of $(N,T) = (100,50)$ data points was generated according to the above design, so $N/T = 2$ in this case. We estimate the unknown function $m(\cdot)$ by WG series estimation and we iterate the entire procedure 1000 times to examine the performances. For series estimation, we use both power series and cubic splines. The number of series functions K is determined to satisfy the order condition discussed in Remark 3.4. For example, when $(N,T) = (100,50)$, we let $K = 4$ for power series, where it satisfies $(NT)^{1/7} \leq K < (NT)^{1/6}$; we let $K = 8$ for regression splines, where it satisfies $(NT)^{1/5} \leq K < (NT)^{1/4}$. Note that for the cubic splines, we allow four knots since the other four terms are cubic polynomials, $(1, y, y^2, y^3)$.

The simulation results are displayed in Table 5.1. The integrated mean square errors (IMSE) and the integrated mean absolute errors (IMAE) are calculated for each case. The integrated mean square error is computed according to the discrete expression[15] $\sum_{j=0}^{121} (0.05) \left\{ (1/1000) \sum_{r=1}^{1000} \left(m(-3+0.05j) - \widehat{m}_r(-3+0.05j) \right)^2 \right\}$, where m is the true nonlinear function and $\widehat{m}_r$ is the estimate in rth replication. The integrated mean absolute error is similarly obtained as $\sum_{j=0}^{121} (0.05) \left\{ (1/1000) \sum_{r=1}^{1000} \left| m(-3+0.05j) - \widehat{m}_r(-3+0.05j) \right| \right\}$. Note that we only look at the estimation results over the bounded interval $[-3,3] = \mathcal{Y}_c$. Table 5.1 exhibits that the IMSE and the IMAE are smaller after bias corrections. A graphical representation is given in Appendix C. The graphs display the average values over 1000 replications. Before bias correction, power series approximation performs better then cubic splines. The bias correction, however, improves the fit for all the cases and the difference between power series and cubic splines becomes much smaller.

	Cubic Splines				Power Series			
	IMSE		IMAE		IMSE		IMAE	
	original	bias-c	original	bias-c	original	bias-c	original	bias-c
model 1	0.9228	0.4842	0.7959	0.5374	0.0450	0.0355	0.1441	0.1375
model 2	0.1010	0.0428	0.2472	0.1429	0.0458	0.0415	0.1416	0.1179
model 3	0.6174	0.3163	0.6318	0.4220	0.1771	0.0784	0.1733	0.0708
model 4	0.1930	0.1134	0.3509	0.2505	0.1341	0.0480	0.1120	0.0451
model 5	0.1111	0.0444	0.2681	0.1559	0.1387	0.0427	0.1162	0.0387

Table 5.1: Within-Group Series Estimation over 1000 iterations with $(N,T) = (100,50)$. "original" displays IMSE and IMAE before bias correction; "bias-c" displays IMSE and IMAE after bias correction.

[15] Ai and Chen (2003) also use the same discrete expressions.

6 Application: Cross-Country Growth Regression

Most of the empirical studies examining cross-country growth equations are based on the assumption that there is an underlying common linear specification as required by the Solow model. However, recent studies question the assumption of linearity and propose nonlinear alternatives allowing for multiple regimes of growth patterns among different countries. These models are consistent with the presence of multiple steady-state equilibria that classify countries into different groups with different convergence characteristics. See Durlauf and Johnson (1995), and Bernard and Durlauf (1996) for further discussion. In this context, the conventional approach is including dummy variables to look at different growth patterns for different groups. On the other hand, Liu and Stengos (1999) employ a semiparametric approach to model the growth equation and show the nonlinear growth patterns. This section also takes the semiparametric approach.

Liu and Stengos (1999) use the pooled cross-country data. As pointed out in Islam (1995), one drawback of the conventional single cross-country regression is that identical aggregate production functions need to be assumed for all the countries. The panel approach, on the other hand, allows for differences in the aggregate production functions across countries by including country-specific effect parameters (i.e., fixed effects). Such an approach will reduce the variable omission bias in the cross-country regression because unobserved country-specific effects will be captured in the fixed effects. Similarly as Islam (1995), we also use panel data to examine the growth patterns. However, this approach is different from Islam (1995) in that it considers a semiparametric model.

For the growth equation, we use the traditional approach based on the Solow type growth model assuming Cobb-Douglas production function (e.g., Mankiw, Romer and Weil, 1992). Combining Liu and Stengos (1999) and Islam (1995), we have the following partial linear growth equation:[16]

$$\Delta \ln y_{i,t} = \mu_i + m\left(\ln y_{i,t-1}\right) + \alpha_2 \ln s_{i,t} + \alpha_3 \ln\left(n_{i,t} + g + \delta\right) + \alpha_4 \ln h_{i,t} + u_{i,t}, \tag{17}$$

where $y_{i,t}$ is the GDP per capita of country i at year t, $\Delta \ln y_{i,t} = \ln y_{i,t} - \ln y_{i,t-1}$, $s_{i,t}$ is the saving rate. $u_{i,t}$ is simply assumed *i.i.d.* $n_{i,t}$ and g are the exogenous growth rates of population and technology, whereas δ is the constant rate of depreciation. Following Islam (1995), $g + \delta$ is set to equal to 0.05 for all i and t. All these variables are obtained from the Penn World Table (version 6.1)[17], which provides (unbalanced) panels for 168 countries from the year 1950 to 2000. $h_{i,t}$ is a proxy for the human capital measure, which is the average schooling years in the total population over age 25. It is obtained from Barro and Lee

[16] We also included time dummies in the regression, but projected them out after taking the within transformation. Whether including the time dummies or not, however, does not effects the results much.

[17] Heston, A., R. Summers, and B. Aten (2002). *Penn World Table Version 6.1*, Center for International Comparisons at the University of Pennsylvania (CICUP).

(2000)[18] for 115 countries in every five years from 1960 to 2000. Recall that in the growth equation (17), if $m(\ln y_{i,t-1}) = \alpha_1 \ln y_{i,t-1}$, then it supports the growth convergence if $\alpha_1 < 0$. Analogously, if the slope of $m(\cdot)$ is negative, then we can interpret that the growth equation supports the growth convergence.

In the empirical analysis, we use a balanced panel set for 73 countries. The list of countries are provided in Table D4 in Appendix D. OECD member countries[19] among the selected 73 countries are marked with asterisks. We conduct semiparametric estimation developed in Section 4.1 for three different sets of samples: entire 73 countries, 24 OECD countries and 49 non-OECD countries. For each data set, we choose two different panel frequencies: the annual panel and the quintannual (every five years) panel. Since the average schooling years h is available only in five-year time intervals, we can look at the effects of the human capital only for the quintannual panel. For the annual data, we use from 1960 to 2000 for the entire countries and the non-OECD countries, whereas we use from 1953 to 2000 for the OECD countries. For the quintannual data, we use the years of 1960, 1965, 1970, 1975, 1980, 1985, 1990, 1995 and 2000 for the entire countries and the non-OECD countries, whereas we use one additional year of 1955 for the OECD countries. In the conventional growth analysis, annual data is not used because they are likely affected by short-run factors. It is therefore difficult to recover long-run dynamics from high frequency data. Taking it into account, we choose to also employ a five year interval, which is the time span used by Islam (1995) among others. On the other hand, we analyze the annual data to increase the number of time series as in Lee, Longmire, Mátyás and Harris (1998). For the analysis with five-year time intervals, saving rates and population growth variables are averaged over each five-year period.

The estimation results are provided in Tables D1 to D3 and Figures D1 to D3 in Appendix D. The tables display the estimation results both for linear specifications (18) and for partial linear specifications (17). For the nonparametric part, we use cubic splines with four knots. The linear specification simply assumes $m(\ln y_{i,t-1}) = \alpha_1 \ln y_{i,t-1}$ in (17):

$$\Delta \ln y_{i,t} = \mu_i + \alpha_1 \ln y_{i,t-1} + \alpha_2 \ln s_{i,t} + \alpha_3 \ln (n_{i,t} + g + \delta) + \alpha_4 \ln h_{i,t} + u_{i,t}. \tag{18}$$

For the linear regressions, the results are close to Islam (1995) and all the estimates for the coefficient of $\ln y_{i,t-1}$ support growth convergence with 1% significance level. The bias correction, which is based on Lee (2005), does not change the results much. For the partial linear regressions, we cannot directly compare the results with the findings in Liu and Stengos (1999) since they estimate the effects of $\ln s_{i,t}$ nonparametrically as well as $\ln y_{i,t-1}$. In most of the cases, however, the estimates for the linear part

[18] Source : www.cid.harvard.edu/ciddata/ciddata.html.

[19] In 2000, the total number of OECD members are 30. But the following six countries are excluded in the analysis since the Penn World Table does not provide balanced panels from 1960 to 2000 for them: Czech Republic, Germany, Hungary, Luxembourg, Poland, Slovak Republic.

(i.e., $\ln s_{i,t}$, $\ln(n_{i,t}+g+\delta)$ and $\ln h_{i,t}$) are close to what we find in the linear growth equation (18) except for non-OECD countries.

Figures D1 to D3 show the nonlinear relations between the GDP growth ($\Delta \ln y_{i,t}$) and the logarithm of lagged GDP ($\ln y_{i,t-1}$) after the other variables–saving rates s, human capital h, population growth n, depreciation rate δ, and technical growth g–are controlled out. Before bias corrections, we can see that the convergence hypothesis is true for any data sets, particularly for countries in the middle to upper income range. This result is identical with the findings in Liu and Stengos (1999). However, after the bias correction,[20] only the OECD countries reveal the convergence patterns. (Figure D2) For the entire 73 countries and for the non-OECD countries, we can only find the convergence in the very upper income range. (Figure D1 and D3)

Finally, we conduct a very similar analysis as in Islam (1995), in that we rank countries based on the country-specific effect estimates. As discussed in Islam (1995), fixed effects reflect the unobserved country-specific effects such as production technology, resource endowments, institutions, and so forth. Though the precise interpretation of fixed effects is not available yet in the literature, we present our findings in Table D4 in Appendix D for comparison purposes with Islam (1995). The ranks are close to what is found in Islam (1995), for the top ranked countries in particular. But some countries show different ranks from Islam (1995): Venezuela and Syria show lower ranks; but Ireland and Barbados are in the top ranks.

7 Concluding Remarks

This paper calls into question the linear autoregressive structure in dynamic panels. This paper develops nonparametric estimation of nonlinear dynamic panels with fixed effects, which is an extension of the standard linear dynamic panel model to a nonparametric form that maintains additive fixed effects. Fixed effects are eliminated by the within transformation; the existing nonparametric estimation techniques based on instrumental variables are infeasible for the within-transformed models. In this case, series approximation is a good alternative since the within transformation of the unknown function is simply replaced by the within-transformed known series functions. It is a global approximation, so it is different from the kernel-based local approximation approach. It does not require instrumental variables; therefore, it enjoys all the good properties of the within-group estimators (or the least squares dummy variable estimators) especially under both large N and T.

Similar to Robinson (1983), we need to presume proper mixing condition to control temporal dependence of a dynamic panel. However, the approach of this paper involves more intensive techniques of the nonlinear time series because of the given data generating form, a nonlinear Markov process. Under proper restrictions

[20] We can use the bias correction formula developed in Section 4.1 because the asymptotic bias does not change whether $\Delta \ln y_{i,t}$ or $\ln y_{i,t}$ is used for the dependent variable. It is also true for the linear case (18).

on the nonlinear dynamics, the nonlinear homogeneous Markov process is stationary β-mixing conditional on the fixed effects. We derive the convergence rates and the asymptotic distribution of the within-group series estimator under large N and T asymptotics. Just as for pooled estimation in linear dynamic panels, an asymptotic bias is present, and a proper bias correction is suggested.

Some extensions of this framework are also considered under exogenous variables and partial linear models, which are relevant in applications. A brief empirical study on nonlinearities in cross-country growth regression is presented by employing the semiparametric approach developed in this paper. The findings seem to be in agreement with that of Liu and Stengos (1999), in that it suggests the presence of multiple regimes in growth patterns. But the convergence pattern in detail appears somewhat differently. In particular, before bias correction, the results exhibit the convergence for the countries in the middle to upper income range, which is in accordance with Liu and Stengos (1999). After bias correction, however, the results support the convergence hypothesis only for the OECD countries and the countries in the upper income range.

Appendix A: Mathematical Proofs

A.1 Useful lemmas

We first look at the following lemmas, which collect the basic building blocks that will be used in proving results in Section 3. We denote the mean deviated process $\underline{g}_K(y) = g_K(y) - \mathbb{E}g_K(y)$ for each K. The proof of each lemma is given in the following section. Lemma A1.1 and A1.2 first provide the convergence rate of the denominator of the within group type estimator $\widehat{\theta}_K$.

Lemma A1.1 *Under Assumptions E1, E2 and W1, for large N and T,*

$$\left\| \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \underline{g}_K(y_{i,t-1}) \underline{g}_K(y_{i,t-1})' - \Gamma_K \right\| = O_p\left(\frac{\zeta_0^2(K) K}{\sqrt{NT}} \right).$$

Lemma A1.2 *Under Assumptions E1, E2 and W1, for large N and T,*

$$\left\| \frac{1}{NT^2} \sum_{i=1}^{N} \sum_{t=1}^{T} \underline{g}_K(y_{i,t-1}) \sum_{s=1}^{T} \underline{g}_K(y_{i,s-1})' \right\| = O_p\left(\frac{\zeta_0^2(K) K}{\sqrt{NT}} \right).$$

Andrews (1991) and Newey (1994 and 1997) show that the variance estimation for linear functions of the series estimator is essentially the same as it is in least squares estimation for fixed K. We thus estimate Γ_K by $\widehat{\Gamma}_K = (1/NT)\sum_{i=1}^{N}\sum_{t=1}^{T} g^0(y_{i,t}) g^0(y_{i,t})'$ for every K. The following lemma shows that $\widehat{\Gamma}_K$ is consistent for Γ_K.

Lemma A1.3 *Under Assumptions E1, E2 and W1,* $\left\|\widehat{\Gamma}_K - \Gamma_K\right\| = O_p\left(\zeta^2(K) K/\sqrt{NT}\right)$ *and* $\left\|\widehat{\Gamma}_K^{-1} - \Gamma_K^{-1}\right\| = O_p\left(\zeta_0^2(K) K/\sqrt{NT}\right)$ *as $N, T \to \infty$, where* $\widehat{\Gamma}_K = (1/NT)\sum_{i=1}^{N}\sum_{t=1}^{T} g_K^0(y_{i,t}) g_K^0(y_{i,t})'$ *and* $\Gamma_K = \mathbb{E}\underline{g}_K(y_{i,t})\underline{g}_K(y_{i,t})'$.

We now look at the convergence rate of the numerator of $\widehat{\theta}_K$. Lemma A1.4 and A1.5 show that the convergence of the numerator $\left(O_p\left(\zeta_0(K) K^{1/2}/\sqrt{NT}\right)\right)$ turns out to be faster then the denominator $\left(O_p\left(\zeta_0^2(K) K/\sqrt{NT}\right)\right)$.

Lemma A1.4 *Under Assumptions E1, E2 and W1, for large N and T,*

$$\begin{aligned} \left\| \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \underline{g}_K(y_{i,t-1}) u_{i,t} \right\| &= O_p\left(\frac{\zeta_0(K) K^{1/2}}{\sqrt{NT}} \right) \quad \text{and} \\ \left\| \frac{1}{NT^2} \sum_{i=1}^{N} \sum_{t=1}^{T} \underline{g}_K(y_{i,t-1}) \sum_{s=1}^{T} u_{i,s} \right\| &= O_p\left(\frac{\zeta_0(K) K^{1/2}}{\sqrt{NT}} \right). \end{aligned}$$

Lemma A1.5 *Under Assumptions E1, E2, W1 and W2, for large N and T,*

$$\begin{aligned} \left\| \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \underline{g}_K(y_{i,t-1}) \left\{ m(y_{i,t-1}) - g_K(y_{i,t-1})'\theta_K \right\} \right\| &= O_p\left(\frac{\zeta_0(K) K^{1/2-\delta}}{\sqrt{NT}} \right) \quad \text{and} \\ \left\| \frac{1}{NT^2} \sum_{i=1}^{N} \sum_{t=1}^{T} \underline{g}_K(y_{i,t-1}) \sum_{s=1}^{T} \left\{ m(y_{i,s-1}) - g_K(y_{i,s-1})'\theta_K \right\} \right\| &= O_p\left(\frac{\zeta_0(K) K^{1/2-\delta}}{\sqrt{NT}} \right) \end{aligned}$$

The following three lemmas establish the building blocks for deriving asymptotic distribution of $\widehat{\theta}_K$.

Lemma A1.6 *Under Assumptions E1, E2 and W1, as $N, T \to \infty$,*

$$\frac{1}{\sqrt{NT}} \sum_{i=1}^{N} \sum_{t=1}^{T} \rho' \Gamma_K^{-1/2} \underline{g}_K(y_{i,t-1}) u_{i,t} \to_d \mathcal{N}\left(0, \sigma^2\right)$$

for some $K \times 1$ random vector ρ satisfying $\|\rho\| = 1$ and $\Gamma_K = \mathbb{E}\underline{g}_K(y_{i,t})\underline{g}_K(y_{i,t})'$.

Lemma A1.7 *Let $\lim_{N,T\to\infty} N/T = \kappa$, where $0 < \kappa < \infty$. Under Assumptions E1, E2, W1 and W2, for large N and T,*

$$\left\| \frac{1}{\sqrt{NT^3}} \sum_{i=1}^{N} \sum_{t=1}^{T} \underline{g}_K \left(y_{i,t-1} \right) \sum_{s=1}^{T} u_{i,s} - \sqrt{\kappa} \Phi_K \right\| = O_p \left(\frac{\zeta_0 \left(K \right) K^{1/2}}{\sqrt{NT}} \right),$$

where $\Phi_K = \sum_{j=0}^{\infty} cov \left(g_K \left(y_{i,t+j} \right), u_{i,t} \right)$ and $\| \Phi_K \| < \infty$ for each K.

Lemma A1.8 *Under Assumptions E1, E2, W1 and W2, for large N and T,*

$$\begin{aligned} \left\| \frac{1}{\sqrt{NT}} \sum_{i=1}^{N} \sum_{t=1}^{T} g_K \left(y_{i,t-1} \right) \left\{ m \left(y_{i,t-1} \right) - g_K \left(y_{i,t-1} \right)' \theta_K \right\} \right\| &= O_p \left(K^{-\delta} \sqrt{NT} \right) \quad \text{and} \\ \left\| \frac{1}{\sqrt{NT^3}} \sum_{i=1}^{N} \sum_{t=1}^{T} g_K \left(y_{i,t-1} \right) \sum_{s=1}^{T} \left\{ m \left(y_{i,s-1} \right) - g_K \left(y_{i,s-1} \right)' \theta_K \right\} \right\| &= O_p \left(K^{-\delta} \sqrt{NT} \right). \end{aligned}$$

Now the following lemmas provide consistency of the estimators for σ^2 and Φ_K. These results justify the bias correction formula in Theorem 3.5.

Lemma A1.9 *Under Assumptions E1, E2, W1 and W2,*

$$\widehat{\sigma}^2 = \frac{1}{NT} \sum_{i=1}^{n} \sum_{t=1}^{T} \left(y_{i,t}^0 - \widehat{m}^0 \left(y_{i,t-1} \right) \right)^2 \to_p \sigma^2$$

as $N, T \to \infty$.

