Nonparametric Estimation of Dynamic Panel Models

Yoonseok Lee[1]

Department of Economics

University of Michigan

First draft: November, 2005; this version: September, 2006

Abstract

This paper investigates stationary β-mixing dynamics in nonlinear panel models and 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. Convergence rates and the asymptotic distribution of the series estimator are derived, in which an asymptotic bias is present and it reduces the mean square convergence rate compared with the cross section case. 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. Using partial linear models, an empirical study on nonlinearity in the cross-country growth regression is presented. 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]An earlier version of this paper is in the second chapter of my dissertation at Yale University. I thank to Peter Phillips, Donald Andrews, Yuichi Kitamura, and seminar participants at Michigan, PSU, Rochester, UBC, UC-Irvine, UVa, UWa-Seattle, Va Tech, Yale, and the 2005 Greater New York Metropolitan Area Econometrics Colloquium in Columbia University for valuable comments. I gratefully acknowledge financial support from the Cowles Foundation under the Carl Arvid Anderson Prize. All errors are solely mine. *E-mail:* `yoolee@umich.edu`. *Address:* Department of Economics, University of Michigan, 611 Tappan Street, 365C Lorch Hall, Ann Arbor, MI 48109-1220.

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. We extend 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 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 time, T, is fixed. Both series and kernel estimations 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-effects models. All of these studies examine 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 including fixed effects, and 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 they only consider random effects. In a similar vein, Hahn and Kuersteiner (2004) examine parametric nonlinear dynamic panel models with fixed effects.

Though several studies have analyzed nonparametric and 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 paper 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 Table) and cross-firm 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 Ai and Chen (2003), Blundell and Powell (2003), Darolles, Florens and Renault (2003), Newey and Powell (2003), and Hall and Horowitz (2005) 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. 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, we develop 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 can be simply approximated by 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.[2]

Specifically, this approach follows earlier works on cross sectional series estimation by Andrews (1991a) and Newey (1997), and generalizes their asymptotic results to dynamic panels. 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), an asymptotic bias is present, which reduces the mean square convergence rate compared with the cross section case. To tackle this problem, we develop bias correction 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 nonlinearity in the cross-country growth regression 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, recent studies question the assumption of linearity in growth

[2] We expect that the within-transformed model based estimation, as long as the asymptotic bias is properly corrected, is more desirable in finite samples than the first-differenced model based estimation. 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, which is closely related to the slow rate of convergence of the IV-based nonparametric estimators.

equations and propose nonlinear alternatives that allow for multiple regimes of growth patterns among different countries. When we analyze a semiparametric dynamic panel growth equation with fixed effects using the Penn World Table, the findings suggest the presence of multiple regimes in growth patterns. In particular, before bias correction, the results exhibit the convergence for the countries in the middle to upper income range. After bias correction, however, the results support the convergence hypothesis only for the OECD countries and the countries in the upper income range.

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, which is 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 finite samples. In Section 6, an empirical study on the nonlinear cross-country growth regression is presented. Section 7 concludes this paper with some remarks. All the mathematical proofs are provided in Appendix.

2 Nonparametric Dynamic Panel Models

2.1 Fixed Effects Models

We consider a panel process $\{y_{i,t}\}$ generated from a nonlinear autoregressive model given by[3]

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

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. The realization of the initial values, $y_{i,0}$, are observed for all i. A fixed (individual) effect, μ_i, is assumed to have finite variance and to satisfy $\mathbb{E}\left(u_{i,t}|\mu_i\right) = 0$ for all i and t, but possibly correlated with $y_{i,t-1}$. Unlike a random effect, the fixed effect captures the omitted and thus unobserved cross sectional heterogeneity, and it is allowed to be correlated with the explanatory variables, $y_{i,t-1}$. On the other hand, it is assumed that $\mathbb{E}\left(u_{i,t}|y_{i,t-1}, \cdots, y_{i,0}\right) = 0$. Therefore, we suppose a common shape of the conditional mean function $m\left(\cdot\right)$ for all i but different in intercepts.

Note that the conditional mean assumption, $\mathbb{E}\left(u_{i,t}|\mu_i\right) = 0$, is important to avoid an endogeneity

[3] One could consider $y_{i,t} = m_\mu\left(y_{i,t-1}; \mu_i\right) + u_{i,t}$, but μ_i and m_μ cannot be separately identifiable without further restrictions on m_μ.

problem. Since the data generating process in (1) implies that $y_{i,t}$ is a function of both μ_i and $\{u_{i,s}\}_{s\leq t}$, the (strict) exogeneity condition $\mathbb{E}(u_{i,t}|y_{i,t-1},\cdots,y_{i,0}) = \mathbb{E}(u_{i,t}|\mu_i, u_{i,t-1}, u_{i,t-2,}\cdots) = 0$ requires that the conditional mean of $u_{i,t}$ on μ_i is zero.[4] The condition, $\mathbb{E}(u_{i,t}|\mu_i) = 0$, on the other hand, does 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, which is a key property of fixed-effects models. To make the notation as simple as possible, we let $\{u_{i,t}\}$ be an independent and identically distributed process with mean zero and finite variance. Further more, we simply assume μ_i to be independent of $u_{i,t}$ for all i and t, instead of assuming $\mathbb{E}(u_{i,t}|\mu_i) = 0$. Therefore, across i, $\{y_{i,t}\}$ is also independent with heterogeneous means. Note that the generalization to serially (weakly) 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.

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, and we will discuss such an extension in Section 4.1. The main analysis of this paper, however, focuses on the simple nonparametric model given in (1).

2.2 Beta-mixing Processes

The stability of the linear autoregressive process is determined by restricting 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)

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

[4] For example, $\{u_{i,t}\}$ needs to be a martingale difference sequence on its natural filtration $\{\mathcal{F}_{i,t}\}$ conditional on μ_i, where $\mathcal{F}_{i,t} = \sigma(u_{i,s} : s\leq t)$.

for $\tau > 0$, where $\mathcal{G}_{i,t_1}^{t_2}$ is the σ-field generated by $\{y_{i,t} : t_1 \leq t \leq t_2\}$ for each i. $\|\cdot\|_{TV}$ is the total variation[5] 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 $\{y_{i,t}\}$:

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

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 a panel process $\{y_{i,t}\}$ 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. See, Doukhan (1994) for example. Moreover, geometric ergodicity implies stationarity of the process $(\{y_{i,t}\})$ if the distribution of the initial values $(y_{i,0})$ are defined by an invariant 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, it[6] becomes a common constant shift in the distribution of the process; therefore, μ_i no longer affects the temporal dependence of $\{y_{i,t}\}$. In 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 positive density almost everywhere and an absolutely continuous marginal distribution with respect to the Lebesgue measure on $\mathbb{R}$. (iii) $u_{i,t}$ is independent of μ_i for all i and t.*

The condition E1 implies that $u_{i,t}$ is independent of $\{y_{i,s}\}_{s \leq t-1}$. We assume the density of $u_{i,t}$ to be positive almost everywhere so that we can minimize restrictions on $m(\cdot)$. We do not need this assumption for a linear autoregressive model. 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 each i and for $y \in \mathcal{Y}$, where $\mathcal{Y} \subseteq \mathbb{R}$ is the support of $\{y_{i,t}\}$.

Assumption E2 *(i) For each i and for the Borel measurable function $\phi_i : \mathcal{Y} \to \mathbb{R}$, 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 $\sup_{y:|y| \leq \overline{y}} |\phi_i(y)| < \infty$, where*

[5] 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 view of Scheffe's theorem. (cf. Liebscher, 2005, p.671)

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

$[-\overline{y}, \overline{y}] \subset \mathcal{Y}$. *(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.*

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 conditioning on μ_i and we can use well-established results for α-mixing processes. The α-mixing condition has been frequently employed in the nonparametric time series literature à la Robinson (1983). In fact, we only require a mixing condition in order to control the temporal dependence in the proof, and thus using more general mixing condition (i.e., using α-mixing condition instead of β-mixing) does not alter any implication of the study. To make this section complete, we define α-mixing coefficients of $\{y_{i,t}\}$ as

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

for $\tau > 0$ and for each i. We let $\alpha(\tau) = \sup_{1 \le i \le N} |\alpha_i(\tau)|$. Since $\alpha(\tau) \le \beta(\tau)$ for each τ, $\alpha(\tau)$ also tends to zero at an exponential rate. That is, we can write $\alpha(\tau) \le C_\alpha a^\tau$ for some a such that $0 < a < 1$ and for 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 $\{w_t\}$ is α-mixing with a mixing coefficient of $O(\tau^{-\epsilon})$ for some $\epsilon > 0$, then $\{\psi(w_t, \cdots, w_{t-k})\}$ is also α-mixing with a mixing coefficient of $O(\tau^{-\epsilon})$.*

The following mixing inequalities will be used frequently in proving the main results. 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 finite and positive constants 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

To avoid incidental parameter problem as N increases, we first need to eliminate individual effects, μ_i, in (1) by employing one of the following methods: 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

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

and the first-differencing transformation of (1) yields

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

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 it is not straightforward to estimate the unknown function $m(\cdot)$ using 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 $y_{i,t}^0 = \ell_{WT}(y_{i,t-1}, \cdots, y_{i,0}) + u_{i,t}^0$, where $\ell_{WT}(y_1, \cdots, y_T) = m(y_1) - (1/T)\sum_{s=1}^{T} m(y_s)$. Then estimating

ℓ_{WT} by kernel regression is infeasible since its dimension increases as $T \to \infty$; moreover, the regression involves an endogeneity problem because $\mathbb{E}\left(u_{i,t}^0 | y_{i,s}\right) \neq 0$ for any $0 \leq s \leq t-1$.[7] 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. Therefore, we need instrumental variables (IV) estimation for the nonparametric models (3) and (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 Ai and Chen (2003), Darolles, Florens and Renault (2003), Newey and Powell (2003), and Hall and Horowitz (2005) among others. Though these results are based on the independent cross section case, the extension to the first differenced dynamic panel in (4) can be done when T is fixed.[8] We will briefly discuss such extension in Section 4.2. One drawback of this approach is, however, slow rate of convergence, which yields efficiency loss in finite samples. Taking it into account, we instead develop a different and novel approach: the within-transform-based nonparametric estimation method based on (3).

For notational convenience, we let $m^0\left(y_{i,t}\right) = m\left(y_{i,t}\right) - (1/T)\sum_{s=1}^{T} m\left(y_{i,s}\right)$ for each i and t,[9] and rewrite the within-transformed model (3) as

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

Note, however, that $m^0\left(y_{i,t-1}\right)$ 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 $m\left(\cdot\right)$, we use series approximation as in Andrews (1991a) and Newey (1997), which approximates an unknown function $m\left(\cdot\right)$ by some linear combination of K known series functions $\{q_{Kk}\}$:

$$m\left(y\right) \approx \sum_{k=1}^{K} \theta_{Kk} q_{Kk}\left(y\right), \tag{5}$$

where $q_{Kk} : \mathcal{Y} \to \mathbb{R}$ are measurable and $\theta_{Kk} \in \mathbb{R}$ for all $k = 1, 2, \cdots, K$. "$\approx$" indicates series approxima-

[7] Since $(1/T)\sum_{s=1}^{T} m\left(y_s\right)$ can be approximated by $\mathbb{E}\left(m\left(y_t\right)\right)$ using the Law of Large Numbers, we can regard $\ell_{WT}\left(y_1, \cdots, y_T\right)$ as a demeaned form of $m\left(y_1\right)$. Then the curse of dimensionality problem disappears asymptotically. The approximation error, however, should be considered in addition to the endogeneity problem (i.e., $\mathbb{E}\left(u_{i,t}^0 | y_{i,s}\right) \neq 0$ for any $0 \leq s \leq t-1$).