Lemma A1.10 *For each K, we let*

$$\widehat{\Phi}_K = \sum_{j=0}^{J} \frac{w \left(j, J \right)}{N \left(T - j \right)} \sum_{i=1}^{n} \sum_{t=1}^{T-j} g_K \left(y_{i,t+j} \right) \widehat{u}_{i,t}^0,$$

where $\widehat{u}_{i,t}^0 = y_{i,t}^0 - \widehat{m}^0 \left(y_{i,t-1} \right)$. If we assume $\sum_{j=1}^{J} \left| w \left(j, J \right) \right| \leq C_w J$ for some constant $0 < C_w < \infty$, where $J = J \left(T \right) \leq O \left(T^{1/3} \right)$, then as $N, T \to \infty$, $\left\| \widehat{\Phi}_K - \Phi_K \right\| \to_p 0$ under Assumptions E1, E2, W1, W2 and NT. Recall that $\Phi_K = \sum_{j=0}^{\infty} cov \left(g_K \left(y_{i,t+j} \right), u_{i,t} \right)$, where $\| \Phi_K \| < \infty$ for each K.

A.2 Proofs of Lemmas in A.1

Proof of Lemma A1.1 By the stationarity over t and independence across i,

$$\begin{aligned} & \mathbb{E} \left\| \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \underline{g}_K \left(y_{i,t-1} \right) \underline{g}_K \left(y_{i,t-1} \right)' - \Gamma_K \right\|^2 \\ = & \sum_{j=1}^{K} \sum_{k=1}^{K} \mathbb{E} \left(\frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \underline{g}_{Kj} \left(y_{i,t-1} \right) \underline{g}_{Kk} \left(y_{i,t-1} \right) - \Gamma_{K,jk} \right)^2 \\ = & \frac{1}{NT} \sum_{j=1}^{K} \sum_{k=1}^{K} \mathbb{E} \left[\underline{g}_{Kj} \left(y_{i,0} \right) \underline{g}_{Kk} \left(y_{i,0} \right) - \Gamma_{K,jk} \right]^2 \\ & + \frac{2}{NT} \sum_{j=1}^{K} \sum_{k=1}^{K} \sum_{\tau=1}^{T-1} \left(1 - \frac{\tau}{T} \right) cov \left(\underline{g}_{Kj} \left(y_{i,0} \right) \underline{g}_{Kk} \left(y_{i,0} \right), \underline{g}_{Kj} \left(y_{i,\tau} \right) \underline{g}_{Kk} \left(y_{i,\tau} \right) \right) \\ \equiv & A_1 \left(N, T, K \right) + A_2 \left(N, T, K \right), \end{aligned}$$

where $\Gamma_{K,jk}$ is the (j,k)th element of the $K \times K$ matrix Γ_K. Note that conditional on μ_i, the stationarity and the mixing property of $\{y_{i,t}\}$ are preserved to $\left\{ \underline{g}_{Kk} \left(y_{i,t} \right) \right\}$ for all k and t by Proposition 2.3 because $g_{Kk} \left(\cdot \right)$ are all measurable functions and the common level shift by its mean does not affect the dependence structure. First note that $\mathbb{E}\underline{g}_{Kj}(y_{i,t})\underline{g}_{Kk}(y_{i,t}) = \Gamma_{K,jk}$ implies[21]

$$\begin{aligned} A_1(N,T,K) &\leq \frac{1}{NT}\sum_{j=1}^{K}\sum_{k=1}^{K}\mathbb{E}\underline{g}_{Kj}^2(y_{i,0})\,\underline{g}_{Kk}^2(y_{i,0}) \\ &= \frac{1}{NT}\mathbb{E}\left(\sum_{j=1}^{K}\underline{g}_{Kj}^2(y_{i,0})\sum_{k=1}^{K}\underline{g}_{Kk}^2(y_{i,0})\right) \\ &\leq \zeta_0^4(K)\,K^2/NT \to 0 \end{aligned}$$

by Assumption W1. Secondly, using Proposition 2.4, under Assumptions E1, E2 and W1[22],

$$\left|cov\left(\underline{g}_{Kj}(y_{i,0})\,\underline{g}_{Kk}(y_{i,0}),\underline{g}_{Kj}(y_{i,\tau})\,\underline{g}_{Kk}(y_{i,\tau})\right)\right| \leq 4\alpha(\tau)\,\zeta_0^4(K)$$

because the boundedness condition $\sup_{y\in\mathcal{Y}_c}\max_{1\leq k\leq K}\left|\underline{g}_{Kk}(y)\right| \leq \zeta_0(K)$ implies $\left|\underline{g}_{Kj}(y)\,\underline{g}_{Kk}(y)\right| \leq \zeta_0^2(K)$ for all j and k. Since we assume $\sum_{\tau\geq 1}\alpha(\tau) < \infty$, we have

$$\begin{aligned} \left|\sum_{\tau=1}^{T-1}\left(1-\frac{\tau}{T}\right)cov\left(\underline{g}_{Kj}(y_{i,0})\,\underline{g}_{Kk}(y_{i,0}),\underline{g}_{Kj}(y_{i,\tau})\,\underline{g}_{Kk}(y_{i,\tau})\right)\right| &\leq 4\zeta_0^4(K)\sum_{\tau=1}^{T-1}\left(1-\frac{\tau}{T}\right)\alpha(\tau) \\ &\leq 4\zeta_0^4(K)\sum_{\tau=1}^{\infty}\alpha(\tau) \end{aligned}$$

using the Kronecker lemma[23]. Therefore,

$$|A_2(N,T,K)| \leq O\left(\zeta_0^4(K)\,K^2/NT\right) \to 0$$

by Assumption W1. It follows that

$$\left\|\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\underline{g}_K(y_{i,t-1})\,\underline{g}_K(y_{i,t-1})' - \Gamma_K\right\| = O_p\left(\zeta_0^2(K)\,K/\sqrt{NT}\right),$$

which is $o_p(1)$ since $\zeta_0^4(K)\,K^2/NT \to 0$ is assumed. ■

Proof of Lemma A1.2 Similarly as Lemma A1.1, we first observe that

$$\begin{aligned} \mathbb{E}\left\|\frac{1}{NT^2}\sum_{i=1}^{N}\sum_{t=1}^{T}\underline{g}_K(y_{i,t-1})\sum_{s=1}^{T}\underline{g}_K(y_{i,s-1})'\right\|^2 &= \frac{1}{NT^4}\sum_{j=1}^{K}\sum_{k=1}^{K}\mathbb{E}\left(\sum_{t=1}^{T}\underline{g}_{Kj}(y_{i,t-1})\sum_{s=1}^{T}\underline{g}_{Kk}(y_{i,t-1})\right)^2 \\ &\leq \frac{\zeta_0^2(K)\,K}{NT^2}\sum_{k=1}^{K}\mathbb{E}\left(\sum_{t=1}^{T}\underline{g}_{Kk}(y_{i,t-1})\right)^2. \end{aligned}$$

[21] Similarly as in Newey (1997), we can derive the sharper upper bound $\zeta_0^2(K)\,K^2/nT$ by assuming $\Gamma_K = I_K$. Letting Γ be the identity matrix does not lose any generality as argued in Newey (1997) since we assume the smallest eigenvalue of Γ_K is bounded above zero. We, however, do not pursue this sharper bound since the covariance term, which is new, cannot achive this sharper bound.

[22] Recall that the mixing inequality should hold conditional on μ_i. However, using law of iterated expectation yields that for each i

$$\begin{aligned} \left|cov\left(\underline{g}_{Kj}(y_{i,0})\,\underline{g}_{Kk}(y_{i,0}),\underline{g}_{Kj}(y_{i,\tau})\,\underline{g}_{Kk}(y_{i,\tau})\right)\right| &\leq \mathbb{E}\left|cov\left(\underline{g}_{Kj}(y_{i,0})\,\underline{g}_{Kk}(y_{i,0}),\underline{g}_{Kj}(y_{i,\tau})\,\underline{g}_{Kk}(y_{i,\tau})\,|\mu_i\right)\right| \\ &\leq \mathbb{E}4\alpha(\tau)\,\zeta_0^4(K) = 4\alpha(\tau)\,\zeta_0^4(K) \end{aligned}$$

since nothing is random any longer. The upper bound obviously is not a function of μ_i, and therefore, the result holds without conditioning on μ_i. I will use this logic in what follows.

[23] If $\sum_{\tau=1}^{T}\alpha(\tau)$ converges, then $(1/T)\sum_{\tau=1}^{T}\tau\alpha(\tau) \to 0$ as $T \to \infty$.

Note that $\mathbb{E}\underline{g}_{Kk}(y_{i,t-1}) = 0$ implies

$$\begin{aligned}
\frac{1}{T}\mathbb{E}\left(\sum_{t=1}^{T}\underline{g}_{Kk}(y_{i,t-1})\right)^2 &\leq \mathbb{E}\underline{g}_{Kk}^2(y_{i,0}) + 2\sum_{\tau=1}^{T-1}\left(1-\frac{\tau}{T}\right)\left|cov\left(\underline{g}_{Kk}(y_{i,0}), \underline{g}_{Kk}(y_{i,\tau})\right)\right| \\
&\leq \zeta_0^2(K) + 8\sum_{\tau=1}^{\infty}\alpha(\tau)\zeta_0^2(K) \\
&= O\left(\zeta_0^2(K)\right) \qquad \text{(a1)}
\end{aligned}$$

similarly as in the proof of Lemma A1.1. Therefore,

$$\mathbb{E}\left\|\frac{1}{NT^2}\sum_{i=1}^{N}\sum_{t=1}^{T}\underline{g}_K(y_{i,t-1})\sum_{s=1}^{T}\underline{g}_K(y_{i,s-1})'\right\|^2 \leq O\left(\zeta_0^4(K)K^2/NT\right) \to 0$$

and it follows that $\left\|(1/NT^2)\sum_{i=1}^{N}\sum_{t=1}^{T}\underline{g}_K(y_{i,t-1})\sum_{s=1}^{T}\underline{g}_K(y_{i,s-1})'\right\| = O_p\left(\zeta_0^2(K)K/\sqrt{NT}\right) = o_p(1)$. ■

Proof of Lemma A1.3 We decompose

$$\sum_{i=1}^{N}\sum_{t=1}^{T}g^0(y_{i,t-1})g^0(y_{i,t-1})' = \sum_{i=1}^{N}\sum_{t=1}^{T}\overline{g}(y_{i,t-1})\overline{g}(y_{i,t-1})' - \frac{1}{T}\sum_{i=1}^{N}\sum_{t=1}^{T}\overline{g}(y_{i,t-1})\sum_{s=1}^{T}\overline{g}(y_{i,s-1})'.$$

Then the first result is easily derived from Lemma A1.1 and A1.2. For the second result, note that

$$\left\|\widehat{\Gamma}_K^{-1}\right\| \leq \left\|\Gamma_K^{-1}\right\| + \left\|\widehat{\Gamma}_K^{-1} - \Gamma_K^{-1}\right\|. \qquad \text{(a2)}$$

With the similar argument of Lewis and Reinsel (1985, Theorem 1) and (Berk, 1974), the first term $\left\|\Gamma_K^{-1}\right\|$ is uniformly bounded over K since the smallest eigenvalue is bounded away from zero and the largest eigenvalue is also bounded above uniformly in K (Assumption W1-(i)). The second term converges to zero in probability if $\zeta_0^4(K)K^2/NT \to 0$. This is because

$$\left\|\widehat{\Gamma}_K^{-1} - \Gamma_K^{-1}\right\| \leq \left\|\widehat{\Gamma}_K^{-1}\right\|\left\|\widehat{\Gamma}_K - \Gamma_K\right\|\left\|\Gamma_K^{-1}\right\| \leq \left(\left\|\Gamma_K^{-1}\right\| + \left\|\widehat{\Gamma}_K^{-1} - \Gamma_K^{-1}\right\|\right)\left\|\widehat{\Gamma}_K - \Gamma_K\right\|\left\|\Gamma_K^{-1}\right\|$$

from (a2), which implies that

$$\begin{aligned}
\left\|\widehat{\Gamma}_K^{-1} - \Gamma_K^{-1}\right\| &\leq \left\|\Gamma_K^{-1}\right\|\left\|\widehat{\Gamma}_K - \Gamma_K\right\|\left\|\Gamma_K^{-1}\right\|\left(1 - \left\|\widehat{\Gamma}_K - \Gamma_K\right\|\left\|\Gamma_K^{-1}\right\|\right)^{-1} \\
&= \left\|\Gamma_K^{-1}\right\|\left\|\widehat{\Gamma}_K - \Gamma_K\right\|\left\|\Gamma_K^{-1}\right\|\left\{1 + \left\|\widehat{\Gamma}_K - \Gamma_K\right\|\left\|\Gamma_K^{-1}\right\| + O\left(\left\|\widehat{\Gamma}_K - \Gamma_K\right\|^2\left\|\Gamma_K^{-1}\right\|^2\right)\right\} \\
&\leq O_p\left(\zeta_0^2(K)K/\sqrt{NT}\right) \to 0 \qquad \text{(a3)}
\end{aligned}$$

by Taylor expansion and using the first result $\left\|\widehat{\Gamma}_K - \Gamma_K\right\| = O_p\left(\zeta^2(K)K/\sqrt{NT}\right)$. Recall that condition that $\zeta_0^4(K)K^2/NT \to 0$. ■

Proof of Lemma A1.4 First note that

$$\mathbb{E}\left\|\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\underline{g}_K(y_{i,t-1})u_{i,t}\right\|^2 = \frac{1}{NT^2}\sum_{k=1}^{K}\mathbb{E}\left(\sum_{t=1}^{T}\underline{g}_{Kk}(y_{i,t-1})u_{i,t}\right)^2,$$

where

$$\frac{1}{T}\mathbb{E}\left(\sum_{t=1}^{T}\underline{g}_{Kk}(y_{i,t-1})u_{i,t}\right)^2 \leq \mathbb{E}\left(\underline{g}_{Kk}(y_{i,0})u_{i,1}\right)^2 + 2\sum_{\tau=1}^{T-1}\left(1-\frac{\tau}{T}\right)\left|cov\left(g_{Kk}(y_{i,0})u_{i,1}, g_{Kk}(y_{i,\tau})u_{i,1+\tau}\right)\right|.$$

The first term is simply $O\left(\zeta_0^2(K)\right)$ since

$$\mathbb{E}\left(\underline{g}_{Kk}(y_{i,0})u_{i,1}\right)^2 = \mathbb{E}\underline{g}_{Kk}^2(y_{i,0})\mathbb{E}\left(u_{i,1}^2|y_{i,0}\right) \leq C\zeta_0^2(K)$$

for some constant $C > 0$ by the law of iterated expectation and Assumptions E1, E2 and W1. For the second term, since $\left\{\underline{g}_{Kk}(y_{i,t})\right\}$ is α-mixing with mixing coefficient $\alpha_i(\tau)$ for each k; and $\{u_{i,t}\}$ is *i.i.d.*, the pair of sequences $\left\{\left(\underline{g}_{Kk}(y_{i,t-1}), u_{i,t}\right)\right\}$ is also α-mixing with the same mixing coefficient $\alpha_i(\tau)$ for each i. It thus follows that the sequence of $\left\{\underline{g}_{Kk}(y_{i,t-1})\, u_{i,t}\right\}$ is also α-mixing with the same mixing coefficient $\alpha_i(\tau)$ since $\underline{g}_k(y_{i,t-1})$ and $u_{i,t}$ are independent for all i and t. Moreover, for some $\nu > 2$, $\mathbb{E}\left|\underline{g}_{Kk}(y_{i,t-1})\, u_{i,t}\right|^{\nu} \leq \zeta_0^{\nu}(K)\,\mathbb{E}|u_{i,t}|^{\nu}$, we have

$$\begin{aligned}
\left|cov\left(g_{Kk}(y_{i,0})\, u_{i,1}, g_{Kk}(y_{i,\tau})\, u_{i,1+\tau}\right)\right| &\leq \mathbb{E}\left|cov\left(g_{Kk}(y_{i,0})\, u_{i,1}, g_{Kk}(y_{i,\tau})\, u_{i,1+\tau}|\mu_i\right)\right| \\
&\leq \mathbb{E}\left[8\alpha(\tau)^{1-2/\nu}\zeta_0^2(K)\left(\mathbb{E}|u_{i,t}|^{\nu}|\mu_i\right)^{2/\nu}\right] \\
&= 8\alpha(\tau)^{1-2/\nu}\zeta_0^2(K)\left(\mathbb{E}|u_{i,t}|^{\nu}\right)^{2/\nu}
\end{aligned}$$

since $u_{i,t}$ is independent of μ_i; and therefore,

$$\frac{1}{T}\mathbb{E}\left(\sum_{t=1}^{T}\underline{g}_{Kk}(y_{i,t-1})\, u_{i,t}\right)^2 \leq C\zeta_0^2(K) + 8\sigma^2\zeta_0^2(K)\left(\mathbb{E}|u_{i,t}|^{\nu}\right)^{2/\nu}\sum_{\tau=1}^{\infty}\alpha(\tau)^{1-2/\nu}$$

and it follows that

$$\mathbb{E}\left\|\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\underline{g}_K(y_{i,t-1})\, u_{i,t}\right\|^2 \leq O\left(\zeta_0^2(K)\, K/NT\right) \to 0$$

since $\sum_{\tau=1}^{\infty}\alpha(\tau)^{1-2/\nu} < \infty$, $\mathbb{E}|u_{i,t}|^{\nu} < \infty$ for $\nu > 4 > 2$ by assumption and $\zeta_0^2(K)\, K/NT \leq \zeta_0^4(K)\, K^2/NT \to 0$.

For the second result, we observe

$$\begin{aligned}
\mathbb{E}\left\|\frac{1}{NT^2}\sum_{i=1}^{N}\sum_{t=1}^{T}\underline{g}_K(y_{i,t-1})\sum_{s=1}^{T}u_{i,s}\right\|^2 &= \frac{1}{NT^4}\sum_{k=1}^{K}\mathbb{E}\left(\sum_{t=1}^{T}\underline{g}_{Kk}(y_{i,t-1})\sum_{s=1}^{T}u_{i,s}\right)^2 \\
&\leq \frac{1}{NT^4}\sum_{k=1}^{K}\mathbb{E}\left(T\zeta_0(K)\sum_{t=1}^{T}u_{i,t}\right)^2 \\
&= \sigma^2\zeta_0^2(K)\, K/NT = O\left(\zeta_0^2(K)\, K/NT\right) \to 0. \quad \blacksquare
\end{aligned}$$

Proof of Lemma A1.5 Note that by Assumption W2,

$$\begin{aligned}
\mathbb{E}\left\|\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\underline{g}_K(y_{i,t-1})\left\{m(y_{i,t-1}) - g_K(y_{i,t-1})'\theta_K\right\}\right\|^2 &= \frac{1}{NT^2}\sum_{k=1}^{K}\mathbb{E}\left(\sum_{t=1}^{T}\underline{g}_{Kk}(y_{i,t-1})\left\{m(y_{i,t-1}) - g_K(y_{i,t-1})'\theta_K\right\}\right)^2 \\
&\leq \frac{1}{NT^2}\sum_{k=1}^{K}\mathbb{E}\left(\sum_{t=1}^{T}\underline{g}_{Kk}(y_{i,t-1})\, C_m K^{-\delta}\right)^2 \\
&\leq O\left(\zeta_0^2(K)\, K^{1-2\delta}/NT\right) \to 0
\end{aligned}$$

because $(1/T)\,\mathbb{E}\left(\sum_{t=1}^{T}\underline{g}_{Kk}(y_{i,t-1})\right)^2 \leq O\left(\zeta_0^2(K)\right)$ as shown in (a1), and $\zeta_0^2(K)\, K^{1-2\delta}/NT \leq \zeta_0^4(K)\, K^2/NT \to 0$ for some $\delta > 0$.

The second result follows similarly since

$$\begin{aligned}
&\mathbb{E}\left\|\frac{1}{NT^2}\sum_{i=1}^{N}\sum_{t=1}^{T}\underline{g}_K(y_{i,t-1})\sum_{s=1}^{T}\left\{m(y_{i,s-1}) - g_K(y_{i,s-1})'\theta_K\right\}\right\|^2 \\
= {}& \frac{1}{NT^4}\sum_{k=1}^{K}\mathbb{E}\left(\sum_{t=1}^{T}\underline{g}_{Kk}(y_{i,t-1})\sum_{s=1}^{T}\left\{m(y_{i,s-1}) - g_K(y_{i,s-1})'\theta_K\right\}\right)^2 \\
\leq {}& \frac{1}{NT^4}\sum_{k=1}^{K}\mathbb{E}\left(\sum_{t=1}^{T}\underline{g}_{Kk}(y_{i,t-1})\, TC_m K^{-\delta}\right)^2 \leq O\left(\zeta_0^2(K)\, K^{1-2\delta}/NT\right). \quad \blacksquare
\end{aligned}$$

Proof of Lemma A1.6 We first let a random variable $Z_{i,t} = \rho' \Gamma_K^{-1/2} \underline{g}_K (y_{i,t-1}) u_{i,t}/\sigma$, then $Z_{i,t}$ is a martingale difference sequence with variance one by construction. Moreover, conditioning on μ_i, $Z_{i,t}$ is α-mixing with the same mixing coefficients $\alpha(\tau)$ of $\{y_{i,t}\}$ since the temporal dependence is solely determined by $\underline{g}_K (y_{i,t-1})$ whereas $u_{i,t}$ is independent. Also note that $|Z_{i,t}| \leq \|\rho\| \left\|\Gamma_K^{-1/2}\right\| \|\underline{g}(y_{i,t-1})\| |u_{i,t}| \leq C_1 K^{1/2} \zeta_0 (K) |u_{i,t}|$ for some constant $C_1 > 0$ since $\|\rho\| = 1$ and by Assumption W1. Thus, for some $r = \nu/2 > 2$, $\mathbb{E}\left|Z_{i,t}^2\right|^r = \mathbb{E}|Z_{i,t}|^{2r} \leq C_2 K^r \zeta_0^{2r}(K) \mathbb{E}|u_{i,t}|^{2r} = O\left(K^r \zeta_0^{2r}(K)\right)$ for some constant $C_2 > 0$ since $\mathbb{E}|u_{i,t}|^{2r} < \infty$ from Assumption E1. Then, similarly as Lemma A1.1, we have $(1/NT)\sum_{i=1}^N \sum_{t=1}^T Z_{i,t}^2 \to_p 1$ because

$$\begin{aligned}
\mathbb{E}\left\|\frac{1}{NT}\sum_{i=1}^N\sum_{t=1}^T Z_{i,t}^2 - 1\right\|^2 &\leq \frac{1}{NT}\left\{\mathbb{E}\left(Z_{i,1}^2 - 1\right)^2 + 2\sum_{\tau=1}^{T-1}\left(1 - \frac{\tau}{T}\right)\left|cov\left(Z_{i,1}^2, Z_{i,\tau+1}^2\right)\right|\right\} \\
&\leq \frac{1}{NT}\left\{\mathbb{E}\left|Z_{i,t}^2\right|^2 + 16\sum_{\tau=1}^{\infty}\alpha(\tau)^{1-2/r}\left(\mathbb{E}\left|Z_{i,t}^2\right|^r\right)^{2/r}\right\} \\
&\leq O\left(K^2\zeta_0^4(K)/NT\right)
\end{aligned}$$

using Proposition 2.4-(2) with $p = q = r > 2$ and $\sum_{\tau=1}^\infty \alpha(\tau)^{1-2/r} < \infty$ by Assumption E1 and E2. Note that the inequality holds without conditioning on μ_i since

$$\begin{aligned}
\left|cov\left(Z_{i,1}^2, Z_{i,\tau+1}^2\right)\right| &\leq \mathbb{E}\left|cov\left(Z_{i,1}^2, Z_{i,\tau+1}^2|\mu_i\right)\right| \leq 8\alpha(\tau)^{1-2/r}\left(\mathbb{E}\left[\left|Z_{i,t}^2\right|^r|\mu_i\right]\right)^{2/r} \\
&\leq 8\alpha(\tau)^{1-2/r}\left(C_2K^r\zeta_0^{2r}(K)\mathbb{E}\left[\mathbb{E}|u_{i,t}|^{2r}|\mu_i\right]\right)^{2/r} \\
&= 8\alpha(\tau)^{1-2/r}\left(C_2K^r\zeta_0^{2r}(K)\mathbb{E}|u_{i,t}|^{2r}\right)^{2/r}
\end{aligned}$$

since $u_{i,t}$ is independent of μ_i similarly as in the proof of Lemma A1.1.

Directly applying the conventional Lindeberg condition as in Theorem 5.23 of White (1984) to the double indexed process $Z_{i,t}$ is not straightforward. Phillips and Moon (1999) develop limit theories for large N and T and examine Lindeberg condition for the Central Limit Theorem of double indexed processes (Theorem 2 and 3), and we adopt their idea to derive the asymptotic normality of $\{Z_{i,t}\}$ as follows[24]. We first define a partial sum process $Z_t = \left(1/\sqrt{N}\right)\sum_{i=1}^N Z_{i,t}$, where $Z_{i,t}$ is *i.i.d.* across i. Then, for any $\epsilon_1, \epsilon_2 > 0$, if we apply Cauchy-Schwartz and Chebyshev's inequalities in turn,

$$\begin{aligned}
\mathbb{E}\left(Z_t^2 \mathbf{1}\left\{|Z_t| > \epsilon_1\sqrt{T}\right\}\right) &= \mathbb{E}\left(Z_t^2\mathbf{1}\left\{Z_t^2 > \epsilon_1^2 T\right\}\right) \\
&\leq \mathbb{E}\left(Z_{i,t}^2\mathbf{1}\left\{Z_{i,t}^2 > NT\epsilon_2\right\}\right) \\
&\leq \left[\mathbb{E}\left(Z_{i,t}^4\right)\right]^{1/2}\left[\mathbb{P}\left\{Z_{i,t}^2 > NT\epsilon_2\right\}\right]^{1/2} \\
&\leq \left[\mathbb{E}\left(Z_{i,t}^4\right)\right]^{1/2}\left[\mathbb{E}\left(Z_{i,t}^4\right)/(NT\epsilon_2)^2\right]^{1/2} \\
&= \mathbb{E}\left|Z_{i,t}^2\right|^2/(NT\epsilon_2) = C_3K^2\zeta_0^4(K)\mathbb{E}|u_{i,t}|^4/NT \to 0
\end{aligned}$$

for some constant $C_3 > 0$, where $\mathbb{E}|u_{i,t}|^4 < \infty$ and $\mathbf{1}\{\cdot\}$ is the binary indicator function. It then follows by Theorem 5.23 of White (1984) that $\left(1/\sqrt{T}\right)\sum_{t=1}^T Z_t = \left(1/\sqrt{NT}\right)\sum_{i=1}^N\sum_{t=1}^T Z_{i,t} \to_d \mathcal{N}(0,1)$ as $N, T \to \infty$. Therefore,

$$\frac{1}{\sqrt{NT}}\sum_{i=1}^N\sum_{t=1}^T \rho'\Gamma_K^{-1/2}\underline{g}_K(y_{i,t-1})u_{i,t} \to_d \mathcal{N}\left(0,\sigma^2\right)$$

as $N, T \to \infty$. ∎

Proof of Lemma A1.7 Note that

$$\mathbb{E}\left\|\frac{1}{\sqrt{NT^3}}\sum_{i=1}^N\sum_{t=1}^T \underline{g}_K(y_{i,t-1})\sum_{s=1}^T u_{i,s} - \sqrt{\kappa}\Phi_K\right\|^2 \leq \frac{N}{T}\mathbb{E}\left\|\frac{1}{NT}\sum_{i=1}^N\sum_{t=1}^T \underline{g}_K(y_{i,t-1})\sum_{s=1}^T u_{i,s} - \Phi_K\right\|^2 + \left(\sqrt{\frac{N}{T}} - \sqrt{\kappa}\right)^2\|\Phi_K\|^2.$$

[24] Alternatively, we can directly apply Theorem 3 of Phillips and Moon (1999) since we already show $\mathbb{E}\left\|\frac{1}{NT}\sum_{i=1}^N\sum_{t=1}^T Z_{i,t}^2 - 1\right\|^2 \leq O\left(K^2\zeta_0^4(K)/NT\right) = o(1)$.

The second part is negligible for large N and T since $\lim_{N,T\to\infty} N/T = \kappa$ and $\|\Phi_K\| < \infty$ for each K. For the first part, $N/T \to \kappa < \infty$ and the following argument shows that $\mathbb{E}\left\|\frac{1}{NT}\sum_{i=1}^N\sum_{t=1}^T \underline{g}_K(y_{i,t-1})\sum_{s=1}^T u_{i,s} - \Phi_K\right\|^2$ is negligible for large N and T. We observe that

$$\begin{aligned}
\mathbb{E}\left(\frac{1}{NT}\sum_{i=1}^N\sum_{t=1}^T \underline{g}_K(y_{i,t-1})\sum_{s=1}^T u_{i,s}\right) &= \frac{1}{T}\sum_{t=1}^T \mathbb{E}\left(\underline{g}_K(y_{i,t-1})\sum_{s=1}^T u_{i,s}\right)\\
&= \frac{1}{T}\left\{\sum_{t=2}^T \mathbb{E}\left(\underline{g}_K(y_{i,t-1})\sum_{s=1}^{t-1} u_{i,s}\right) + \sum_{t=1}^T \mathbb{E}\left(\underline{g}_K(y_{i,t-1})\sum_{s=t}^T u_{i,s}\right)\right\}\\
&= \frac{1}{T}\sum_{t=2}^T \mathbb{E}\left(\underline{g}_K(y_{i,t-1})\sum_{s=1}^{t-1} u_{i,s}\right)\\
&= \frac{1}{T}\sum_{t=2}^T\sum_{j=1}^{t-1} \mathbb{E}\underline{g}_K(y_{i,t-j})\,u_{i,1}\\
&= \sum_{j=1}^{T-1}(1-j/T)\,cov\left(\underline{g}(y_{i,j}), u_{i,1}\right)
\end{aligned}$$

by the stationarity. Therefore,

$$\begin{aligned}
&\mathbb{E}\left\|\frac{1}{NT}\sum_{i=1}^N\sum_{t=1}^T \underline{g}_K(y_{i,t-1})\sum_{s=1}^T u_{i,s} - \Phi_K\right\|^2\\
\leq\ &\mathbb{E}\left\|\frac{1}{NT}\sum_{i=1}^N\sum_{t=1}^T \underline{g}_K(y_{i,t-1})\sum_{s=1}^T u_{i,s} - \mathbb{E}\left(\frac{1}{NT}\sum_{i=1}^N\sum_{t=1}^T \underline{g}_K(y_{i,t-1})\sum_{s=1}^T u_{i,s}\right)\right\|^2\\
&+\left\|\sum_{j=1}^{T-1}(1-j/T)\,cov\left(g_K(y_{i,j}), u_{i,1}\right) - \Phi_K\right\|^2\\
\equiv\ &B_1(N,T,K) + B_2(N,T,K).
\end{aligned}$$

By the Kronecker lemma, $B_2(N,T,K)$ is negligible for large T since $\Phi_K = \sum_{j=1}^\infty cov\left(g_K(y_{i,t-1}), u_{i,t-j}\right)$. Moreover, if we let a $K\times 1$ vector $\Psi_K \equiv \mathbb{E}\left((1/NT)\sum_{i=1}^N\sum_{t=1}^T \underline{g}_K(y_{i,t-1})\sum_{s=1}^T u_{i,s}\right)$ and its kth element as Ψ_{Kk}, then we have

$$\begin{aligned}
B_1(N,T,K) &= \frac{1}{NT^2}\sum_{k=1}^K \mathbb{E}\left(\sum_{t=1}^T\left[\underline{g}_{Kk}(y_{i,t-1})\sum_{s=1}^T u_{i,s} - \Psi_{Kk}\right]\right)^2\\
&= \frac{1}{NT^2}\sum_{k=1}^K\sum_{t=1}^T \mathbb{E}\left(\underline{g}_{Kk}(y_{i,t-1})\sum_{s=1}^T u_{i,s} - \Psi_{Kk}\right)^2\\
&+\frac{1}{NT^2}\sum_{k=1}^K\sum_{t=1}^{T-1}\sum_{\tau=1}^{T-t-1} \mathbb{E}\left(\underline{g}_{Kk}(y_{i,t-1})\,\underline{g}_{Kk}(y_{i,t-1+\tau})\left(\sum_{s=1}^T u_{i,s}\right)^2\right)\\
&+\frac{1}{NT^2}\sum_{k=1}^K\sum_{t=2}^T\sum_{\tau=1}^{t-1} \mathbb{E}\left(\underline{g}_{Kk}(y_{i,t-1-\tau})\,\underline{g}_{Kk}(y_{i,t-1})\left(\sum_{s=1}^T u_{i,s}\right)^2\right)
\end{aligned}$$

Note that

$$\sum_{t=1}^T \mathbb{E}\left(\underline{g}_{Kk}(y_{i,t-1})\sum_{s=1}^T u_{i,s} - \Psi_{Kk}\right)^2 \leq \sum_{t=1}^T \mathbb{E}\left(\underline{g}_{Kk}(y_{i,t-1})\sum_{s=1}^T u_{i,s}\right)^2 \leq T^2\sigma^2\zeta_0^2(K);$$

and by Assumption E1 and by Cauchy-Schwartz inequality

$$\begin{aligned}
\left|\mathbb{E}\left(\underline{g}_{Kk}\left(y_{i,t-1}\right)\underline{g}_{Kk}\left(y_{i,t-1+\tau}\right)\left(\sum_{s=1}^{T}u_{i,s}\right)^2\right)\right| &\leq \left[\mathbb{E}\left(\underline{g}_{Kk}^2\left(y_{i,t-1}\right)\underline{g}_{Kk}^2\left(y_{i,t-1+\tau}\right)\right)\right]^{1/2}\left[\mathbb{E}\left(\sum_{s=1}^{T}u_{i,s}\right)^4\right]^{1/2} \\
&\leq \left[4\alpha\left(\tau\right)\zeta_0^4\left(K\right)\right]^{1/2}\left[T\mathbb{E}\left|u_{i,t}\right|^4+3T\left(T-1\right)\sigma^4\right]^{1/2} \\
&= C_1\alpha\left(\tau\right)^{1/2}\zeta_0^2\left(K\right)T,
\end{aligned}$$

and similarly

$$\left|\mathbb{E}\left(\underline{g}_{Kk}\left(y_{i,t-1-\tau}\right)\underline{g}_{Kk}\left(y_{i,t-1}\right)\left(\sum_{s=1}^{T}u_{i,s}\right)^2\right)\right| \leq C_2\alpha\left(\tau\right)^{1/2}\zeta_0^2\left(K\right)T,$$

where C_1 and C_2 are some positive constants; $u_{i,t}$ is *i.i.d.* with $\mathbb{E}\left|u_{i,t}\right|^4<\infty$. It follows that

$$\begin{aligned}
B_1\left(N,T,K\right) &= \frac{1}{NT^2}\sum_{k=1}^{K}T^2\sigma^2\zeta_0^2\left(K\right)+\frac{1}{NT^2}\sum_{k=1}^{K}\sum_{t=1}^{T-1}\sum_{\tau=1}^{T-t-1}C_1\alpha\left(\tau\right)^{1/2}\zeta_0^2\left(K\right)T+\frac{1}{NT^2}\sum_{k=1}^{K}\sum_{t=2}^{T}\sum_{\tau=1}^{t-1}C_2\alpha\left(\tau\right)^{1/2}\zeta_0^2\left(K\right)T \\
&\leq O\left(\zeta_0^2\left(K\right)K/N\right),
\end{aligned}$$

and therefore, $\mathbb{E}\left\|\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\underline{g}_K\left(y_{i,t-1}\right)\sum_{s=1}^{T}u_{i,s}-\Phi_K\right\|^2=o\left(1\right)$ since the assumption $\lim_{N,T\to\infty}N/T=\kappa$ with $0<\kappa<\infty$ implies that $\left(\zeta_0^2\left(K\right)K/N\right)^2=\left(\zeta_0^4\left(K\right)K^2/NT\right)\left(T/N\right)\to 0$ as $N,T\to\infty$ by Assumption W1. The result is then following since $\lim_{N,T\to\infty}N/T=\kappa<\infty$ implies $O\left(1/N\right)=O\left(1/\sqrt{NT}\right)$. ■

Proof of Lemma A1.8 Note that Assumption W2 implies

$$\left\|\frac{1}{\sqrt{NT}}\sum_{i=1}^{N}\sum_{t=1}^{T}g_K\left(y_{i,t-1}\right)\left\{m\left(y_{i,t-1}\right)-g_K\left(y_{i,t-1}\right)'\theta_K\right\}\right\| \leq \left\|\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}g_K\left(y_{i,t-1}\right)\right\|\left(C_mK^{-\delta}\sqrt{NT}\right),$$

and by the ergodic theorem for α-mixing process (e.g., see White (1984))

$$\left\|\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}g_K\left(y_{i,t-1}\right)-\mathbb{E}g_K\left(y_{i,t-1}\right)\right\|\to_{a.s.}0$$

as $N,T\to\infty$ with $\left\|\mathbb{E}g_K\left(y_{i,t+j}\right)\right\|<\infty$ since $\mathbb{E}\left|g_{Kk}\left(y_{i,t-1}\right)\right|<\infty$ thanks to the boundedness $g_{Kk}\left(y\right)$ for all k. More precisely,

$$\begin{aligned}
\mathbb{E}\left\|\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}g_K\left(y_{i,t-1}\right)-\mathbb{E}g_K\left(y_{i,t-1}\right)\right\|^2 &= \frac{1}{NT^2}\sum_{k=1}^{K}\mathbb{E}\left(\sum_{t=1}^{T}g_K\left(y_{i,t-1}\right)-\mathbb{E}g_K\left(y_{i,t-1}\right)\right)^2 \\
&= \frac{K}{NT}\left\{\mathbb{E}\left(g_K\left(y_{i,0}\right)-\mathbb{E}g_K\left(y_{i,0}\right)\right)^2+2\sum_{\tau=1}^{T-1}\left(1-\tau/T\right)cov\left(g_K\left(y_{i,0}\right),g_K\left(y_{i,\tau}\right)\right)\right\} \\
&\leq O\left(\zeta_0^2\left(K\right)K/NT\right)\to 0.
\end{aligned}$$

Therefore,

$$\left\|\frac{1}{\sqrt{NT}}\sum_{i=1}^{N}\sum_{t=1}^{T}g_K\left(y_{i,t-1}\right)\left\{m\left(y_{i,t-1}\right)-g_K\left(y_{i,t-1}\right)'\theta_K\right\}\right\|\leq O_p\left(K^{-\delta}\sqrt{NT}\right)=o_p\left(1\right)$$

since we assume $K^{-\delta}\sqrt{NT}\to 0$.