[8] More careful analysis is required when both N and T are large. We leave this case as a future study.

[9] In what follows, any variables or functions with superscript ^{0} indicate that they are within-transformed.

tion; namely, it means "is approximately equal for large K." The choice of the sequence must be such that the approximation to $m(\cdot)$ improves as K gets larger, where $K = K(N,T)$ and $K \to \infty$ as $N,T \to \infty$. Using the series approximation in (5), 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 $q_{Kk}^0(y_{i,t}) = q_{Kk}(y_{i,t}) - (1/T)\sum_{s=1}^{T} q_{Kk}(y_{i,s})$.

Note that we need an additional condition for identifying μ_i and $m(\cdot)$ 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.[10]

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

In Porter (1996), it is instead assumed that[11]

$$\mathbb{E}\mu_i = 0 \tag{6}$$

or $\sum_{i=1}^{N} \mu_i = 0$ if μ_i's are regarded as fixed parameters. The condition (6) allows the level of $m(0)$ unrestricted, but normalizes the sum of individual effects μ_i to zero. Under this assumption, $m(0)$ 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. This condition implies that μ_i absorbs both homogeneous and heterogeneous intercepts, and thus it merely shifts a common function $m(\cdot)$ vertically for each i. If $m(0) \neq 0$, we can 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 (6) explains why Porter (1996) is only able to identify $m(\cdot)$ up to a constant addition.[12] For example, when we consider the first-differenced model

[10] To meet this condition, the series functions $\{q_{Kk}\}$ are chosen to satisfy $\sum_{k=1}^{K} \theta_{Kk} q_{Kk}(0) = 0$ for each K.

[11] As noted in Porter (1996), the condition (6) is weaker than $\mathbb{E}(\mu_i | \mathcal{G}_{i,t-1}) = 0$, where $\mathcal{G}_{i,t} = \sigma\left(\{y_{i,s}\}_{s \leq t}\right)$, that assumes away any potential correlation between individual effects and regressors. Thus, under $\mathbb{E}(\mu_i | \mathcal{G}_{i,t-1}) = 0$, heterogeneity bias is no longer an issue. This is the situation of random-effects models.

[12] 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

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 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).

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.[13] 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 $\mathcal{Y}_c = \{y : |y| \leq \overline{y}\} \subset \mathcal{Y}$ 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 $y_{i,t-1}$ 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) A sequence of trimming functions $\{\lambda(y_{i,t})\}$ are defined as $\lambda(y_{i,t}) = 1\{y_{i,t} \in \mathcal{Y}_c\}$ for some compact $Y_c \subset Y$, where $1\{\cdot\}$ is the binary indicator function.

Definition TR along with properly chosen series functions, such as power series or splines,[14] guarantees that $\lambda(y) q_{Kk}(y)$ are uniformly bounded over a bounded subset $\mathcal{Y}_c$.[15] 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 $\{y_{i,t}\}$

(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).

[13] 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 the case of the autoregressive model (1) because 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.

[14] 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. Hermite polynomial is an example of orthogonal polynomial. 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

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

where $K = 1 + r + L$; $(z)_+ = z$ if $z > 0$ and zero otherwise. For example, $r = 3$ for cubic splines. Note that we will omit the case $k = 0$ since $m(0) = 0$ is assumed.

[15] 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.

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 (Assumption W1), 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 then defined 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 $\mathcal{Y}_c$.

3.2 Regularity conditions

In this subsection, we list and discuss regularity conditions on which we base all the main 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 Table) and cross-firm studies.[16]

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

[16] Of course, when T is large and N is small, the dynamic panel model becomes Vector Autoregressive (VAR) model with parameter restrictions.

Assumption W1 *(i) For every K, there exist positive integers N^* and T^* such that for all $N \geq N^*$ and $T \geq T^*$, the $NT \times K$ vector $\left(g_K^0(y_{1,0}), \cdots, g_K^0(y_{N,T-1})\right)'$ is of full column rank K almost surely. (ii) 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 bounded largest eigenvalue, where all the elements of $\mathbb{E}g_K(y)$ are finite. (iii) For every K, there is a sequence of $\zeta_0(K)$ 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-(iii) seems stronger than Newey (1997), who assumes $\zeta_{0*}^2(K) K/NT \to 0$, where $\zeta_{0*}(K)$ is the uniform bound of the norm 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 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 $k = 1, 2, \cdots, K$ from 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 $\sum_{\tau=1}^{\infty} \alpha(\tau)^{1-2/\nu_g} < \infty$ is still satisfied for some $\nu_g > 4$. More precisely, we can have an alternative condition to Assumption W1 as follows.

Assumption W1b *(i) For every K, there exist positive integers N^* and T^* such that for all $N \geq N^*$ and $T \geq T^*$, the $NT \times K$ vector $\left(g_K^0(y_{1,0}), \cdots, g_K^0(y_{N,T-1})\right)'$ is of full column rank K almost surely. (ii) 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 bounded largest eigenvalue, where all the elements of $\mathbb{E}g_K(y)$ are finite. (iii) For every K and for some $\nu_g > 4$, there is a sequence of $\zeta_{0\nu}(K)$ satisfying $\zeta_{0\nu}(K) \geq \max_{1 \leq k \leq K} \mathbb{E}\,|g_{Kk}(y)|^{\nu_g/2}$ and $K = K(N,T)$ such that $\zeta_{0\nu}(K)^{2/\nu_g} K^2/NT \to \infty$ as $N, T \to \infty$.*

Using either of the conditions, W1 or W1b, 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 W1b. Note that, unlike Robinson (1983), Assumption W1 implies Assumption W1b 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* $\theta_K \in \mathbb{R}^K$ *and a constant* $\delta > 0$ *satisfying* $\sup_{y \in \mathcal{Y}_c} \left| m(y) - g_K(y)' \theta_K \right| = O\left(K^{-\delta}\right)$ *for every* K.