The second result can be derived similarly since

$$\begin{aligned}
\left\|\frac{1}{\sqrt{NT^3}}\sum_{i=1}^{N}\sum_{t=1}^{T}g_K\left(y_{i,t-1}\right)\sum_{s=1}^{T}\left\{m\left(y_{i,s-1}\right)-g_K\left(y_{i,s-1}\right)'\theta_K\right\}\right\| &\leq \left\|\frac{1}{\sqrt{NT^3}}\sum_{i=1}^{N}\sum_{t=1}^{T}g_K\left(y_{i,t-1}\right)\right\|\left(TC_mK^{-\delta}\right) \\
&= \left\|\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}g_K\left(y_{i,t-1}\right)\right\|\left(C_mK^{-\delta}\sqrt{NT}\right). \quad ■
\end{aligned}$$

Proof of Lemma A1.9 We have

$$\begin{aligned}
\mathbb{E}\left|\widehat{\sigma}^2-\sigma^2\right|^2 &= \mathbb{E}\left|\frac{1}{NT}\sum_{i=1}^{n}\sum_{t=1}^{T}\left(y_{i,t}^0-\widehat{m}^0\left(y_{i,t-1}\right)\right)^2-\sigma^2\right|^2 \\
&\le \mathbb{E}\left|\frac{1}{NT}\sum_{i=1}^{n}\sum_{t=1}^{T}\left(y_{i,t}^0-m^0\left(y_{i,t-1}\right)\right)^2-\sigma^2\right|^2+\mathbb{E}\left|\frac{1}{NT}\sum_{i=1}^{n}\sum_{t=1}^{T}\left(\widehat{m}^0\left(y_{i,t-1}\right)-m^0\left(y_{i,t-1}\right)\right)^2\right|^2 \\
&\quad +\mathbb{E}\left|\frac{1}{NT}\sum_{i=1}^{n}\sum_{t=1}^{T}\left(y_{i,t}^0-m^0\left(y_{i,t-1}\right)\right)\left(\widehat{m}^0\left(y_{i,t-1}\right)-m^0\left(y_{i,t-1}\right)\right)\right|^2 \\
&= B_1(N,T)+B_2(N,T)+B_3(N,T).
\end{aligned}$$

We first observe that since $y_{i,t}^0-m^0\left(y_{i,t-1}\right)=u_{i,t}^0=u_{i,t}-(1/T)\sum_{s=1}^{T}u_{i,s}$,

$$B_1(N,T)\le\mathbb{E}\left|\frac{1}{NT}\sum_{i=1}^{n}\sum_{t=1}^{T}\left(u_{i,t}^2-\sigma^2\right)\right|^2+\mathbb{E}\left|\frac{1}{NT^2}\sum_{i=1}^{n}\left(\sum_{t=1}^{T}u_{i,t}\right)^2\right|^2,$$

where the first term is simply $(1/NT)\,\mathbb{E}\left(u_{i,t}^2-\sigma^2\right)^2=(1/NT)\left\{\mathbb{E}u_{i,t}^4-\sigma^4\right\}=O(1/NT)$ since $\mathbb{E}u_{i,t}^4<\infty$ from Assumption E1 and $u_{i,t}$ is *i.i.d.* with mean zero and $\mathbb{E}u_{i,t}^2=\sigma^2$. For the second term

$$\begin{aligned}
\mathbb{E}\left|\frac{1}{NT^2}\sum_{i=1}^{n}\left(\sum_{t=1}^{T}u_{i,t}\right)^2\right|^2 &= \frac{1}{NT^4}\mathbb{E}\left(\sum_{t=1}^{T}u_{i,t}\right)^4+\frac{1}{N^2T^4}\sum_{i\neq j}\mathbb{E}\left(\sum_{t=1}^{T}u_{i,t}\right)^2\mathbb{E}\left(\sum_{t=1}^{T}u_{j,t}\right)^2 \\
&= \frac{1}{NT^4}\left\{T\mathbb{E}u_{i,t}^4+\frac{T(T-1)}{2}\sigma^4\right\}+\frac{1}{N^2T^4}\left\{\frac{N(N-1)}{2}T\sigma^2\right\}^2 \\
&= O\left(1/NT^2\right)+O\left(1/T^2\right).
\end{aligned}$$

Therefore, $B_1(N,T)=o(1)$ for large N and T. Now, from Theorem 3.1, it is following that for any $y\in\mathcal{Y}_c$, $\mathbb{E}\left(\widehat{m}^0(y)-m^0(y)\right)^2\le O\left(K/NT+K^{-2\delta}+\zeta_0^2(K)\,K/NT\right)$. Therefore,

$$\left|\frac{1}{NT}\sum_{i=1}^{n}\sum_{t=1}^{T}\left(\widehat{m}^0\left(y_{i,t-1}\right)-m^0\left(y_{i,t-1}\right)\right)^2\right|\le O_p\left(\frac{K}{NT}+K^{-2\delta}+\frac{\zeta_0^2(K)\,K}{NT}\right)=o(1)$$

and $B_2(N,T)=o(1)$. Finally, if we also use the result in Theorem 3.1

$$\begin{aligned}
B_3(N,T) &= \mathbb{E}\left|\frac{1}{NT}\sum_{i=1}^{n}\sum_{t=1}^{T}u_{i,t}^0\left(\widehat{m}^0\left(y_{i,t-1}\right)-m^0\left(y_{i,t-1}\right)\right)\right|^2 \\
&\le \mathbb{E}\left|\frac{1}{NT}\sum_{i=1}^{n}\sum_{t=1}^{T}u_{i,t}^0\right|^2 O\left(\frac{K}{NT}+K^{-2\delta}+\frac{\zeta_0^2(K)\,K}{NT}\right)=o(1)
\end{aligned}$$

since $\sum_{i=1}^{n}\sum_{t=1}^{T}u_{i,t}^0=0$. ■

Proof of Lemma A1.10 We first decompose

$$\begin{aligned}
\left\|\widehat{\Phi}_K-\Phi_K\right\| &\le \left\|\sum_{j=0}^{J}\frac{w(j,J)}{N(T-j)}\sum_{i=1}^{n}\sum_{t=1}^{T-j}g_K\left(y_{i,t+j}\right)\left(\widehat{u}_{i,t}^0-u_{i,t}\right)\right\| && \text{(a4)} \\
&\quad +\left\|\sum_{j=0}^{J}\frac{w(j,J)}{N(T-j)}\sum_{i=1}^{n}\sum_{t=1}^{T-j}\left(g_K\left(y_{i,t+j}\right)u_{i,t}-\mathbb{E}g_K\left(y_{i,t+j}\right)u_{i,t}\right)\right\| && \text{(a5)} \\
&\quad +\left\|\sum_{j=0}^{J}\frac{w(j,J)}{N(T-j)}\sum_{i=1}^{n}\sum_{t=1}^{T-j}\mathbb{E}g_K\left(y_{i,t+j}\right)u_{i,t}-\sum_{j=0}^{\infty}cov\left(g_K\left(y_{i,t+j}\right),u_{i,t}\right)\right\|. && \text{(a6)}
\end{aligned}$$

The third term (a6) simply converges to zero as $J \to \infty$ using Kronecker lemma since we assume that

$$\left\| \sum_{j=0}^{\infty} cov\left(g_K\left(y_{i,t+j}\right), u_{i,t}\right) \right\| = \left\| \sum_{j=0}^{\infty} \mathbb{E}\left(g_K\left(y_{i,t+j}\right) u_{i,t}\right) \right\| = \|\Phi_K\| < \infty.$$

For the second term (a5), if we use the similar technique as in Newey and West (1987, Proof of Theorem 2), for any $\varepsilon > 0$, we have

$$\begin{aligned}
& \mathbb{P}\left(\left\| \sum_{j=0}^{J} \frac{w(j,J)}{N(T-j)} \sum_{i=1}^{n} \sum_{t=1}^{T-j} \left(g_K\left(y_{i,t+j}\right) u_{i,t} - \mathbb{E} g_K\left(y_{i,t+j}\right) u_{i,t}\right) \right\| > \varepsilon \right) \\
\leq \; & \mathbb{P}\left(\sum_{j=0}^{J} |w(j,J)| \left\| \frac{1}{N(T-j)} \sum_{i=1}^{n} \sum_{t=1}^{T-j} \left(g_K\left(y_{i,t+j}\right) u_{i,t} - \mathbb{E} g_K\left(y_{i,t+j}\right) u_{i,t}\right) \right\| > \varepsilon \right) \\
\leq \; & \sum_{j=1}^{J} \mathbb{P}\left(\left\| \frac{1}{N(T-j)} \sum_{i=1}^{n} \sum_{t=1}^{T-j} \left(g_K\left(y_{i,t+j}\right) u_{i,t} - \mathbb{E} g_K\left(y_{i,t+j}\right) u_{i,t}\right) \right\| > \frac{\varepsilon}{C_w J} \right) \\
\leq \; & \sum_{j=1}^{J} \left(C_w J/\varepsilon\right)^2 \mathbb{E} \left\| \frac{1}{N(T-j)} \sum_{i=1}^{n} \sum_{t=1}^{T-j} \left(g_K\left(y_{i,t+j}\right) u_{i,t} - \mathbb{E} g_K\left(y_{i,t+j}\right) u_{i,t}\right) \right\|^2, \qquad \text{(a7)}
\end{aligned}$$

where the third inequality is using Chebyshev's inequality. We assume $\sum_{j=1}^{J} |w(j,J)| \leq C_w J$ for some constant $0 < C_w < \infty$, However,

$$\begin{aligned}
& \mathbb{E} \left\| \frac{1}{N(T-j)} \sum_{i=1}^{n} \sum_{t=1}^{T-j} \left(g_K\left(y_{i,t+j}\right) u_{i,t} - \mathbb{E} g_K\left(y_{i,t+j}\right) u_{i,t}\right) \right\|^2 \\
= \; & \frac{1}{N(T-j)^2} \sum_{k=1}^{K} \mathbb{E} \left(\sum_{t=1}^{T-j} \left(g_{Kk}\left(y_{i,t+j}\right) u_{i,t} - \mathbb{E} g_{Kk}\left(y_{i,t+j}\right) u_{i,t}\right) \right)^2 \\
= \; & \frac{1}{N(T-j)} \sum_{k=1}^{K} \mathbb{E} \left(g_{Kk}\left(y_{i,t+j}\right) u_{i,t} - \mathbb{E} g_{Kk}\left(y_{i,t+j}\right) u_{i,t}\right)^2 \\
& + \frac{2}{N(T-j)^2} \sum_{\tau=1}^{T-j} (T-j-\tau) \left| cov\left(g_{Kk}\left(y_{i,t+j}\right) u_{i,t}; g_{Kk}\left(y_{i,t+j+\tau}\right) u_{i,t+\tau}\right) \right| \\
\leq \; & C_\Phi \zeta_0^2(K) K / N(T-j)
\end{aligned}$$

for some constant $0 < C_\Phi < \infty$ since $\mathbb{E}\left(g_{Kk}\left(y_{i,t+j}\right) u_{i,t} - \mathbb{E} g_{Kk}\left(y_{i,t+j}\right) u_{i,t}\right)^2 \leq \mathbb{E}\left(g_{Kk}\left(y_{i,t+j}\right) u_{i,t}\right)^2 \leq \zeta_0^2(K) \mathbb{E} u_{i,t}^2 = \zeta_0^2(K) \sigma^2$ and the second term is properly bounded using mixing inequality as in the proof of A1.4. In consequence, the formula in (a7) converges to zero if $J = J(T) = O\left(T^{1/3}\right)$ since

$$\begin{aligned}
\sum_{j=1}^{J} \left(\frac{C_w J}{\varepsilon}\right)^2 \mathbb{E} \left\| \frac{1}{N(T-j)} \sum_{i=1}^{n} \sum_{t=1}^{T-j} \left(g_K\left(y_{i,t+j}\right) u_{i,t} - \mathbb{E} g_K\left(y_{i,t+j}\right) u_{i,t}\right) \right\|^2 \; & \leq \; \frac{C_w^2 C_\Phi}{\varepsilon^2} \cdot \frac{\zeta_0^2(K) K}{N} \left(\sum_{j=1}^{J} \frac{J^2}{T-j} \right) \\
& \leq \; C_\varepsilon \left(\frac{\zeta_0^2(K) K}{N} \right) \left(\frac{J^3}{T} \right) \to 0
\end{aligned}$$

for some constant $0 < C_\varepsilon < \infty$, as $N, T \to \infty$ with $N/T \to \kappa \in (0, \infty)$. Note that since N and T are comparable, $\zeta_0^2(K) K/N \approx \zeta_0^2(K) K/\sqrt{NT} \to 0$ for large N and T.

Lastly, for the first term (a4), note that

$$\widehat{u}_{i,t}^0 - u_{i,t} = \left(y_{i,t}^0 - \widehat{m}^0\left(y_{i,t-1}\right)\right) - u_{i,t} = \left(m^0\left(y_{i,t-1}\right) - \widehat{m}^0\left(y_{i,t-1}\right)\right) - \left(\frac{1}{T} \sum_{s=1}^{T} u_{i,s}\right).$$

Therefore,

$$\begin{aligned}
&\left\| \sum_{j=0}^{J} \frac{w(j,J)}{N(T-j)} \sum_{i=1}^{n} \sum_{t=1}^{T-j} g_K(y_{i,t+j}) \left(\widehat{u}_{i,t}^{0} - u_{i,t} \right) \right\| \\
\leq & \left\| \sum_{j=0}^{J} \frac{w(j,J)}{N(T-j)} \sum_{i=1}^{n} \sum_{t=1}^{T-j} g_K(y_{i,t+j}) \left(m^0(y_{i,t-1}) - \widehat{m}^0(y_{i,t-1}) \right) \right\| + \left\| \sum_{j=0}^{J} \frac{w(j,J)}{NT(T-j)} \sum_{i=1}^{n} \sum_{t=1}^{T-j} g_K(y_{i,t+j}) \sum_{s=1}^{T} u_{i,s} \right\| .
\end{aligned}$$

Similarly as in the (a7), for any $\varepsilon > 0$, the first part is

$$\begin{aligned}
&\mathbb{P}\left(\left\| \sum_{j=0}^{J} \frac{w(j,J)}{N(T-j)} \sum_{i=1}^{n} \sum_{t=1}^{T-j} g_K(y_{i,t+j}) \left(m^0(y_{i,t-1}) - \widehat{m}^0(y_{i,t-1}) \right) \right\| > \varepsilon \right) \\
\leq & \sum_{j=1}^{J} \left(\frac{C_w J}{\varepsilon} \right)^2 \mathbb{E} \left\| \frac{1}{N(T-j)} \sum_{i=1}^{n} \sum_{t=1}^{T-j} g_K(y_{i,t+j}) \left(m^0(y_{i,t-1}) - \widehat{m}^0(y_{i,t-1}) \right) \right\|^2 \\
\leq & \sum_{j=1}^{J} \left(\frac{C_w J}{\varepsilon} \right)^2 \mathbb{E} \left\| \frac{1}{N(T-j)} \sum_{i=1}^{n} \sum_{t=1}^{T-j} g_K(y_{i,t+j}) \right\|^2 O\left(\frac{K}{NT} + K^{-2\delta} + \frac{\zeta_0^2(K)K}{NT} \right) \\
\leq & \sum_{j=1}^{J} \left(\frac{C_w J}{\varepsilon} \right)^2 O(1)\, O\left(\frac{K}{NT} + K^{-2\delta} + \frac{\zeta_0^2(K)K}{NT} \right) \to 0 \text{ as } J \to \infty,
\end{aligned}$$

using Theorem 3.1 and since $\left\| (1/(N(T-j))) \sum_{i=1}^{n} \sum_{t=1}^{T-j} g_K(y_{i,t+j}) - \mathbb{E} g_K(y_{i,t+j}) \right\| \to_{a.s.} 0$ with $\| \mathbb{E} g_K(y_{i,t+j}) \| < \infty$ by the Law of large numbers in mixing process as in the proof of A1.8. Because $J \leq O\left(T^{1/3}\right)$, with the similar argument as in the proof of (a5), the first part is $o(1)$. The second part also converges to zero as $J \to \infty$ since

$$\begin{aligned}
&\mathbb{P}\left(\left\| \sum_{j=0}^{J} \frac{w(j,J)}{NT(T-j)} \sum_{i=1}^{n} \sum_{t=1}^{T-j} g_K(y_{i,t+j}) \sum_{s=1}^{T} u_{i,s} \right\| > \varepsilon \right) \\
\leq & \sum_{j=1}^{J} \left(\frac{C_w J}{\varepsilon} \right)^2 \mathbb{E} \left\| \sum_{j=0}^{J} \frac{1}{N(T-j)} \sum_{i=1}^{n} \sum_{t=1}^{T-j} g_K(y_{i,t+j}) \sum_{s=1}^{T} u_{i,s} \right\|^2 \\
\leq & \sum_{j=1}^{J} \left(\frac{C_w J}{\varepsilon} \right)^2 O\left(\frac{\zeta_0^2(K)K}{NT} \right) \to 0
\end{aligned}$$

with the same argument on J. ■

A.3 Within group type estimator

Using lemmas in A.1, we now prove the main results in Section 3. The basic idea of the proof of Theorem 3.1 is mainly obtained from Newey (1997).

Proof of Theorem 3.1 As in Section 4.1, for notational convenience, we define $NT \times K$ matrices $\underline{\mathbf{g}}_K = \left(\underline{g}_K(y_{1,0}), \cdots, \underline{g}_K(y_{N,T-1}) \right)'$ and $\mathbf{g}_K^0 = \left(g_K^0(y_{1,0}), \cdots, g_K^0(y_{N,T-1}) \right)'$; $NT \times 1$ vectors $\mathbf{u} = (u_{1,1}, \cdots, u_{N,T})'$, $\mathbf{u}^0 = \left(u_{1,1}^0, \cdots, u_{N,T}^0 \right)'$, $\mathbf{m} = (m(y_{1,0}), \cdots, m(y_{N,T-1}))'$ and $\mathbf{m}^0 = \left(m^0(y_{1,0}), \cdots, m^0(y_{N,T-1}) \right)'$. Then, we can write

$$\widehat{\theta}_K - \theta_K = \left(\mathbf{g}_K^{0\prime} \mathbf{g}_K^0 / NT \right)^{-1} \left(\mathbf{g}_K^{0\prime} \mathbf{u}^0 / NT \right) + \left(\mathbf{g}_K^{0\prime} \mathbf{g}_K^0 / NT \right)^{-1} \left(\mathbf{g}_K^{0\prime} \left(\mathbf{m}^0 - \mathbf{g}_K^0 \theta_K \right) / NT \right).$$

Also note that using Lemma A1.1 to A1.4, we have

$$\begin{aligned}
\left(\mathbf{g}_K^{0\prime} \mathbf{g}_K^0 / NT \right)^{-1/2} &= \left(\underline{\mathbf{g}}_K' \underline{\mathbf{g}}_K / NT \right)^{-1/2} + O_p\left(\zeta_0^2(K) K / \sqrt{NT} \right) \\
\mathbf{g}_K^{0\prime} \mathbf{u}^0 / NT &= \underline{\mathbf{g}}_K' \mathbf{u} / NT + O_p\left(\zeta_0(K) K^{1/2} / \sqrt{NT} \right) \\
\mathbf{g}_K^{0\prime} \left(\mathbf{m}^0 - \mathbf{g}_K^0 \theta_K \right) / NT &= \underline{\mathbf{g}}_K' \left(\mathbf{m} - \mathbf{g}_K \theta_K \right) / NT + O_p\left(\zeta_0(K) K^{1/2-\delta} / \sqrt{NT} \right),
\end{aligned}$$

where the first result is due to the Taylor expansion and the fact that $\underline{\mathbf{g}}_K'\underline{\mathbf{g}}_K/NT = O_p(1)$. Moreover, with the similar argument as (a3), $\left\|\widehat{\Gamma}_K^{-1/2}\right\| = O_p(1)$.

First observe that

$$\mathbb{E}\left\|\Gamma_K^{-1/2}\left(\underline{\mathbf{g}}_K'\mathbf{u}/NT\right)\right\|^2 = \mathbb{E}\left(\mathbf{u}'\underline{\mathbf{g}}_K\Gamma_K^{-1}\underline{\mathbf{g}}_K'\mathbf{u}\right)/(NT)^2 = tr\left[\Gamma_K^{-1/2}\mathbb{E}\left(\underline{\mathbf{g}}_K'\mathbf{u}\mathbf{u}\underline{\mathbf{g}}_K\right)\Gamma_K^{-1/2}\right]/(NT)^2,$$

where

$$\begin{aligned}
\mathbb{E}\left(\underline{\mathbf{g}}_K'\mathbf{u}\mathbf{u}\underline{\mathbf{g}}_K\right) &= \mathbb{E}\left(\sum_{i=1}^N\sum_{t=1}^T\underline{g}_K(y_{i,t-1})u_{i,t}\right)\left(\sum_{i=1}^N\sum_{t=1}^T u_{i,t}\underline{g}_K'(y_{i,t-1})\right) \\
&= N\mathbb{E}\left(\sum_{t=1}^T\underline{g}_K(y_{i,t-1})u_{i,t}\right)\left(\sum_{s=1}^T u_{i,s}\underline{g}_K'(y_{i,s-1})\right) \\
&= NT\mathbb{E}\left(\underline{g}_K(y_{i,0})u_{i,1}^2\underline{g}_K'(y_{i,0})\right) + 2NT\sum_{\tau=1}^{T-1}(1-\tau/T)\mathbb{E}\left(\underline{g}_K(y_{i,0})u_{i,1}u_{i,1+\tau}\underline{g}_K'(y_{i,\tau})\right).
\end{aligned}$$

The first term is simply $NT\sigma^2\Gamma_K$ by the law of iterated expectations. For the second term, similarly as the proof in Lemma A1.3,

$$\left|2NT\sum_{\tau=1}^{T-1}(1-\tau/T)\mathbb{E}\left(\underline{g}_K(y_{i,0})u_{i,1}u_{i,1+\tau}\underline{g}_K'(y_{i,\tau})\right)\right| \le 2NT\Gamma\sum_{\tau=1}^{\infty}\alpha(\tau).$$

Therefore,

$$\mathbb{E}\left\|\Gamma_K^{-1/2}\left(\underline{\mathbf{g}}_K'\mathbf{u}/NT\right)\right\|^2 \le tr\left[\Gamma_K^{-1/2}\left\{\sigma^2\Gamma_K + 2\Gamma\sum_{\tau=1}^{\infty}\alpha(\tau)\right\}\Gamma_K^{-1/2}\right]/NT = O(K/NT),$$

since $\sum_{\tau=1}^{\infty}\alpha(\tau) < \infty$. Substituting $\widehat{\Gamma}_K$ for Γ_K does not change the result since

$$\begin{aligned}
\left\|\widehat{\Gamma}_K^{-1/2}\left(\underline{\mathbf{g}}_K'\mathbf{u}/NT\right)\right\|^2 &\le \left\|\Gamma_K^{-1/2}\left(\underline{\mathbf{g}}_K'\mathbf{u}/NT\right)\right\|^2 + \left\|\left(\widehat{\Gamma}_K^{-1/2} - \Gamma_K^{-1/2}\right)\left(\underline{\mathbf{g}}_K'\mathbf{u}/NT\right)\right\|^2 \\
&\le \left\|\Gamma_K^{-1/2}\left(\underline{\mathbf{g}}_K'\mathbf{u}/NT\right)\right\|^2 + \left\|\widehat{\Gamma}_K^{-1/2}\right\|^2\left\|\Gamma_K^{1/2} - \widehat{\Gamma}_K^{1/2}\right\|^2\left\|\Gamma_K^{-1/2}\left(\underline{\mathbf{g}}_K'\mathbf{u}/NT\right)\right\|^2 \\
&= O_p(K/NT) \qquad \text{(a8)}
\end{aligned}$$

for $\left\|\widehat{\Gamma}_K - \Gamma_K\right\| \to_p 0$ with $\|\Gamma_K\| < \infty$ by Lemma 3.1. It follows that

$$\begin{aligned}
\left\|\left(\mathbf{g}_K^{0\prime}\mathbf{g}_K^0/NT\right)^{-1}\left(\mathbf{g}_K^{0\prime}\mathbf{u}^0/NT\right)\right\|^2 &\le \left\|\widehat{\Gamma}_K^{-1/2}\right\|^2\left\|\widehat{\Gamma}_K^{-1/2}\left(\underline{\mathbf{g}}_K'\mathbf{u}/NT + O_p\left(\zeta_0(K)K^{1/2}/\sqrt{NT}\right)\right)\right\|^2 \\
&\le O_p\left(K/NT + \zeta_0^2(K)K/NT\right).
\end{aligned}$$

since $\left\|\widehat{\Gamma}_K^{-1/2}\right\| = O_p(1)$.

Secondly, using Lemma A1.4 and since $\underline{\mathbf{g}}\left(\underline{\mathbf{g}}_K'\underline{\mathbf{g}}_K\right)^{-1}\underline{\mathbf{g}}_K$ is idempotent[25],

$$\begin{aligned}
&\left\|\widehat{\Gamma}_K^{-1/2}\left(\left(\underline{\mathbf{g}}_K'(\mathbf{m} - \mathbf{g}_K\theta_K)/NT\right)\right)\right\|^2 \\
=& \left\|\left(\left(\underline{\mathbf{g}}_K'\underline{\mathbf{g}}_K/NT\right)^{-1/2} + O_p\left(\zeta_0^2(K)K/\sqrt{NT}\right)\right)\left(\underline{\mathbf{g}}_K'(\mathbf{m} - \mathbf{g}_K\theta_K)/NT\right)\right\|^2 \\
\le& \left\|\left(\underline{\mathbf{g}}_K'\underline{\mathbf{g}}_K/NT\right)^{-1/2}\left(\underline{\mathbf{g}}_K'(\mathbf{m} - \mathbf{g}_K\theta_K)/NT\right)\right\|^2 + O_p\left(\zeta_0^4(K)K^2/NT\right)\left\|\underline{\mathbf{g}}_K'(\mathbf{m} - \mathbf{g}_K\theta_K)/NT\right\|^2 \\
\le& \left((\mathbf{m} - \mathbf{g}_K\theta_K)'\underline{\mathbf{g}}_K\left(\underline{\mathbf{g}}_K'\underline{\mathbf{g}}_K\right)^{-1}\underline{\mathbf{g}}_K'(\mathbf{m} - \mathbf{g}_K\theta_K)\right)/NT + O_p\left(\zeta_0^4(K)K^2/NT\right)O_p\left(\zeta_0^2(K)K^{1-2\delta}/NT\right) \\
\le& \left((\mathbf{m} - \mathbf{g}_K\theta_K)'(\mathbf{m} - \mathbf{g}_K\theta_K)\right)/NT + O_p\left(\zeta_0^6(K)K^{3-2\delta}/(NT)^2\right) \\
=& O_p\left(K^{-2\delta} + \zeta_0^6(K)K^{3-2\delta}/(NT)^2\right),
\end{aligned}$$

[25] Since all the eigenvalues of any idempotent matrix P is either zero or one, $x'Px \le x'Ix$ for non-zero vector x and the identity matrix I with conformable dimensions.

giving

$$\begin{aligned}
&\left\|\left(\mathbf{g}_K^{0\prime}\mathbf{g}_K^0/NT\right)^{-1}\left(\mathbf{g}_K^{0\prime}\left(\mathbf{m}^0-\mathbf{g}_K^0\theta_K\right)/NT\right)\right\|^2 \\
\leq\ & \left\|\widehat{\Gamma}_K^{-1/2}\right\|^2\left\|\widehat{\Gamma}_K^{-1/2}\left(\underline{\mathbf{g}}_K'\left(\mathbf{m}-\mathbf{g}_K\theta_K\right)/NT+O_p\left(\zeta_0\left(K\right)K^{1/2-\delta}/\sqrt{NT}\right)\right)\right\|^2 \\
\leq\ & \left\|\widehat{\Gamma}_K^{-1/2}\right\|^2\left\|\widehat{\Gamma}_K^{-1/2}\left(\underline{\mathbf{g}}_K'\left(\mathbf{m}-\mathbf{g}_K\theta_K\right)/NT\right)\right\|^2+\left\|\widehat{\Gamma}_K^{-1/2}\right\|^4 O_p\left(\zeta_0^2\left(K\right)K^{1-2\delta}/NT\right) \\
\leq\ & O_p\left(K^{-2\delta}+\zeta_0^2\left(K\right)K^{1-2\delta}/NT\right)
\end{aligned}$$

since $\left\|\widehat{\Gamma}_K^{-1/2}\right\|=O_p\left(1\right)$ and $\zeta_0^6\left(K\right)K^{3-2\delta}/\left(NT\right)^2=\left(\zeta_0^4\left(K\right)K^2/NT\right)\left(\zeta_0^2\left(K\right)K^{1-2\delta}/NT\right)=o\left(1\right)\left(\zeta_0^2\left(K\right)K^{1-2\delta}/NT\right)<\zeta_0^2\left(K\right)K^{1-2\delta}/NT$. Therefore,

$$\begin{aligned}
\left\|\widehat{\theta}_K-\theta_K\right\|^2 &\leq \left\|\left(\mathbf{g}_K^{0\prime}\mathbf{g}_K^0/NT\right)^{-1}\left(\mathbf{g}_K^{0\prime}\mathbf{u}^0/NT\right)\right\|^2+\left\|\left(\mathbf{g}_K^{0\prime}\mathbf{g}_K^0/NT\right)^{-1}\left(\mathbf{g}_K^{0\prime}\left(\mathbf{m}^0-\mathbf{g}_K^0\theta_K\right)/NT\right)\right\|^2 \\
&= O_p\left(K/NT+K^{-2\delta}+\zeta_0^2\left(K\right)K/NT\right)
\end{aligned}$$

since $\zeta_0^2\left(K\right)K^{1-2\delta}/NT$ is dominated by $\zeta_0^2\left(K\right)K/NT$ for $\delta>0$. Next, by the triangular inequality,

$$\begin{aligned}
\int_{y\in\mathcal{Y}_c}\left[\widehat{m}\left(y\right)-m\left(y\right)\right]^2dF\left(y\right) &= \int_{y\in\mathcal{Y}_c}\left[g_K\left(y\right)'\left(\widehat{\theta}_K-\theta_K\right)+\left(g_K\left(y\right)'\theta_K-m\left(y\right)\right)\right]^2dP\left(y\right) \\
&\leq \left\|\widehat{\theta}_K-\theta_K\right\|^2+\int_{y\in\mathcal{Y}_c}\left[\left(g_K\left(y\right)'\theta_K-m\left(y\right)\right)\right]^2dP\left(y\right) \\
&= O_p\left(K/NT+K^{-2\delta}+\zeta_0^2\left(K\right)K/NT\right)+O\left(K^{-2\delta}\right) \\
&= O_p\left(K/NT+K^{-2\delta}+\zeta_0^2\left(K\right)K/NT\right).
\end{aligned}$$

For the uniform convergence rate, if we use the triangular inequality and Cauchy-Schwartz inequalities, we have

$$\begin{aligned}
&\sup_{y\in\mathcal{Y}_c}\max_{s\leq D}\left|d^s\left(\widehat{m}\left(y\right)-m\left(y\right)\right)/dy^s\right| \\
\leq\ & \sup_{y\in\mathcal{Y}_c}\max_{s\leq D}\left\|d^s\left(g_K\left(y\right)'\left(\widehat{\theta}_K-\theta_K\right)\right)/dy^s\right\|+\sup_{y\in\mathcal{Y}_c}\max_{s\leq D}\left\|d^s\left(g_K\left(y\right)'\theta_K-m\left(y\right)\right)/dy^s\right\| \\
\leq\ & K^{1/2}\zeta_D\left(K\right)\left\|\widehat{\theta}_K-\theta_K\right\|+O\left(K^{-\delta}\right) \\
=\ & O_p\left(K^{1/2}\zeta_D\left(K\right)\left(K^{1/2}/\sqrt{NT}+K^{-\delta}+\zeta_0\left(K\right)K^{1/2}/\sqrt{NT}\right)\right)
\end{aligned}$$

by Assumption W2. ■

Proof of Theorem 3.2 The within group type estimator for $m\left(\cdot\right)$ can be written as

$$\widehat{m}\left(y\right)-m\left(y\right)=g_K\left(y\right)'\left(\widehat{\theta}_K-\theta_K\right)-\left(m\left(y\right)-g_K\left(y\right)'\theta_K\right)$$

or

$$\begin{aligned}
\frac{\sqrt{NT}\left(\widehat{m}\left(y\right)-m\left(y\right)+\left(1/T\right)g_K\left(y\right)'\Gamma_K^{-1}\Phi_K\right)}{\sqrt{g_K\left(y\right)'\widehat{\Gamma}_K^{-1}g_K\left(y\right)}} &= \frac{g_K\left(y\right)'\sqrt{NT}\left(\widehat{\theta}_K-\theta_K+\left(1/T\right)\Gamma_K^{-1}\Phi_K\right)}{\sqrt{g_K\left(y\right)'\widehat{\Gamma}_K^{-1}g_K\left(y\right)}} \\
&\quad -\frac{\sqrt{NT}\left(m\left(y\right)-g_K\left(y\right)'\theta_K\right)}{\sqrt{g_K\left(y\right)'\widehat{\Gamma}_K^{-1}g_K\left(y\right)}}. \qquad \text{(a9)}
\end{aligned}$$

By Assumption W2, the second term in (a9) is negligible since

$$\left\|\frac{\sqrt{NT}\left(m\left(y\right)-g_K\left(y\right)'\theta_K\right)}{\sqrt{g_K\left(y\right)'\widehat{\Gamma}_K^{-1}g_K\left(y\right)}}\right\|\leq O_p\left(1\right)O_p\left(K^{-\delta}\sqrt{NT}\right)=O_p\left(K^{-\delta}\sqrt{NT}\right)\to 0.$$

Therefore, the asymptotic distribution of $\widehat{m}(\cdot)$ is determined by the asymptotic behavior of the first term in (a9), which is given by

$$
\begin{aligned}
&= \frac{g_K(y)' \sqrt{NT}\left(\widehat{\theta}_K - \theta_K + (1/T)\Gamma_K^{-1}\Phi_K\right)}{\sqrt{g_K(y)'\widehat{\Gamma}_K^{-1} g_K(y)}} \qquad \text{(a10)}\\
&= \left(\frac{1}{\sqrt{NT}}\sum_{i=1}^{N}\sum_{t=1}^{T}\widehat{\rho}'\widehat{\Gamma}_K^{-1/2}\underline{g}_K(y_{i,t-1})\,u_{i,t}\right) - \widehat{\rho}'\widehat{\Gamma}_K^{-1/2}\left(\frac{1}{\sqrt{NT^3}}\sum_{i=1}^{N}\sum_{t=1}^{T}\underline{g}_K(y_{i,t-1})\sum_{s=1}^{T}u_{i,t} - \sqrt{\frac{N}{T}}\Phi_K\right)\\
&\quad + \widehat{\rho}'\widehat{\Gamma}_K^{-1/2}\left(\frac{1}{\sqrt{NT}}\sum_{i=1}^{N}\sum_{t=1}^{T} g_K(y_{i,t-1})\left\{m^0(y_{i,t-1}) - g_K^0(y_{i,t-1})'\theta_K\right\}\right),
\end{aligned}
$$

where $\widehat{\rho} = g_K(y)'\widehat{\Gamma}_K^{-1/2}/\sqrt{g_K(y)'\widehat{\Gamma}_K^{-1}g_K(y)}$. By construction, $\|\widehat{\rho}\| = 1$. We look at the asymptotic distribution of (a10) in the following three steps.

[Step 1] We first consider the infeasible case that Γ_K is known. We have

$$
\begin{aligned}
&\frac{\sqrt{NT}\left(\widehat{m}(y) - m(y) + (1/T)\, g_K(y)'\Gamma_K^{-1}\Phi_K\right)}{\sqrt{g_K(y)'\Gamma_K^{-1}g_K(y)}}\\
&= \left(\frac{1}{\sqrt{NT}}\sum_{i=1}^{N}\sum_{t=1}^{T}\rho'\Gamma_K^{-1/2}\underline{g}_K(y_{i,t-1})\,u_{i,t}\right) - \rho'\Gamma_K^{-1/2}\left(\frac{1}{\sqrt{NT^3}}\sum_{i=1}^{N}\sum_{t=1}^{T}\underline{g}_K(y_{i,t-1})\sum_{s=1}^{T}u_{i,t} - \sqrt{\frac{N}{T}}\Phi_K\right)\\
&\quad + \rho'\Gamma_K^{-1/2}\left(\frac{1}{\sqrt{NT}}\sum_{i=1}^{N}\sum_{t=1}^{T} g_K(y_{i,t-1})\left\{m^0(y_{i,t-1}) - g_K^0(y_{i,t-1})'\theta_K\right\}\right),
\end{aligned}
$$

where $\rho = g_K(y)'\Gamma_K^{-1/2}/\sqrt{g_K(y)'\Gamma_K^{-1}g_K(y)}$ and $\|\rho\| = 1$ by construction. The first term converges in distribution to $\mathcal{N}\left(0,\sigma^2\right)$ by Lemma A1.5. The second term becomes negligible as $N, T \to \infty$ with $\lim_{N,T\to\infty} N/T \to \kappa$, $0 < \kappa < \infty$, since $\left\|\rho'\Gamma_K^{-1/2}\right\| \leq \|\rho\|\left\|\Gamma_K^{-1/2}\right\| < \infty$ from Assumption W1 and

$$
\left\|\frac{1}{\sqrt{NT^3}}\sum_{i=1}^{N}\sum_{t=1}^{T}\underline{g}_K(y_{i,t-1})\sum_{s=1}^{T}u_{i,t} - \sqrt{\frac{N}{T}}\Phi_K\right\| \leq \left\|\frac{1}{\sqrt{NT^3}}\sum_{i=1}^{N}\sum_{t=1}^{T}\underline{g}_K(y_{i,t-1})\sum_{s=1}^{T}u_{i,t} - \sqrt{\kappa}\Phi_K\right\| + \left\|\sqrt{\frac{N}{T}} - \sqrt{\kappa}\right\|\|\Phi\| \to_p 0,
$$

where the first part is $o_p(1)$ by Lemma A1.6; $\left\|\sqrt{N/T} - \sqrt{\kappa}\right\| \to 0$ for $N/T \to \kappa$; and $\|\Phi_K\| < \infty$ from Assumption W2. Finally, the third term also converges in probability to zero using Lemma A1.7. The asymptotic normality thus simply follows by adding these three results:

$$
\frac{\sqrt{NT}\left(\widehat{m}(y) - m(y) + (1/T)\, g_K(y)'\Gamma_K^{-1}\Phi_K\right)}{\sqrt{g_K(y)'\Gamma_K^{-1}g_K(y)}} \to_d \mathcal{N}\left(0,\sigma^2\right).
$$