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. In Assumption W2, 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 $m(\cdot)$ 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, $\delta\,(= D)$ corresponds to the smoothness of $m(y)$ and the following condition can replace Assumption W2. Assumption W2b is intuitively more appealing in that the smoother $m(y)$, the easier it is to approximate it.

Assumption W2b *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,*

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

as $N, T \to \infty$, *where* $P(y)$ *denotes the cumulative distribution function of* $y_{i,t}$.[17]

[17] 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).

Theorem 3.1 implies that the probability limit of $\int_{y\in\mathcal{Y}_c}\left[\widehat{m}\left(y\right)-m\left(y\right)\right]^2 dP\left(y\right)$ is zero since $K^{-2\delta}\to 0$ and $\zeta_0^2\left(K\right)K/NT\to 0$. For the mean square convergence rate (9), the first term, K/NT, essentially corresponds to the convergence rate of the variance, whereas the remaining terms, $K^{-2\delta}$ and $\zeta_0^2\left(K\right)K/NT$, correspond to the convergence rate of the bias. The third term, $\zeta_0^2\left(K\right)K/NT$, 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 $\zeta_0\left(K\right)$ is a nondecreasing function of K.

If we assume $g_K\left(\cdot\right)$ and $m\left(\cdot\right)$ are differentiable up to D-th order as in Assumption W2′, and if we introduce $\zeta_D\left(K\right)\geq \sup_{y\in\mathcal{Y}_c}\max_{1\leq k\leq K}\max_{s\leq D}\left|d^s g_{Kk}\left(y\right)/dy^s\right|$, which is assumed to be larger than $O\left(K^{-1/2}\right)$ and to exist, then we have a uniform convergence rate of $\widehat{m}\left(y\right)$ given by

$$\sup_{y\in\mathcal{Y}_c}\max_{s\leq D}\left|d^s\left(\widehat{m}\left(y\right)-m\left(y\right)\right)/dy^s\right|=O_p\left(K^{1/2}\zeta_D\left(K\right)\left[\zeta_0\left(K\right)K^{1/2}/\sqrt{NT}+K^{-\delta}\right]\right).$$

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 for the *i.i.d.* cross section case under stronger conditions. 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 $m\left(\cdot\right)$ as follows. Note that "$\to_d$" means convergence in distribution; $\left\|B\right\|=\left(B'B\right)^{1/2}$ if B is a vector and $\left\|B\right\|=\left(tr\left(B'B\right)\right)^{1/2}$ if B is a matrix, where $tr\left(\cdot\right)$ is the trace operator.

Theorem 3.2 (Asymptotic normality) *Let $\Phi_K=\sum_{j=0}^{\infty}cov\left(g_K\left(y_{i,t+j}\right),u_{i,t}\right)$ satisfy $\left\|\Phi_K\right\|<\infty$ for each K. If Assumptions NT, E1, E2, W1 and W2 are satisfied and $\sqrt{NT}K^{-\delta}\to 0$,*[18] *then as $N,T\to\infty$*

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

for $y\in\mathcal{Y}_c$, where $v\left(y,K,N,T\right)=\sigma^2 g_K\left(y\right)'\Gamma_K^{-1}g_K\left(y\right)/NT$ and $b_K\left(y\right)=g_K\left(y\right)'\Gamma_K^{-1}\Phi_K$. The asymptotic normality still holds using a consistent estimator $\widehat{v}\left(y,K,N,T\right)=\widehat{\sigma}^2 g_K\left(y\right)'\widehat{\Gamma}_K^{-1}g_K\left(y\right)/NT$, where[19] *$\widehat{\Gamma}_K=\left(1/NT\right)\sum_{i=1}^{N}\sum_{t=1}^{T}g_K^0\left(y_{i,t}\right)g_K^0\left(y_{i,t}\right)'$ and $\widehat{\sigma}^2=\left(1/NT\right)\sum_{i=1}^{N}\sum_{t=1}^{T}\left(y_{i,t}^0-\widehat{m}^0\left(y_{i,t-1}\right)\right)^2$.*

[18] 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 $K=K\left(N,T\right)$ is chosen to satisfy reasonably small rate with respect to N and T, e.g., $K=O\left(\left(NT\right)^{1/6}\right)$, 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 2.3.4.

[19] For obtaining $\widehat{\Gamma}_K$ and $\widehat{\sigma}^2$, we can normalize them using $1/\left(NT-N-K\right)$ by adjusting the degrees of freedom.

The pointwise asymptotic distribution result in Theorem 3.2 is similar to the *i.i.d.* cross section cases as in Andrews (1991a) 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 $v(y,K,N,T)^{-1/2} = \sqrt{NT}\left(\sigma^2 g_K(y)' \Gamma_K^{-1} g_K(y)\right)^{-1/2}$. For example, since we assume the smallest eigenvalue of Γ_K is bounded away from zero and its largest eigenvalue is bounded for every K, if we simply let Γ_K be the identity matrix I_K, then the rate of convergence is given by $\sqrt{NT/K}$.