[Step 2] We now consider another infeasible case that

$$
\begin{aligned}
&\frac{\sqrt{NT}\left(\widehat{m}(y) - m(y) + (1/T)\, g_K(y)'\Gamma_K^{-1}\Phi_K\right)}{\sqrt{g_K(y)'\Gamma_K^{-1}g_K(y)}} \qquad \text{(a11)}\\
&= \left(\frac{1}{\sqrt{NT}}\sum_{i=1}^{N}\sum_{t=1}^{T}\widetilde{\rho}'\widehat{\Gamma}_K^{-1/2}\underline{g}_K(y_{i,t-1})\,u_{i,t}\right) - \widetilde{\rho}'\widehat{\Gamma}_K^{-1/2}\left(\frac{1}{\sqrt{NT^3}}\sum_{i=1}^{N}\sum_{t=1}^{T}\underline{g}_K(y_{i,t-1})\sum_{s=1}^{T}u_{i,t} - \sqrt{\frac{N}{T}}\Phi_K\right)\\
&\quad + \widetilde{\rho}'\widehat{\Gamma}_K^{-1/2}\left(\frac{1}{\sqrt{NT}}\sum_{i=1}^{N}\sum_{t=1}^{T} g_K(y_{i,t-1})\left\{m^0(y_{i,t-1}) - g_K^0(y_{i,t-1})'\theta_K\right\}\right),
\end{aligned}
$$

where $\widetilde{\rho} = g_K(y)'\widehat{\Gamma}_K^{-1/2}/\sqrt{g_K(y)'\Gamma_K^{-1}g_K(y)}$. If we use the matrix notation defined in the proof of Theorem 3.1, the first term is

$$
\widetilde{\rho}'\Gamma_K^{-1/2}\underline{\mathbf{g}}_K'\mathbf{u}/\sqrt{NT} = \rho'\Gamma_K^{-1/2}\underline{\mathbf{g}}_K'\mathbf{u}/\sqrt{NT} + \left[g_K(y)'\Gamma_K^{-1}g_K(y)\right]^{-1/2} g_K(y)'\left[\widehat{\Gamma}_K^{-1} - \Gamma_K^{-1}\right]\underline{\mathbf{g}}_K'\mathbf{u}/\sqrt{NT},
$$

where the residual term is

$$
\begin{aligned}
& \left\| \left[g_K(y)' \Gamma_K^{-1} g_K(y) \right]^{-1/2} g_K(y)' \left[\widehat{\Gamma}_K^{-1} - \Gamma_K^{-1} \right] \underline{\mathbf{g}}_K' \mathbf{u} / \sqrt{NT} \right\| \\
= & \left\| \left[g_K(y)' \Gamma_K^{-1} g_K(y) \right]^{-1/2} g_K(y)' \widehat{\Gamma}_K^{-1} \left[\Gamma_K - \widehat{\Gamma}_K \right] \Gamma_K^{-1} \underline{\mathbf{g}}_K' \mathbf{u} / \sqrt{NT} \right\| \\
\leq & \left\| \left[g_K(y)' \Gamma_K^{-1} g_K(y) \right]^{-1/2} g_K(y)' \widehat{\Gamma}_K^{-1} \right\| \left\| \left[\Gamma_K - \widehat{\Gamma}_K \right] \Gamma_K^{-1/2} \right\| \left\| \Gamma_K^{-1/2} \underline{\mathbf{g}}_K' \mathbf{u} / \sqrt{NT} \right\| \\
\leq & O_p(1) O_p\left(\zeta_0^2(K) K / \sqrt{NT} \right) O_p\left(K^{1/2} / \sqrt{NT} \right) \\
= & O_p\left(\zeta_0^2(K) K^{3/2} / NT \right) \to 0
\end{aligned}
$$

since $\left\| \left[g_K(y)' \Gamma_K^{-1} g_K(y) \right]^{-1/2} g_K(y)' \widehat{\Gamma}_K^{-1} \right\| \leq \|\widetilde{\rho}\| \left\| \widehat{\Gamma}_K^{-1/2} \right\| = O_p(1)$; $\left\| \left[\Gamma_K - \widehat{\Gamma}_K \right] \Gamma_K^{-1/2} \right\| \leq \left\| \Gamma_K - \widehat{\Gamma}_K \right\| \left\| \Gamma_K^{-1/2} \right\| \leq O_p\left(\zeta_0^2(K) K / \sqrt{NT} \right)$ by Lemma A1.3; and $\left\| \Gamma_K^{-1/2} \underline{\mathbf{g}}_K' \mathbf{u} / \sqrt{NT} \right\| \leq O_p\left(K^{1/2} / \sqrt{NT} \right)$ as (a8). Note that $|\|\widetilde{\rho}\| - 1| = o_p(1)$ by Lemma A1.3 and $\zeta_0^2(K) K^{3/2}/NT \leq \zeta_0^4(K) K^2/NT \to 0$. Therefore, using [Step 1], $\widetilde{\rho}' \Gamma_K^{-1/2} \underline{g}' \mathbf{u} / \sqrt{NT} \to_d \mathcal{N}(0, \sigma^2)$. Now the rest two terms in (a11) are still asymptotically negligible similarly as in [Step 1] since

$$
\begin{aligned}
\left\| \widetilde{\rho}' \widehat{\Gamma}_K^{-1/2} - \rho' \Gamma_K^{-1/2} \right\| & = \left\| \left[g_K(y)' \Gamma_K^{-1} g_K(y) \right]^{-1/2} g_K(y)' \widehat{\Gamma}_K^{-1} - \left[g_K(y)' \Gamma_K^{-1} g_K(y) \right]^{-1/2} g_K(y)' \Gamma_K^{-1} \right\| \\
& \leq \left\| \left[g_K(y)' \Gamma_K^{-1} g_K(y) \right]^{-1/2} g_K(y)' \Gamma_K^{-1/2} \right\| \left\| \Gamma_K^{1/2} \right\| \left\| \widehat{\Gamma}_K^{-1} - \Gamma_K^{-1} \right\| \\
& = \|\rho\| \left\| \Gamma_K^{1/2} \right\| \left\| \widehat{\Gamma}_K^{-1} - \Gamma_K^{-1} \right\| \leq O_p\left(\zeta_0^2(K) K / \sqrt{NT} \right) \to 0.
\end{aligned}
$$

[Step 3] We finally consider the feasible case[26] given by

$$
\begin{aligned}
& \frac{\sqrt{NT}\left(\widehat{m}(y) - m(y) + (1/T) g_K(y)' \Gamma_K^{-1} \Phi_K \right)}{\sqrt{g_K(y)' \widehat{\Gamma}_K^{-1} g_K(y)}} \\
= & \left(\frac{1}{\sqrt{NT}} \sum_{i=1}^N \sum_{t=1}^T \widehat{\rho}' \widehat{\Gamma}_K^{-1/2} \underline{g}_K(y_{i,t-1}) u_{i,t} \right) - \widehat{\rho}' \widehat{\Gamma}_K^{-1/2} \left(\frac{1}{\sqrt{NT^3}} \sum_{i=1}^N \sum_{t=1}^T \underline{g}_K(y_{i,t-1}) \sum_{s=1}^T u_{i,t} - \sqrt{\frac{N}{T}} \Phi_K \right) \\
& + \widehat{\rho}' \widehat{\Gamma}_K^{-1/2} \left(\frac{1}{\sqrt{NT}} \sum_{i=1}^N \sum_{t=1}^T g_K(y_{i,t-1}) \left\{ m^0(y_{i,t-1}) - g_K^0(y_{i,t-1})' \theta_K \right\} \right),
\end{aligned}
$$

where $\widehat{\rho} = \widehat{\Gamma}_K^{-1/2} g_K(y) / \sqrt{g_K(y)' \widehat{\Gamma}_K^{-1} g_K(y)}$ and $\|\widehat{\rho}\| = 1$ by construction. Notice that the only difference between [Step 2] and [Step 3] lies in the difference between $\widetilde{\rho}$ and $\widehat{\rho}$. Similarly as the proof in [Step 2], we first look at

$$
\widehat{\rho}' \Gamma_K^{-1/2} \underline{\mathbf{g}}_K' \mathbf{u} / \sqrt{NT} = \widetilde{\rho}' \Gamma_K^{-1/2} \underline{\mathbf{g}}_K' \mathbf{u} / \sqrt{NT} + \left\{ \left[g_K(y)' \widehat{\Gamma}_K^{-1} g_K(y) \right]^{-1/2} - \left[g_K(y)' \Gamma_K^{-1} g_K(y) \right]^{-1/2} \right\} g_K(y)' \widehat{\Gamma}_K^{-1} \underline{\mathbf{g}}_K' \mathbf{u} / \sqrt{NT},
$$

where the residual term is

$$
\begin{aligned}
& \left\| \left\{ \left[g_K(y)' \widehat{\Gamma}_K^{-1} g_K(y) \right]^{-1/2} - \left[g_K(y)' \Gamma_K^{-1} g_K(y) \right]^{-1/2} \right\} g_K(y)' \widehat{\Gamma}_K^{-1} \underline{\mathbf{g}}_K' \mathbf{u} / \sqrt{NT} \right\| \\
\leq & \left\| \left[g_K(y)' \widehat{\Gamma}_K^{-1} g_K(y) \right]^{-1/2} g_K(y)' - \left[g_K(y)' \Gamma_K^{-1} g_K(y) \right]^{-1/2} g_K(y)' \right\| \left\| \widehat{\Gamma}_K^{-1} \right\| \left\| \Gamma_K^{1/2} \right\| \left\| \Gamma_K^{-1/2} \underline{\mathbf{g}}_K' \mathbf{u} / \sqrt{NT} \right\| \\
\leq & \left\{ \left\| \left[g_K(y)' \widehat{\Gamma}_K^{-1} g_K(y) \right]^{-1/2} g_K(y)' \widehat{\Gamma}_K^{-1/2} \right\| \left\| \widehat{\Gamma}_K^{1/2} \right\| \right. \\
& \left. + \left\| \left[g_K(y)' \Gamma_K^{-1} g_K(y) \right]^{-1/2} g_K(y)' \Gamma_K^{-1/2} \right\| \left\| \Gamma_K^{1/2} \right\| \right\} \left\| \widehat{\Gamma}_K^{-1} \right\| \left\| \Gamma_K^{1/2} \right\| \left\| \Gamma_K^{-1/2} \underline{\mathbf{g}}_K' \mathbf{u} / \sqrt{NT} \right\| \\
= & \left\{ \left\| \widehat{\Gamma}_K^{1/2} \right\| + \left\| \Gamma_K^{1/2} \right\| \right\} \left\| \widehat{\Gamma}_K^{-1} \right\| \left\| \Gamma_K^{1/2} \right\| \left\| \Gamma_K^{-1/2} \underline{\mathbf{g}}_K' \mathbf{u} / \sqrt{NT} \right\| \\
= & O_p\left(K^{1/2} / \sqrt{NT} \right) \to 0.
\end{aligned}
$$

[26] We, however, still assume the asymptotic bias is of known form.

Therefore, using the proof in [Step 2], $\widetilde{\rho}'\Gamma_K^{-1/2}\underline{g}'\mathbf{u}/\sqrt{NT}\rightarrow_d \mathcal{N}\left(0,\sigma^2\right)$. Now the rest two terms are still asymptotically negligible similarly as in [Step 2] since

$$\begin{aligned}
\left\|\widehat{\rho}'\widehat{\Gamma}_K^{-1/2}-\widetilde{\rho}'\widehat{\Gamma}_K^{-1/2}\right\| &= \left\|\left[g_K(y)'\widehat{\Gamma}_K^{-1}g_K(y)\right]^{-1/2}g_K(y)'\widehat{\Gamma}_K^{-1}-\left[g_K(y)'\Gamma_K^{-1}g_K(y)\right]^{-1/2}g_K(y)'\widehat{\Gamma}_K^{-1}\right\| \\
&\leq \left\|\left[g_K(y)'\widehat{\Gamma}_K^{-1}g_K(y)\right]^{-1/2}g_K(y)'\widehat{\Gamma}_K^{-1/2}\right\|\left\|\widehat{\Gamma}_K^{1/2}\right\|+\left\|\left[g_K(y)'\Gamma_K^{-1}g_K(y)\right]^{-1/2}g_K(y)'\Gamma_K^{-1/2}\right\|\left\|\Gamma_K^{1/2}\right\| \\
&= \left\{\left\|\widehat{\Gamma}_K^{1/2}\right\|+\left\|\Gamma_K^{1/2}\right\|\right\}\left\|\widehat{\Gamma}_K^{-1}\right\| \\
&= O_p(1).
\end{aligned}$$

The desired result then follows using Lemma A1.9. ■

Proof of Theorem 3.3 First observe that

$$v(K,N,T)^{-1/2}\left(\widetilde{m}(y)-m(y)\right)=v(K,N,T)^{-1/2}\left(\widehat{m}(y)-m(y)+\frac{1}{T}b_K(y)\right)+\frac{1}{T}v(K,N,T)^{-1/2}\left(\widehat{b}_K(y)-b_K(y)\right),$$

where the first part converges in distribution to the standard normal as $N,T\rightarrow\infty$ by Theorem 3.3. For the second part, we will show that $\left\|(1/T)\,v(K,N,T)^{-1/2}\left(\widehat{b}_K(y)-b_K(y)\right)\right\|\rightarrow_p 0$ as $N,T\rightarrow\infty$ to complete the proof. Note that

$$\begin{aligned}
&\left\|\frac{1}{T}v(K,N,T)^{-1/2}\left(\widehat{b}_K(y)-b_K(y)\right)\right\| \\
\leq\; &\frac{1}{T}\left|\frac{g_K(y)'\Gamma_K^{-1}g_K(y)}{NT}\right|^{-1/2}\left\|g_K(y)'\widehat{\Gamma}_K^{-1}\widehat{\Phi}_K-g_K(y)'\Gamma_K^{-1}\widehat{\Phi}_K\right\| \\
&+\frac{1}{T}\left|\frac{g_K(y)'\Gamma_K^{-1}g_K(y)}{NT}\right|^{-1/2}\left\|g_K(y)'\Gamma_K^{-1}\widehat{\Phi}_K-g_K(y)'\Gamma_K^{-1}\Phi_K\right\| \\
=\; &\sqrt{\frac{N}{T}}\left\|\frac{g_K(y)'\left(\widehat{\Gamma}_K^{-1}-\Gamma_K^{-1}\right)\widehat{\Phi}_K}{\sqrt{g_K(y)'\Gamma_K^{-1}g_K(y)}}\right\|+\sqrt{\frac{N}{T}}\left\|\frac{g_K(y)'\Gamma_K^{-1}\left(\widehat{\Phi}_K-\Phi_K\right)}{\sqrt{g_K(y)'\Gamma_K^{-1}g_K(y)}}\right\| \\
=\; &B_1(N,T,K)+B_2(N,T,K). \qquad\qquad \text{(a12)}
\end{aligned}$$

The second term $B_2(N,T,K)$ is simply $o(1)$ since $N/T\rightarrow\kappa<\infty$ and

$$\left\|\frac{g_K(y)'\Gamma_K^{-1}\left(\widehat{\Phi}_K-\Phi_K\right)}{\sqrt{g_K(y)'\Gamma_K^{-1}g_K(y)}}\right\|\leq\left\|\frac{g_K(y)'\Gamma_K^{-1/2}}{\sqrt{g_K(y)'\Gamma_K^{-1}g_K(y)}}\right\|\left\|\Gamma_K^{-1/2}\right\|\left\|\widehat{\Phi}_K-\Phi_K\right\|\rightarrow 0,$$

where for each K, the first norm is one by construction, the second norm is bounded by Assumption W1, and the third norm converges to zero in probability as $N,T\rightarrow\infty$ by Lemma A1.10. For the first term $B_1(N,T,K)$ in (a12), observe that

$$\begin{aligned}
\left\|\frac{g_K(y)'\left(\widehat{\Gamma}_K^{-1}-\Gamma_K^{-1}\right)\widehat{\Phi}_K}{\sqrt{g_K(y)'\Gamma_K^{-1}g_K(y)}}\right\| &\leq \left\|\frac{g_K(y)'\left(\widehat{\Gamma}_K^{-1}-\Gamma_K^{-1}\right)\left(\widehat{\Phi}_K-\Phi_K\right)}{\sqrt{g_K(y)'\Gamma_K^{-1}g_K(y)}}\right\|+\left\|\frac{g_K(y)'\left(\widehat{\Gamma}_K^{-1}-\Gamma_K^{-1}\right)\Phi_K}{\sqrt{g_K(y)'\Gamma_K^{-1}g_K(y)}}\right\| \\
&\leq \left\|\frac{g_K(y)'\Gamma_K^{-1/2}}{\sqrt{g_K(y)'\Gamma_K^{-1}g_K(y)}}\right\|\left\|\Gamma_K^{1/2}\right\|\left\|\widehat{\Gamma}_K^{-1}-\Gamma_K^{-1}\right\|\left\{\left\|\widehat{\Phi}_K-\Phi_K\right\|+\left\|\Phi_K\right\|\right\}\rightarrow 0
\end{aligned}$$

since for each K, the first norm is one by construction, the second norm is bounded by Assumption W1, the third norm converges to zero in probability as $N,T\rightarrow\infty$ by Lemma A1.3, the fourth norm also converges to zero in probability as $N,T\rightarrow\infty$ by Lemma A1.10, and the fifth norm is bounded by assumption. ■

Proof of Theorem 4.1 First note that for $y \in \mathcal{Y}_c$,

$$\begin{aligned}
& \sqrt{NT}\left(\widehat{m}\left(y\right)-m\left(y\right)\right) \\
= & \sqrt{NT}g_K\left(y\right)'\left(\widehat{\theta}_K-\theta_K\right) \\
= & \sqrt{NT}g_K\left(y\right)'\left(\mathbf{g}_K^{0\prime}M_x\mathbf{g}_K^0\right)^{-1}\mathbf{g}_K^{0\prime}M_x\left(\mathbf{m}^0-\mathbf{g}_K^{0\prime}\theta_K\right)+\sqrt{NT}g_K\left(y\right)'\left(\mathbf{g}_K^{0\prime}M_x\mathbf{g}_K^0\right)^{-1}\mathbf{g}_K^{0\prime}M_x\mathbf{u}^0 \qquad \text{(a13)}
\end{aligned}$$

and

$$\sqrt{NT}\left(\widehat{\gamma}-\gamma\right)=\sqrt{NT}\left(\mathbf{x}^{0\prime}M_g\mathbf{x}^0\right)^{-1}\mathbf{x}^{0\prime}M_g\mathbf{u}^0. \qquad \text{(a14)}$$

Similarly as Lemma A1.3, we have $\left\|\widehat{\Sigma}-\Sigma\right\|\to 0$ as $N,T\to\infty$, where $\widehat{\Sigma}=(1/NT)\left[\mathbf{g}_K^{0\prime},\mathbf{x}^{0\prime}\right]'\left[\mathbf{g}_K^{0\prime},\mathbf{x}^{0\prime}\right]$. Therefore, the first term of (a13) is simply negligible as in Lemma A1.5 from Assumption W2. For the second term in (a13) and the formula (a14), the result readily follows if we use the result of partitioned regressions. Since we approximate the unknown function $m(\cdot)$ using a linear combination of series functions, the estimation is just a partitioned regression. The detailed proof is, therefore, a straightforward extension of the proof of Theorem 3.4, and we simply discuss the heuristic idea of the proof here. By combining the second term in (a13) and the formula (a14), we have

$$\begin{pmatrix} \sqrt{NT}g_K\left(y\right)'\left(\mathbf{g}_K^{0\prime}M_x\mathbf{g}_K^0\right)^{-1}\mathbf{g}_K^{0\prime}M_x\mathbf{u}^0 \\ \sqrt{NT}\left(\mathbf{x}^{0\prime}M_g\mathbf{x}^0\right)^{-1}\mathbf{x}^{0\prime}M_g\mathbf{u}^0 \end{pmatrix}=\begin{pmatrix} g_K\left(y\right) \\ 1 \end{pmatrix}'\begin{pmatrix} \mathbf{g}_K^{0\prime}\mathbf{g}_K^0/NT & \mathbf{g}_K^{0\prime}\mathbf{x}^0/NT \\ \mathbf{x}^{0\prime}\mathbf{g}_K^0/NT & \mathbf{x}^{0\prime}\mathbf{x}^0/NT \end{pmatrix}^{-1}\begin{pmatrix} \mathbf{g}_K^{0\prime}\mathbf{u}^0/\sqrt{NT} \\ \mathbf{x}^{0\prime}\mathbf{u}^0/\sqrt{NT} \end{pmatrix}.$$

Since $x_{i,t}$ is strictly exogenous for all i and t, the limit distribution of $\begin{pmatrix} \mathbf{g}_K^{0\prime}\mathbf{g}_K^0/NT & \mathbf{g}_K^{0\prime}\mathbf{x}^0/NT \\ \mathbf{x}^{0\prime}\mathbf{g}_K^0/NT & \mathbf{x}^{0\prime}\mathbf{x}^0/NT \end{pmatrix}^{-1}\begin{pmatrix} \mathbf{g}_K^{0\prime}\mathbf{u}^0/\sqrt{NT} \\ \mathbf{x}^{0\prime}\mathbf{u}^0/\sqrt{NT} \end{pmatrix}$ is approximately normal with mean $\Sigma^{-1}\begin{pmatrix} -\sqrt{\kappa}\Phi_K \\ 0 \end{pmatrix}$ and variance $\sigma^2\Sigma^{-1}$ from Theorem 3.4 if we keep K fixed. By using the inverse matrix formula of the partitioned matrix, however,

$$\begin{aligned}
\Sigma^{-1} &= \begin{pmatrix} \Sigma_{gg} & \Sigma_{gx} \\ \Sigma_{xg} & \Sigma_{xx} \end{pmatrix}^{-1} \\
&= \begin{pmatrix} \Sigma_{gg\cdot x}^{-1} & -\Sigma_{gg\cdot x}^{-1}\Sigma_{gx}\Sigma_{xx}^{-1} \\ -\Sigma_{xx}^{-1}\Sigma_{xg}\Sigma_{gg\cdot x}^{-1} & \Sigma_{xx}^{-1}+\Sigma_{xx}^{-1}\Sigma_{xg}\Sigma_{gg\cdot x}^{-1}\Sigma_{gx}\Sigma_{xx}^{-1} \end{pmatrix}=\begin{pmatrix} \Sigma_{gg}^{-1}+\Sigma_{gg}^{-1}\Sigma_{gx}\Sigma_{xx\cdot g}^{-1}\Sigma_{xg}\Sigma_{gg}^{-1} & -\Sigma_{gg}^{-1}\Sigma_{gx}\Sigma_{xx\cdot g}^{-1} \\ -\Sigma_{xx\cdot g}^{-1}\Sigma_{xg}\Sigma_{gg}^{-1} & \Sigma_{xx\cdot g}^{-1} \end{pmatrix},
\end{aligned}$$

and we have the desired result using this expression. ■

Appendix B: Two Stage IV Estimation

Similarly as Newey and Powell (2003), we consider the general setting given by

$$\mathbb{E}\left[\rho\left(y_{i,t}, y_{i,t-1}, y_{i,t-2}; m\right)|z_{i,t}\right] = 0,$$

where $\rho\left(y_{i,t}, y_{i,t-1}, y_{i,t-2}; m\right) = \Delta y_{i,t} - \left\{m\left(y_{i,t-1}\right) - m\left(y_{i,t-2}\right)\right\}$. The first stage series estimator is then given by

$$\widehat{\rho}\left(z_{i,t}; m\right) = \left(\sum_{j=1}^{N}\sum_{s=1}^{T}\rho\left(y_{j,s}, y_{j,s-1}, y_{j,s-2}; m\right) q_J\left(z_{j,s}\right)\right)\left(\sum_{j=1}^{N}\sum_{s=1}^{T} q_J\left(z_{j,s}\right) q_J\left(z_{j,s}\right)'\right)^{-1} q_J\left(z_{i,t}\right).$$

The minimization problem given in (16) can be rewritten as

$$m\left(\widehat{\theta}_K\right) = \arg\min_{\theta_K}\widehat{Q}\left(\theta_K\right) = \sum_{i=1}^{N}\widehat{\rho}\left(z_i; m\right)' R\widehat{\rho}\left(z_i; m\right),$$

where $\widehat{\rho}\left(z_i; m\right)$ is the $(T \times 1)$ stack of $\widehat{\rho}\left(z_{i,t}; m\right)$. We assume the following conditions.

Assumption I1 *$\{y_{i,t}\}$ satisfies the stability conditions in Section 2, where the support $\mathcal{Y}$ of $y_{i,t}$ is bounded.*

The stationarity and mixing condition over t is only necessary when $T \to \infty$. The bounded support of $y_{i,t}$ is necessary to avoid any complications. For the details, refer to Newey and Powell (2003, p.1569). The next condition is the identification condition for m.

Assumption I2 *$m \in \Theta$ is uniquely identified satisfying $\mathbb{E}\left[\rho\left(y_{i,t}, y_{i,t-1}, y_{i,t-2}; m\right)|z_{i,t}\right] = 0$.*

Assumption I3 *There is a metric $\|\cdot\|_*$ such that Θ is compact under $\|\cdot\|_*$.*

Assumption I4 *For any $m(y) \in \Theta$ with $y \in \mathcal{Y}$, there exists a series approximation $g_K(y)'\theta_K$ such that $\left\|m(y) - g_K(y)'\theta_K\right\|_* \to 0$ as $K \to \infty$.*

Assumption I5 *$\mathbb{E}\left[\left|\rho\left(y_{i,t}, y_{i,t-1}, y_{i,t-2}; m\right)\right|^2|z_{i,t}\right]$ is bounded and $\rho\left(y_{i,t}, y_{i,t-1}, y_{i,t-2}; m\right)$ is Hölder continuous in $m \in \Theta$, i.e., there exists $M\left(y_1, y_2, y_3\right)$, $\nu > 0$ such that for all $m_1, m_2 \in \Theta$, $\left|\rho\left(y_1, y_2, y_3; m_1\right) - \rho\left(y_1, y_2, y_3; m_2\right)\right| \leq M\left(y_1, y_2, y_3\right)\left\|m_1 - m_2\right\|_*^{\nu}$ and $\mathbb{E}\left[\left|M\left(y_1, y_2, y_3\right)\right|^2|z_{i,t}\right]$ is bounded.*

The following condition assumes that the first stage series approximation can approximate any function with finite mean-square.

Assumption I6 *(i) For every J, there exists $\mathbb{E}q_J\left(z_{i,t}\right) q_J\left(z_{i,t}\right)'$, whose smallest eigenvalue is bounded away from zero uniformly in J and whose largest eigenvalue is bounded above uniformly in J; and (ii) for any $b(z)$ with $\mathbb{E}\left(b(z)\right)^2 < \infty$ there exist $q_J(z)$ and φ with $\mathbb{E}\left(b(z) - q_J(z)'\varphi\right)^2 \to 0$ as $J \to \infty$, where $J/N \to 0$ if T is fixed; $J/NT \to 0$ if T tends to infinity.*

We provide the consistency result as in Newey and Powell (2003) without proof.

Theorem B (Consistency: Newey and Powell (2003)) *Under Assumptions I1 to I6, $\left\|\widehat{\ell}(y) - \ell(y)\right\|_* \to_p 0$ and thus $\left\|\widehat{m}(y) - m(y)\right\|_* \to_p 0$ for any $y \in \mathcal{Y}$ as $K \to \infty$.*

Notice that Theorem B holds as long as $K \to \infty$ with $N \to \infty$, independent of $T \to \infty$ or not. However, there still remain more challenges when the length of time T is large. This is because the number of instruments increases as T goes to infinity, which generates the large number of moment conditions problem.

Appendix C: Simulation Results

Model 1 : $y_{i,t} = \mu_i + \{0.6y_{i,t-1}\} + u_{i,t}$

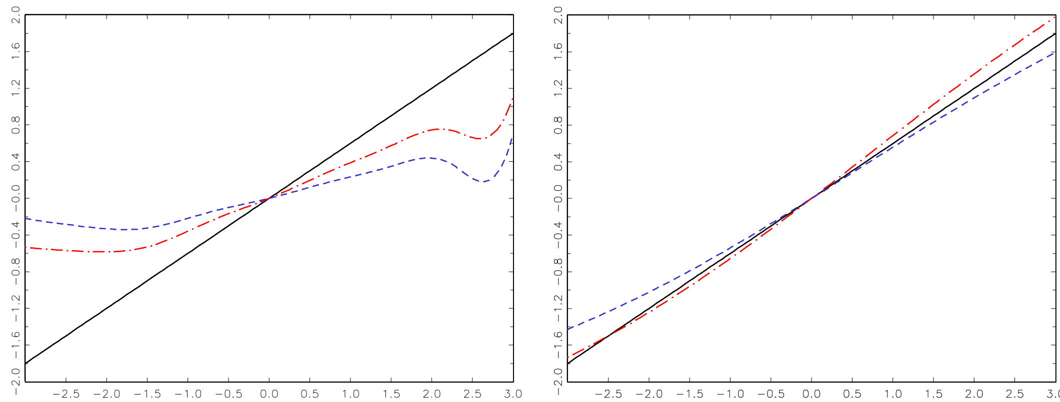

Figure C1 : Nonparametric estimation - Cubic splines (left, 4 knots) v.s. Power series (right, 4th polynomial). Solid (black) line: true; dotted (blue) line: series estimate; dashed (red) line: bias corrected.[27]

Model 2 : $y_{i,t} = \mu_i + \{\exp(y_{i,t-1}) / (1 + \exp(y_{i,t-1})) - 0.5\} + u_{i,t}$

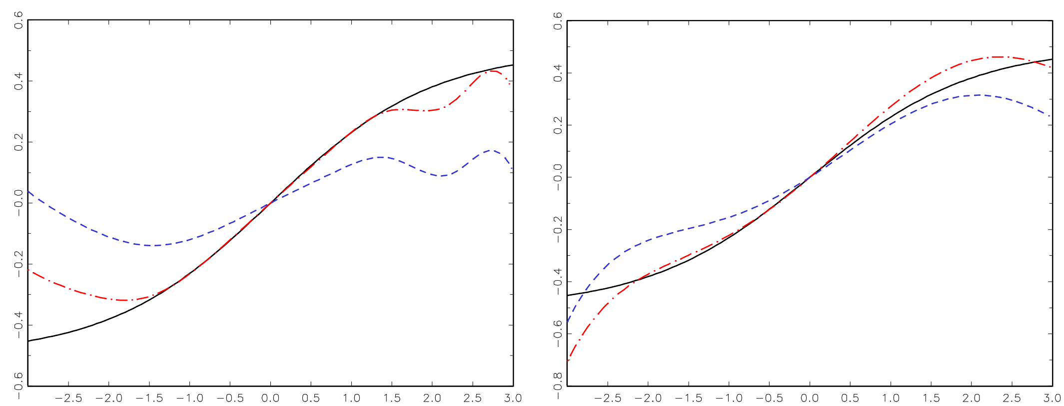

Figure C2 : Nonparametric estimation - Cubic splines (left, 4 knots) v.s. Power series (right, 4th polynomial). Solid (black) line: true; dotted (blue) line: series estimate; dashed (red) line: bias corrected.

[27] $(N, T) = (100, 50)$ and the values are averages over 1000 replications.

Model 3 : $y_{i,t} = \mu_i + \{\ln(|y_{i,t-1} - 1| + 1)\, sgn\,(y_{i,t-1} - 1) + \ln 2\} + u_{i,t}$

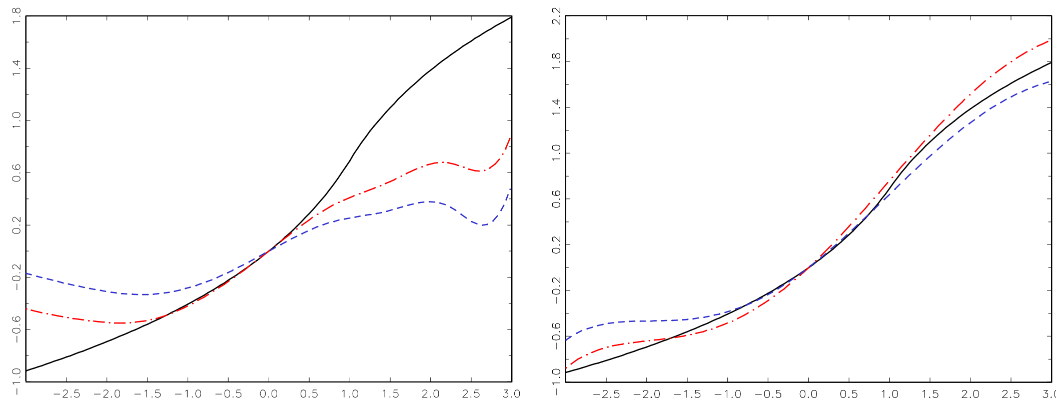

Figure C3 : Nonparametric estimation - Cubic splines (left, 4 knots) v.s. Power series (right, 4th polynomial). Solid (black) line: true; dotted (blue) line: series estimate; dashed (red) line: bias corrected.

Model 4 : $y_{i,t} = \mu_i + \{0.6y_{i,t-1} - 0.9y_{i,t-1}/\,(1 + \exp(y_{i,t-1} - 2.5))\} + u_{i,t}$

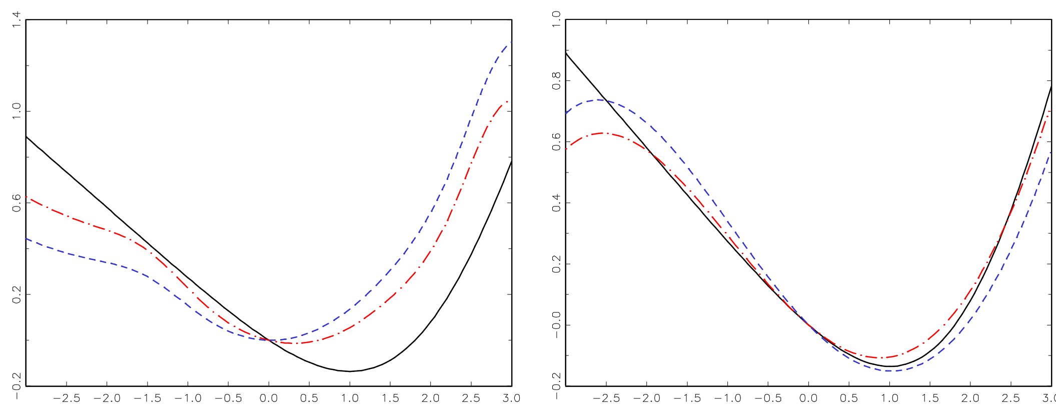

Figure C4 : Nonparametric estimation - Cubic splines (left, 4 knots) v.s. Power series (right, 4th polynomial). Solid (black) line: true; dotted (blue) line: series estimate; dashed (red) line: bias corrected.

Model 5 : $y_{i,t} = \mu_i + \left\{0.3y_{i,t-1} \exp\left(-0.1y_{i,t-1}^2\right)\right\} + u_{i,t}$

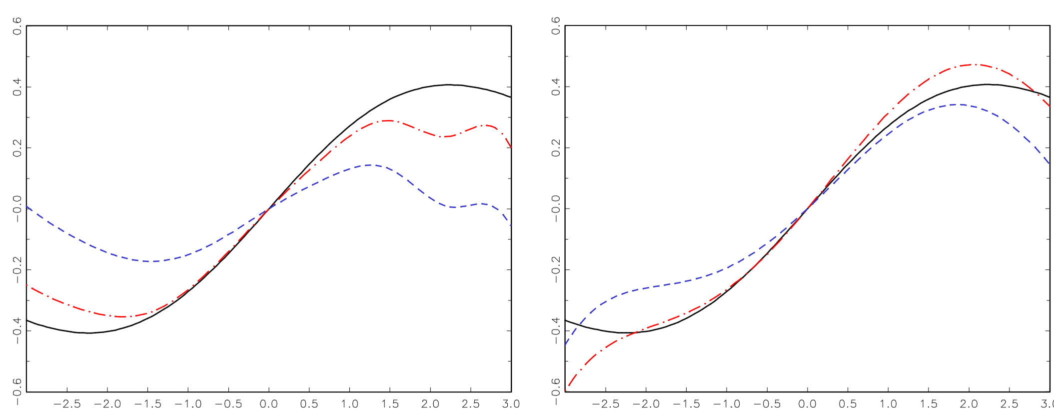

Figure C5 : Nonparametric estimation - Cubic splines (left, 4 knots) v.s. Power series (right, 4th polynomial). Solid (black) line: true; dotted (blue) line: series estimate; dashed (red) line: bias corrected.

Appendix D: Growth Regression Results

Annual Data						
	Linear			Semiparametric		
	WG	WG_c	s.e.	WG	WG_c	s.e.
ALL ($N = 73; T = 40$ from 1961 to 2000)						
$\log y_{i,t-1}$	-0.0366	-0.0365	0.0045			
$\log s$	0.0172	0.0171	0.0027	0.0164	0.0147	0.0029
$\log(n+g+\delta)$	-0.0430	-0.0426	0.0111	-0.0446	-0.0383	0.0117
R^2	0.0370					
OECD ($N = 24; T = 47$ from 1954 to 2000)						
$\log y_{i,t-1}$	-0.0495	-0.0495	0.0055			
$\log s$	0.0652	0.0652	0.0046	0.0588	0.0587	0.0049
$\log(n+g+\delta)$	-0.0231	-0.0230	0.0105	-0.0104	-0.0051	0.0115
R^2	0.1674					
Non-OECD ($N = 49; T = 40$ from 1961 to 2000)						
$\log y_{i,t-1}$	-0.0394	-0.0393	0.0059			
$\log s$	0.0127	0.0126	0.0033	0.0117	0.0113	0.0035
$\log(n+g+\delta)$	-0.0480	-0.0477	0.0149	-0.0489	-0.0439	0.0155
R^2	0.0322					

Table D1 : Growth regression results with annual panel data. WG is the within-group type estimates and WG_c is the within-group type estimates after bias correction. Standard errors (s.e.) are of the bias corrected estimates.

Every-5-year Data (without Human Capital)						
	Linear			Semiparametric		
	WG	WG$_c$	s.e.	WG	WG$_c$	s.e.
ALL ($N = 73; T = 8$ from 1965 to 2000)						
$\log y_{i,t-1}$	-0.2351	-0.2322	0.0235			
$\log s$	0.1219	0.1198	0.0157	0.1217	0.1130	0.0200
$\log(n+g+\delta)$	-0.1229	-0.1133	0.0805	-0.1389	-0.0759	0.1037
R^2	0.2308					
OECD ($N = 24; T = 9$ from 1960 to 2000)						
$\log y_{i,t-1}$	-0.2147	-0.2126	0.0316			
$\log s$	0.2360	0.2351	0.0313	0.1949	0.1695	0.0426
$\log(n+g+\delta)$	0.0259	0.0288	0.0794	0.1232	0.2137	0.1052
R^2	0.2900					
Non-OECD ($N = 49; T = 8$ from 1965 to 2000)						
$\log y_{i,t-1}$	-0.2495	-0.2462	0.0302			
$\log s$	0.1146	0.1127	0.0186	0.1144	0.1094	0.0231
$\log(n+g+\delta)$	-0.1625	-0.1543	0.1100	-0.1921	-0.1474	0.1370
R^2	0.2307					

Table D2 : Growth regression results with quintannual panel data (without Human Capital variables). WG is the within-group type estimates and WG$_c$ is the within-group type estimates after bias correction. Standard errors (s.e.) are of the bias corrected estimates.

Every-5-year Data (with Human Capital)						
	Linear			Semiparametric		
	WG	WG_c	s.e.	WG	WG_c	s.e.
ALL ($N = 73; T = 8$ from 1965 to 2000)						
$\log y_{i,t-1}$	-0.2479	-0.2441	0.0240			
$\log s$	0.1287	0.1273	0.0159	0.1251	0.1113	0.0203
$\log(n+g+\delta)$	-0.1223	-0.1132	0.0801	-0.1328	-0.0784	0.1037
$\log h$	-0.0517	-0.0540	0.0230	-0.0336	0.0159	0.0309
R^2	0.2383					
OECD ($N = 24; T = 9$ from 1960 to 2000)						
$\log y_{i,t-1}$	-0.2077	-0.2036	0.0328			
$\log s$	0.2417	0.2414	0.0321	0.2016	0.1796	0.0427
$\log(n+g+\delta)$	0.0124	0.0164	0.0810	0.0933	0.1656	0.1070
$\log h$	-0.0375	-0.0410	0.0472	-0.0987	-0.1514	0.0722
R^2	0.2924					
Non-OECD ($N = 49; T = 8$ from 1965 to 2000)						
$\log y_{i,t-1}$	-0.2587	-0.2544	0.0307			
$\log s$	0.1196	0.1185	0.0188	0.1166	0.1093	0.0234
$\log(n+g+\delta)$	-0.1532	-0.1464	0.1098	-0.1864	-0.1473	0.1372
$\log h$	-0.0432	-0.0455	0.0290	-0.0232	0.0002	0.0360
R^2	0.2357					

Table D3 : Growth regression results with quintannual panel data (with Human Capital variables). WG is the within-group type estimates and WG_c is the within-group type estimates after bias correction. Standard errors (s.e.) are of the bias corrected estimates.