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* $\widehat{u}_{i,t}^0 = y_{i,t}^0 - \widehat{m}^0(y_{i,t-1})$. *The 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 $b_K(y) = g_K(y)' \Gamma_K^{-1} \Phi_K$ as shown in Theorem 3.2, Theorem 3.3 follows by consistently estimating $b_K(y)$ with $\widehat{b}_K(y) = g_K(y)' \widehat{\Gamma}_K^{-1} \widehat{\Phi}_K$. In Appendix A.1, it is shown that $\left\|\widehat{\Gamma}_K^{-1} - \Gamma_K^{-1}\right\| = o_p(1)$ and

$$\widehat{\Phi}_K = \frac{1}{NT}\sum_{i=1}^{N}\sum_{j=0}^{J} w(j,J) \sum_{t=1}^{T-j} g_K(y_{i,t+j})\widehat{u}_{i,t}^0 \tag{11}$$

is a consistent estimator for the one-side long-run covariance Φ_K so that $\left\|\widehat{\Phi}_K - \Phi_K\right\| = o_p(1)$ 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 (1991b) 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 for J to grow slower than $T^{1/3}$. We can modify the simple weight function $w(j,J)=(1-j/(J+1))$ using kernel functions as in Andrews (1991b).

Remark 3.4 (Determining the order of K) In nonparametric analysis, the smoothing parameters are conventionally 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). As usual, 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.[20]

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 rate condition 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[21] 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 $K<O\left((NT)^{1/6}\right)$ for B-splines.

[20] Recall that $\zeta_0^2(K)K/NT$ dominates K/NT.

[21] 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 $K^4/NT+\sqrt{NT}/K^{\delta}\to 0$.

Remark 3.5 (Testing linearity) When we approximate the unknown function $m(\cdot)$ using (orthogonal) polynomials, testing linearity for $m(\cdot)$ becomes straightforward. Since $q_{K1}(y) = y$ in this case, testing the linearity is identical to testing $\theta_{Kk} = 0$ for all $k = 2, 3, \cdots, K$ in (5), where $K \to \infty$ as $N, T \to \infty$. Therefore, we can construct a Wald statistic as $W_{K-1} = \left(R_K \widetilde{\theta}_K\right)' \left[R_K \widehat{\Gamma}_K^{-1} R_K'\right]^{-1} \left(R_K \widetilde{\theta}_K\right) / \left(\widehat{\sigma}^2/NT\right)$, where $R_K = [0_{K\times 1}; I_{K-1}]$ is a $(K-1) \times K$ matrix and $\widetilde{\theta}_K = \left(\widetilde{\theta}_{K1}, \widetilde{\theta}_{K2}, \cdots, \widetilde{\theta}_{KK}\right)'$, $\widehat{\Gamma}_K$ and $\widehat{\sigma}^2$ are defined as in Theorem 3.3.[22] By the similar argument as in the proof of Theorem 3.3 in Appendix A.3, it is following that $W_{K-1} \to_d \lim_{K\to\infty} \mathcal{X}^2_{K-1}$ as $N, T \to \infty$. The critical values can be found by applying the well-known normal approximation results such as $X^2_K(\vartheta) \approx (1/2)\left\{Z(\vartheta) + \sqrt{2K-1}\right\}^2$ (Fisher, 1925) or $X^2_K(\vartheta) \approx K\left\{1 - (2/9K) + Z(\vartheta)\sqrt{2/9K}\right\}^3$ (Wilson and Hilferty, 1931) for large K, where $X^2_K(\vartheta)$ and $Z(\vartheta)$ denote the 100ϑ percentage point of the $\mathcal{X}^2$ distribution with K degrees of freedom and the standard normal distribution, respectively. On the other hand, if we approximate $m(\cdot)$ using other functionals, we need to consider more general nonparametric specification tests such as the general likelihood ratio test for nonparametric models (e.g., Fan and et al., 2001). We leave further details as a future project.

4 Extensions

4.1 Partial linear models

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

$$y_{i,t} = m(y_{i,t-1}) + \gamma' x_{i,t} + \mu_i + u_{i,t}, \tag{12}$$

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: $y_t = \rho y_{t-1} + m(x_t) + u_t$. The purpose of such a model is to control out the effects from x_t nonparametrically. In (12), on the other hand, we are rather interested in the partial effects of exogenous variable $x_{i,t}$ to $y_{i,t}$, whereas 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 $m(y_{i,t-1}) = \rho y_{i,t-1}$. In some cases, moreover, we are more interested in uncovering the unknown shape of dynamics in $y_{i,t}$ (i.e., $m(\cdot)$) with controlling other characteristics $x_{i,t}$ linearly. Such examples can be found in semiparametric cross-country

[22] That is $\widetilde{\theta}_K = \widehat{\theta}_K + (1/T)\left(\sum_{i=1}^N \sum_{t=1}^T g_K^0(y_{i,t}) g_K^0(y_{i,t})'\right)^{-1} \sum_{i=1}^N \sum_{j=0}^J \sum_{t=1}^{T-j} \left(1 - \frac{j}{J+1}\right) g_K(y_{i,t+j}) \widehat{u}^0_{i,t}$.

growth regressions as in Liu and Stengos (1999).

We could relax the linear part in (12) so that it is also fully nonparametric in both $y_{i,t-1}$ and $x_{i,t}$ as

$$y_{i,t} = m\left(y_{i,t-1}, x_{i,t}\right) + \mu_i + u_{i,t}. \tag{13}$$

However, such generalization is limited in empirical studies because of the curse of dimensionality. Therefore, we need more restriction on $m(\cdot,\cdot)$, such as single index or additivity, to reduce the dimension of $m(\cdot,\cdot)$. For example, a number of studies on semiparametric estimation 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 $m_x : \mathbb{R}^r \to \mathbb{R}^1$, and use marginal integration to estimate the additive nonparametric components.[23] If we assume similar conditions on the series approximation for both m_x and m_y as in the previous section, we can derive the asymptotic distribution of the series estimator for (13). Note that the conditions for m_x should correspond to those in Porter (1996) since $x_{i,t}$ is strictly exogenous. The following condition guarantees that the autoregressive process $\{y_{i,t}\}$ with exogenous variables satisfying (13) is stationary and mixing as in Section 2. This condition also ensures the stationarity and mixing for $\{y_{i,t}\}$ in (12) since the partial linear model (12) is a special case of (13). Though we provide general conditions for (13) in the following assumption, we will mainly examine its special case, the partial linear specification (12), since it is more relevant in empirical studies. We let $\phi_i(z) = \mu_i + m(z)$, where $z = (y_1, ..., y_p; x_1, ..., x_q) \in \mathbb{R}^p \times \mathbb{R}^{qr}$.