Country	rank		Country	rank		Country	rank	
	w/ h	w/o h		w/ h	w/o h		w/ h	w/o h
Algeria	48	45	Iceland*	10	8	Panama	39	41
Argentina	30	31	India	56	57	Paraguay	38	38
Australia*	7	7	Indonesia	49	48	Peru	47	49
Austria*	16	15	Iran, I.R. of	44	44	Philippines	54	55
Bangladesh	66	65	Ireland*	4	3	Portugal*	26	23
Barbados	5	4	Israel	21	22	Senegal	67	67
Belgium*	14	13	Italy*	20	16	South Africa	29	28
Bolivia	57	60	Jamaica	55	56	Spain*	24	24
Brazil	34	34	Japan*	6	6	Sri Lanka	53	53
Cameroon	60	58	Jordan	50	50	Sweden*	13	18
Canada*	3	5	Kenya	64	64	Switzerland*	8	10
Chile	31	32	Korea*	25	25	Syria	51	51
Colombia	37	37	Lesotho	62	69	Thailand	45	47
Costa Rica	36	36	Malawi	68	70	Togo	70	71
Denmark*	9	9	Malaysia	33	33	Trinidad & Tob.	19	11
Dominican Rep.	46	46	Mali	71	66	Turkey*	35	35
Ecuador	52	52	Mauritius	22	21	Uganda	43	42
El Salvador	41	40	Mexico*	32	30	United Kingdom*	15	17
Finland*	17	19	Mozambique	63	61	United States*	1	1
France*	18	20	Nepal	69	63	Uruguay	28	29
Ghana	59	59	Netherlands*	12	12	Venezuela	40	39
Greece*	27	27	New Zealand*	23	26	Zambia	73	73
Guatemala	42	43	Niger	72	72	Zimbabwe	61	62
Honduras	65	68	Norway*	11	14			
Hong Kong	2	2	Pakistan	58	54			

Table D4 : Ranking of 73 countries based on estimated country specific effects. (24 OECD countries are marked with $*$.) "w/ h" means "with Human Capital variables"; "w/o h" means "without Human Capital variables."

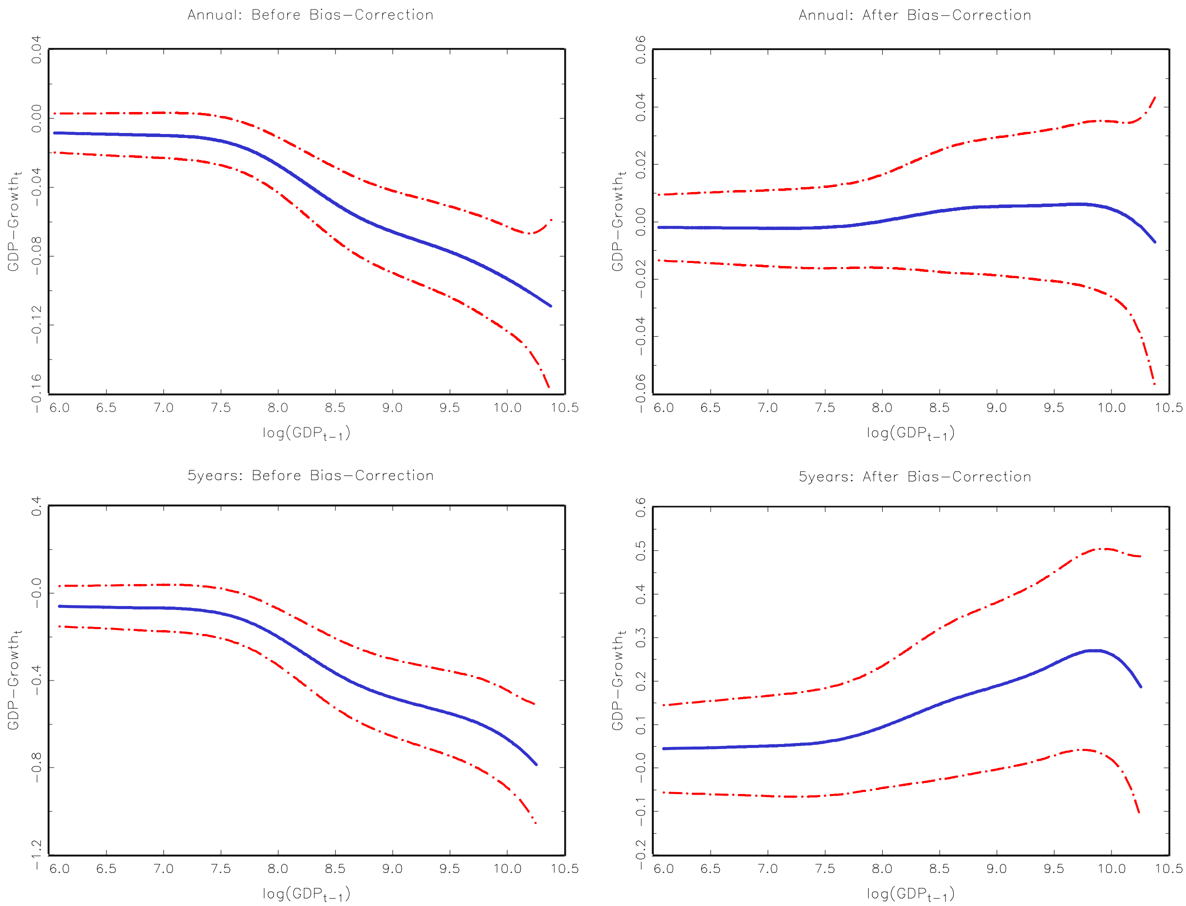

Figure D1 : GDP growth versus $\log(GDP_{t-1})$ of all 73 countries. The vertical axis represents the GDP growth after controlling saving rates, population growth rate, depreciation rate, technical growth rate and human capital (for bottom two graphs only). Top two graphs are based on the annual-frequency-panal; bottom two graphs are based on 5-year-frequency-panel with human capital variables. Bold lines are the curve estimates using cubic splines with 4 knots; dashed lines show the pointwise 5% confidence regions.

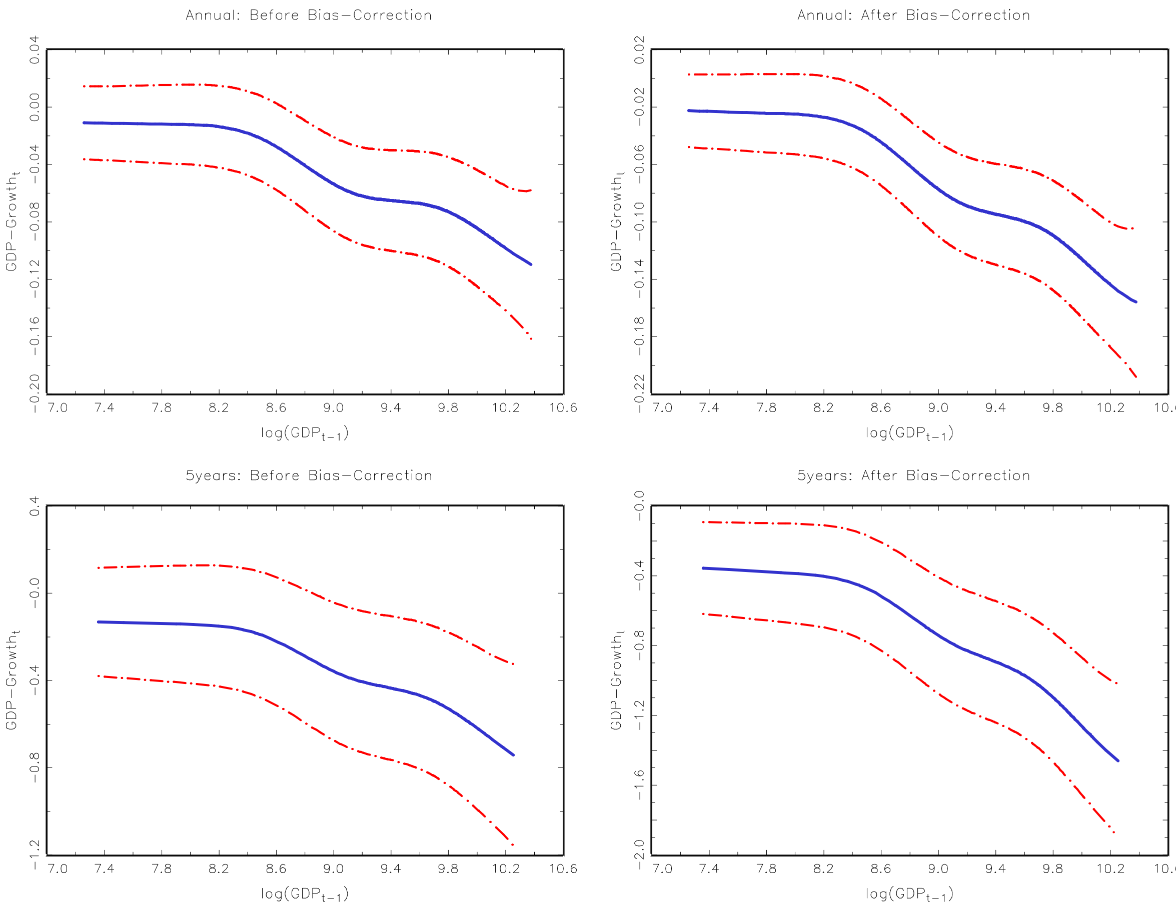

Figure D2 : GDP growth versus log (GDP_{t-1}) of the 24 OECD countries. The vertical axis represents the GDP growth after controlling saving rates, population growth rate, depreciation rate, technical growth rate and human capital (for bottom two graphs only). Top two graphs are based on the annual-frequency-panal; bottom two graphs are based on 5-year-frequency-panel with human capital variables. Bold lines are the curve estimates using cubic splines with 4 knots; dashed lines show the pointwise 5% confidence regions.

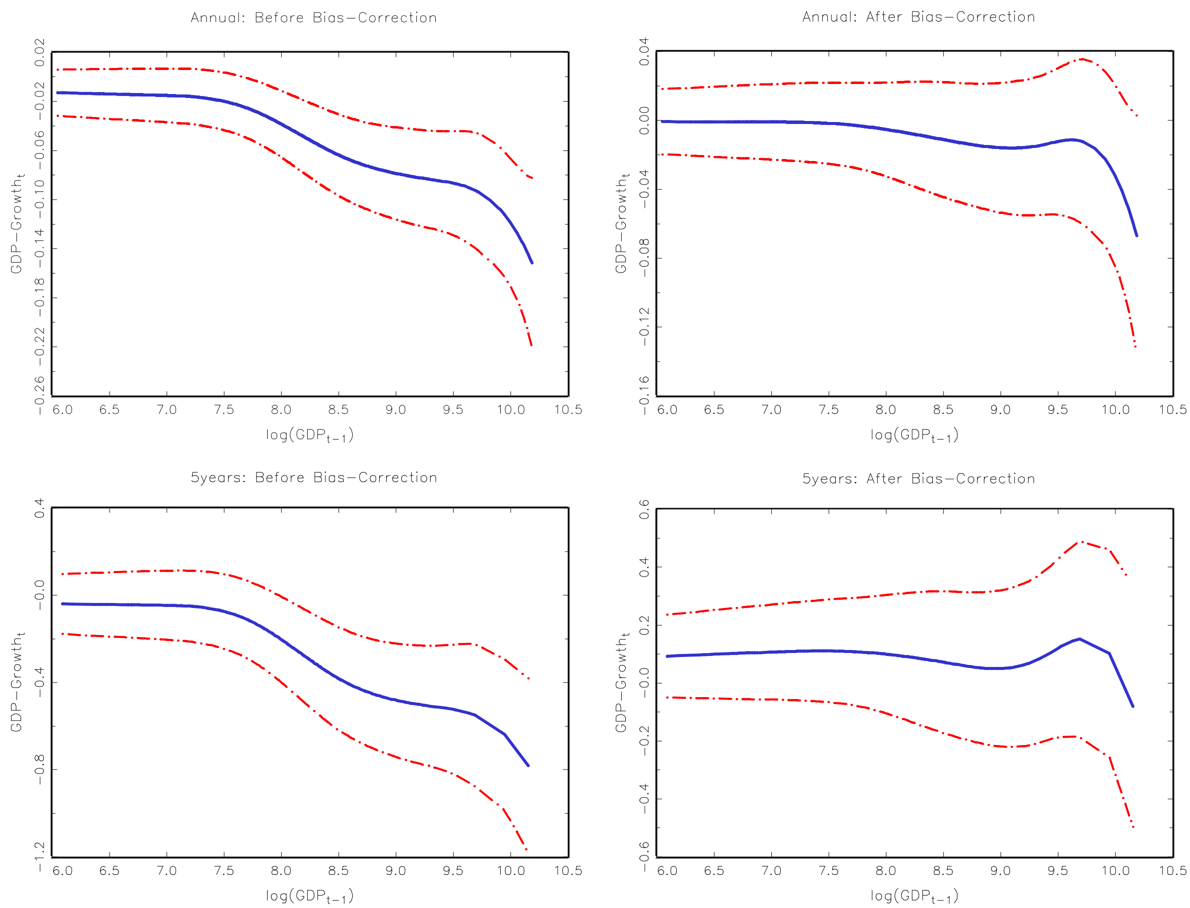

Figure D3 : GDP growth versus log (GDP_{t-1}) of 49 non-OECD countries. The vertical axis represents the GDP growth after controlling saving rates, population growth rate, depreciation rate, technical growth rate and human capital (for bottom two graphs only). Top two graphs are based on the annual-frequency-panal; bottom two graphs are based on 5-year-frequency-panel with human capital variables. Bold lines are the curve estimates using cubic splines with 4 knots; dashed lines show the pointwise 5% confidence regions.

References

Ai, C., and X. Chen (2003). Efficient estimation of models with conditional moment restrictions containing unknown functions, *Econometrica*, 71, 1795-1843.

Alvarez, J., and M. Arellano (2003). The time series and cross-section asymptotics of dynamic panel data estimators, *Econometrica*, 71, 1121-1159.

An, H.Z. and F.C. Huang (1996). The geometrical ergodicity of nonlinear autoregressive models, *Statistica Sinica*, 6, 943-956.

Andrews, D.W.K. (1991). Asymptotic normality of series estimators for nonparametric and semiparametric regression models, *Econometrica*, 59, 307-345.

Andrews, D.W.K. (1991). Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation, *Econometrica*, 59, 817-858.

Baltagi, B.H., and D. Li (2002). Series estimation of partially linear panel data models with fixed effects, *Annals of Economics and Finance*, 3, 103-116.

Barro, R.J. and J.-W. Lee (2000). International data on educational attainment updates and implications, *NBER Working Papers*, 7911, NBER.

Berk, K.N. (1974). Consistent autoregressive spectral estimates, *Annals of Statistics*, 2, 489-502.

Bernard, A.B. and S.N. Durlauf (1996). Interpreting tests of the convergence hypothesis, *Journal of Econometrics*, 71, 161-173.

Bierens, H.J. (1994). *Topics in advanced econometrics*, Cambridge: Cambridge University Press.

Billingsley, P. (1968). *Convergence of probability measures*, New York: Wiley.

Blundell, R., and J. Powell (2003). Endogeneity in Nonparametric and Semiparametric Regression Models, in *Advances in Economics and Econonometrics: Theory and Applications, Eighth World Congress*, Vol. II, M. Dewatripont, L.P. Hansen and S.J. Turnovsky, eds. Cambridge: Cambridge University Press.

Darolles, S., J.-P. Florens, and E. Renault (2003). Nonparametric instrumental regression, *mimeo*.

Davydov, Y. (1973). Mixing conditions for Markov chains, *Theory of Probability and Its Applications*, 18, 312-328.

De Jong, R.M. (2002). A note on "Convergence rates and asymptotic normality for series estimators": uniform convergence rates, *Journal of Econometrics*, 111, 1-9.

Doukhan, P. (1994). *Mixing: Properties and Examples*, New York: Springer-Verlag.

Durlauf, S.N. and P.A. Johnson (1995). Multiple regimes and cross-country growth behaviour, *Journal of Applied Econometrics*, 10, 365-84.

Fan, J., and Q. Yao (2003). *Nonlinear Time Series: Nonparametric and Parametric Methods*, New York: Springer-Verlag.

Hahn, J. and G. Kuersteiner (2002). Asymptotically unbiased inference for a dynamic panel model with fixed effects, *Econometrica*, 70, 1639-1657.

Hahn, J., and G. Kuersteiner (2004). Bias reduction for dynamic nonlinear panel models with fixed effects, *mimeo.*

Hall, P. and J.L. Horowitz (2003). Nonparametric methods for inference in the presence of instrumental variables, *working paper*, Department of Economics, Northwestern University.

Henderson, D.J. and A. Ullah (2005). A nonparametric random effects estimator, *Economics Letters*, 88, 403-407.

Islam, N. (1995). Growth empirics: A panel data approach, *Quarterly Journal of Economics*, 110, 1127-1170.

Lee, M., R. Longmire, L. Mátyás, and M. Harris (1998). Growth convergence: Some panel data evidence, *Applied Economics*, 30, 907-912.

Lee, Y. (2005). A general approach to bias correction in dynamic panels under time series misspecification, *mimeo.*

Lewis, R., and G.C. Reinsel (1985). Prediction of multivariate time-series by autoregressive model-fitting, *Journal of Multivariate Analysis*, 16, 393-411.

Li, Q. and T.J. Kniesner (2002). Nonlinearity in dynamic adjustment: Semiparametric estimation of panel labor supply, *Empirical Economics*, 27, 131-148.

Li, Q. and T. Stengos (1996). Semiparametric estimation of partially linear panel data models, *Journal of Econometrics*, 71, 389-397.

Liebscher, E. (2005). Towards a unified approach for proving geometric ergodicity and mixing properties of nonlinear autoregressive processes, *Journal of Time Series Analysis*, 26, 669-689.

Liu, Z. and T. Stengos (1999). Non-linearities in cross-country growth regressions: A semiparametric approach, *Journal of Applied Econometrics*, 14, 527-38.

Luukkonen, R. and T. Teräsvirta (1991). Testing linearity of economic time series against cyclical asymmetry, *Annales d'Economie et de Statistiques*, 20/21, 125-142.

Mankiw, N.G., D. Romer, and D.N. Weil (1992). A contribution to the empirics of economic growth, *The Quarterly Journal of Economics*, 107, 407-437.

Mundra, K. (2005). Nonparametric slope estimators for fixed-effect panel data, *mimeo*.

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

Newey, W.K. (1997). Convergence rates and asymptotic normality for series estimators, *Journal of Econometrics*, 79, 147-168.

Newey, W.K., and J.L. Powell (2003). Instrumental variables estimation for nonparametric models, *Econometrica*, 71, 1565-1578.

Newey, W.K., and K.D. West (1987). A simple positive semi-definite heteroskedasticity and autocorrelation consistent covariance matrix, *Econometrica*, 55, 703-708.

Phillips, P.C.B. and H.R. Moon (1999). Linear Regression Limit Theory for Nonstationary Panel Data, *Econometrica*, 67, 1057-1111.

Phillips, P.C.B., and D. Sul (2004). Bias in dynamic panel estimation with fixed effects, incidental trends and cross section dependence, *Cowles Foundation Discussion Paper*, No. 1438.

Porter, J.R. (1996). *Essays in Econometrics*, Ph.D. dissertation, MIT.

Robinson, P.M. (1983). Nonparametric estimators for time series, *Journal of Time Series Analysis*, 4, 185-207.

Robinson, P.M. (1988). Root-N consistent semiparametric regression, *Econometrica*, 56, 931-954.

Stone, C.J. (1982). Optimal global rates of convergence for nonparametric regression, *Annals of Statistics*, 10, 1040-1053.

Tong, H. (1990). *Non-linear Time Series: A Dynamic System Approach*, New York: Oxford University Press.

Ullah A., and N. Roy (1998). Nonparametric and semiparametric econometrics of panel data, *Handbook of Applied Economics Statistics*, 579-604.

White, H. (1984). *Asymptotic Theory for Econometricians*, Orlando: Academic Press.

White, H., and I. Domowitz (1984). Nonlinear regression with dependent observations, *Econometrica*, 52, 143-161.

Wooldridge, J.M. (2002). *Econometric Analysis of Cross Section and Panel Data*, Cambridge: MIT Press.