Assumption E3 (Stability condition) *(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 positive density almost everywhere and an absolutely continuous marginal distribution with respect to the Lebesgue measure on $\mathbb{R}$. (iii) For each i, there exist constants $c_i > 0$, $\overline{z} > 0$ and $a_i, \cdots, a_p \geq 0$, and a locally bounded measurable function $f : \mathbb{R}^r \to [0,\infty)$ such that $|\phi_i(z)| \leq \sum_{j=1}^p a_j |y_j| + \sum_{h=1}^q f(x_h) - c_i$ if $\|z\|_\infty > \overline{z}$ and $\sup_{z:\|z\|_\infty \leq \overline{z}} |\phi_i(z)| < \infty$, where $w^p - a_1 w^{p-1} - \cdots - a_p \neq 0$ for $|w| \geq 1$ and $\|z\|_\infty = \max\{|y_1|, ..., |y_p|, |x_1|, ..., |x_q|\}$. (iv) $\mathbb{E}f(x_{i,1}) + \mathbb{E}|u_{i,1}|^\nu < \infty$ 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 an 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 assumption 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 presumes 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.

[23] 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.

For example, Doukhan (1994) considers stationary Markov $\{x_{i,t}\}$, in fact. We conjecture that conditional on μ_i, provided $\{x_{i,t}\}$ is mixing with mixing coefficients decaying faster than or as fast as those of $\{y_{i,t}\}$ for each i, 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 $\xi(\cdot,\cdot) \in \mathbb{R}^r$ is a measurable function and $x_{i,t}^*$ is a stable stochastic process independent of μ_i.

In the partial linear model (12), we first eliminate fixed effects, μ_i, by the within transformation:

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

Note that (12) 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 (12). To show this, we take conditional expectations on (12) to have

$$\mathbb{E}\left(y_{i,t}|y_{i,t-1}\right) = m\left(y_{i,t}\right) + \gamma'\mathbb{E}\left(x_{i,t}|y_{i,t-1}\right) + \mathbb{E}\left(\mu_i|y_{i,t-1}\right) \tag{14}$$

since $\mathbb{E}\left(u_{i,t}|y_{i,t-1}\right) = 0$ by assumption. We subtract (14) from (12), and obtain

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

in which we cannot eliminate $\left[\mu_i - \mathbb{E}\left(\mu_i|y_{i,t-1}\right)\right]$ either by the within transformation or the first-differencing transformation. This is because $\left[\mu_i - \mathbb{E}\left(\mu_i|y_{i,t-1}\right)\right]$ is a function of not only μ_i but also $y_{i,t-1}$, which still depends on the time index t. This illustration suggests that we need to eliminate fixed effects at the very first stage. Porter (1996), for example, 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.

For notational convenience, we introduce the following vectors and matrices. We define an $NT \times K$ vector $\mathbf{g}_K^0 = \left(g_K^0\left(y_{1,0}\right), \cdots, g_K^0\left(y_{N,T-1}\right)\right)'$, and $NT \times 1$ vectors $\mathbf{y}^0 = \left(y_{1,1}, \cdots, y_{N,T}\right)'$, $\mathbf{x}^0 = \left(x_{1,1}^0, \cdots, x_{N,T}^0\right)'$ and $\widehat{\mathbf{m}}^0 = \left(\widehat{m}^0\left(y_{1,0}\right), \cdots, \widehat{m}^0\left(y_{N,T-1}\right)\right)'$. We also define $NT \times NT$ matrices such as $M_x = I_{NT} - \mathbf{x}^0\left(\mathbf{x}^{0\prime}\mathbf{x}^0\right)^{-1}\mathbf{x}^{0\prime}$ 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 that both $\mathbf{x}^{0\prime}\mathbf{x}^0$ and $\mathbf{g}_K^{0\prime}\mathbf{g}_K^0$ are nonsingular (at least almost surely). 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 $\left(g_K\left(y_{i,t-1}\right)', x_{i,t}'\right)'$, whose smallest eigenvalue is bounded above zero and the largest eigenvalue is bounded for every K. We decompose it into

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

conformably as $\left(g_K\left(y_{i,t-1}\right)', x_{i,t}'\right)'$. Recall that the conditional variance of $g_K\left(y_{i,t-1}\right)$ 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 $g_K\left(y_{i,t-1}\right)$ is defined as $\Sigma_{xx\cdot g} = \Sigma_{xx} - \Sigma_{xg}\Sigma_{gg}^{-1}\Sigma_{gx}$. We summarize the additional conditions in the following assumption.

Assumption W3 *(i)* $\mathbf{x}^0$ *is of a full column rank* r*. (ii) For every* K*,* Σ *has the smallest eigenvalue bounded above zero and the bounded largest eigenvalue.*

We now derive the asymptotic distribution of the partial linear model estimators in (12).

Theorem 4.1 (Partial linear model) *Under Assumptions NT, E3, W1, W2 and W3, as* $N, T \to \infty$

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

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

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

where $v_x\left(y, K, N, T\right) = \sigma^2 g_K\left(y\right)' \Sigma_{gg\cdot x}^{-1} g_K\left(y\right)/NT$ *and* $V_x = \sigma^2\Sigma_{xx\cdot g}^{-1}$*. The results still hold after replacing* $v_x\left(y, K, N, T\right)$ *and* V_x *with their consistent estimators,* $\widehat{v}_x\left(y, K, N, T\right) = \widehat{\sigma}^2 g_K\left(y\right)' \widehat{\Sigma}_{gg\cdot x}^{-1} g_K\left(y\right)$ *and* $\widehat{V}_x = \widehat{\sigma}^2\widehat{\Sigma}_{xx\cdot g}^{-1}$*, where* $\widehat{\sigma}^2 = \left(1/NT\right)\sum_{i=1}^N\sum_{t=1}^T\left(y_{i,t}^0 - \widehat{m}^0\left(y_{i,t-1}\right) - \widehat{\gamma}' x_{i,t}^0\right)^2$ *and the conditional variance estimators,* $\widehat{\Sigma}_{gg\cdot x}^{-1}$ *and* $\widehat{\Sigma}_{xx\cdot g}^{-1}$*, are obtained from* $\widehat{\Sigma} = \left(1/NT\right)\sum_{i=1}^N\sum_{t=1}^T\left(g_K^0\left(y_{i,t-1}\right)', x_{i,t}^{0\prime}\right)'\left(g_K^0\left(y_{i,t-1}\right)', x_{i,t}^{0\prime}\right)$.

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 $\widehat{m}\left(y\right)$. As in Theorem 3.3, bias correction can be conducted as follows.

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

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

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

$$\widetilde{m}(y) = \widehat{m}(y) + \frac{1}{T} g_K(y)' \widehat{\Sigma}_{gg \cdot x}^{-1} \widehat{\Phi}_K = g_K(y)' \left(\mathbf{g}_K^{0\prime} M_x \mathbf{g}_K^0 \right)^{-1} \left\{ \mathbf{g}_K^{0\prime} M_x \mathbf{y}^0 + \frac{1}{T} \widehat{\Phi}_K \right\},$$
$$\widetilde{\gamma} = \widehat{\gamma} - \frac{1}{T} \widehat{\Sigma}_{xx \cdot g}^{-1} \widehat{\Sigma}_{xg} \widehat{\Sigma}_{gg}^{-1} \widehat{\Phi}_K = \left(\mathbf{x}^{0\prime} M_g \mathbf{x}^0 \right)^{-1} \mathbf{x}^{0\prime} \left\{ M_g \mathbf{y}^0 - \frac{1}{T} \mathbf{g}_K^0 \left(\mathbf{g}_K^{0\prime} \mathbf{g}_K^0 \right)^{-1} \widehat{\Phi}_K \right\}$$

and $\widehat{\Phi}_K$ *is defined as in (11).*

4.2 Two stage instrumental variables estimator

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

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

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 $\mathbb{E}(x_{i,t} | z_{i,t}) \neq 0$ for all i and t. In dynamic panel regressions, instruments conventionally consist of the lagged observations of $y_{i,t-s}$ for $s \geq 2$. Using instruments $z_{i,t}$, 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 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 could generate a bias problem.[24] In this case, the within-transformation-based method seems to be more appropriate.

The second problem is restoring the estimator of the original function $\widehat{m}(\cdot)$ from $\widehat{\ell}(\cdot)$. The identification problem in a fixed-effect model is closely discussed in Porter (1996), where he uses the partial integration

[24] 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 (2003), 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, e.g., $(\log T)^2 / N \to 0$.

method as in Newey (1994): to restore the original function $m(y_1)$, he integrates $\ell(y_1, y_2)$ 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: $\ell(y_1, y_2) = m(y_1) - m(y_2)$. Porter (1996) employs the structural information by imposing additional restrictions $\ell(y_1, y_2) = -\ell(y_2, y_1)$ and $\ell(y, y) = 0$. This method, however, can only identify $m(\cdot)$ up to a constant addition by $\mathbb{E}m(\cdot)$. We suggest an alternative method that, under the normalization condition ID (i.e., $m(0) = 0$), which is introduced in Section 3.1, we can easily restore $\widehat{m}(\cdot)$ from $\widehat{\ell}(\cdot)$ using the additive structure, $\ell(y_1, y_2) = m(y_1) - m(y_2)$. That is, $\widehat{m}(y)$ can be obtained from $\widehat{\ell}(y_1, y_2)$ by letting the second argument be zero because $\ell(y_1, 0) = m(y_1) - m(0) = m(y_1)$.

Remark 4.3 (Identification) The identification of ℓ from the conditional expectation, $\mathbb{E}(\Delta y|z) = \mathbb{E}(\ell(y_1, y_2)|z)$, can be discussed in a more general setup as follows. The conditional expectation of the first-differenced model given instruments z yields

$$\eta(z) = \mathbb{E}(\Delta y|z) = \int \ell(y_1, y_2) P(dy_1|z), \tag{15}$$

where $y_2 \in z$ and $P(y_1|z)$ is the conditional distribution of y_1 given z. As noted in Newey and Powell (2003), $\eta(z)$ and $P(dy_1|z)$ are identified because they are functionals of the distribution function for the observations (y_1, y_2, z). Identification of $\ell(y_1, y_2)$ from the integral equation (15), however, is not straightforward. We need the following condition to solve this problem:

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

This completeness condition guarantees the uniqueness of the solution $\ell(y_1, y_2)$ of the integral equation (15) if its existence is presumed. Another important condition is the continuity assumption to avoid the ill-posed inverse problem in estimation. As noted by Newey and Powell (2003) or Florens (2003), if $\widehat{\ell}$, the estimator of ℓ, is not continuous in $\widehat{\eta}$ and $\widehat{P}$, which are the estimators of η and P, then the consistency of $\widehat{\ell}$ does not follow from the consistency of $\widehat{\eta}$ and $\widehat{P}$. One solution to avoid ill-posed problem is to assume that m (or ℓ) belongs to a compact subset of a normed set of functions and to restrict the estimator $\widehat{m}$ (or $\widehat{\ell}$) to lie in this compact set. Since integration is a continuous mapping, compactness implies that inverse is continuous. We also employ this approach to eliminate the ill-posed inverse problem. In our case, since we consider nonparametric estimation only over a compact subset $\mathcal{Y}_c$ of the support of y, restricting m and $\widehat{m}$ to be in a compact set is not difficult. As noted in Gallant and Nychka (1987), and Ai and Chen (2003),

when the infinite dimensional parameter space $\mathcal{M}_c$, such that $m \in \mathcal{M}_c$, consists of bounded and smooth functions, then there exists a metric $\|\cdot\|_c$ such that $\mathcal{M}_c$ is compact under $\|\cdot\|_c$. Note that Assumption E2-(i) implies m is bounded over $\mathcal{Y}_c$ and Assumption W2 (or W2$'$) implies m is smooth up to order D on $\mathcal{Y}_c$. For further discussions on this type of regularization, see Tikhonov et al. (1995), Ai and Chen (2003), Blundell and Powell (2003), Newey and Powell (2003), and references therein. More general treatment using a ridge-type regularization can be found in Darolles et al. (2003), Florens (2003), and Hall and Horowitz (2005) among others.

We now 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. If we use the series approximation as (5), we have

$$m(y_1) - m(y_2) \approx \sum_{k=1}^{K} \theta_{Kk} \left[g_{Kk}(y_1) - g_{Kk}(y_2) \right], \tag{16}$$

and

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

In the first stage, we estimate the conditional expectation by any nonparametric estimation method to have $\widehat{\mathbb{E}}\left[g_{Kk}(y_{i,t-1}) - g_{Kk}(y_{i,t-2}) \,|z_{i,t}\right] \equiv \Delta\widehat{g}_{Kk}(z_{i,t})$. In the second stage, if define a $K \times 1$ vector $\Delta\widehat{g}_K(z_{i,t}) = \left(\Delta\widehat{g}_{K1}(z_{i,t}), \cdots, \Delta\widehat{g}_{KK}(z_{i,t})\right)'$, we can estimate $\theta_K = (\theta_{K1}, \cdots, \theta_{KK})'$ by solving the minimization problem:[25]

$$\widehat{\theta}_K = \arg\min_{\theta_K} \sum_{i=1}^{N} \left(\Delta y_i - \Delta\widehat{g}_K(z_i)\,\theta_K\right)' H \left(\Delta y_i - \Delta\widehat{g}_K(z_i)\,\theta_K\right), \tag{17}$$

where $\Delta y_i = (\Delta y_{i,1}, \cdots, \Delta y_{i,T})'$, $\Delta\widehat{g}_K(z_i) = \left(\Delta\widehat{g}_K(z_{i,1}), \cdots, \Delta\widehat{g}_K(z_{i,T})\right)'$, and H is a $T \times T$ matrix given by

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

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

[25] 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.

H 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 (17) is therefore a simple GLS problem with a known covariance matrix. The pointwise consistency of $\widehat{m}(\cdot)$ for large N can be derived similarly as Newey and Powell (2003), and Ai and Chen (2003) under proper regularity conditions and a metric. Their 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 for consistency are discussed in Appendix B when $N \to \infty$.[26]

Remark 4.4 (Partial linear models) 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

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

where $x_{i,t}$ and $w_{i,t}$ do not need to be exogenous. In order to prevent any problem with large dimension, we simply let $m(\cdot,\cdot)$ 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 panel models with fixed effects of five different dynamic structures given by

$$\begin{aligned}
&\text{(M1)}: \quad y_{i,t} = \{0.6 y_{i,t-1}\} + \mu_i + u_{i,t} \\
&\text{(M2)}: \quad y_{i,t} = \{\exp(y_{i,t-1}) / (1 + \exp(y_{i,t-1})) - 0.5\} + \mu_i + u_{i,t} \\
&\text{(M3)}: \quad y_{i,t} = \{\ln(|y_{i,t-1} - 1| + 1)\, sgn(y_{i,t-1} - 1) + \ln 2\} + \mu_i + u_{i,t}
\end{aligned}$$

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

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

We estimate $\widehat{h}_{Kk}(z) = \widehat{\mathbb{E}}[h_{Kk}(y_1,y_2)|z]$ using any nonparametric method and conduct series estimation such as $\widehat{\theta}_K = \arg\min_{\theta_K} \sum_{i=1}^{N} \left(\Delta y_i - \Delta \widehat{h}_K(z_i)'\theta_K\right)' H \left(\Delta y_i - \Delta \widehat{h}_K(z_i)'\theta_K\right)$, 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 still has an identification problem of restoring $\widehat{m}$ from $\widehat{\ell}$, whereas the first approach in (16) does not have such a problem.

$$\text{(M4)}: \quad y_{i,t} = \{0.6y_{i,t-1} - 0.9y_{i,t-1}/\left(1+\exp\left(y_{i,t-1}-2.5\right)\right)\} + \mu_i + u_{i,t}$$
$$\text{(M5)}: \quad y_{i,t} = \left\{0.3y_{i,t-1}\exp\left(-0.1y_{i,t-1}^2\right)\right\} + \mu_i + u_{i,t}$$

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 is also investigated in Ai and Chen (2003) in the cross section case. 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 as in the conventional discrete 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).

Samples of $(N,T) = (100, 50)$ data points were generated, 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. For series estimation, we use both power series and cubic splines. Orthogonal (Hermite) polynomial is used for the power series. 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 use four knots since the other four terms are cubic polynomials, $\left(1, y, y^2, y^3\right)$. We do not consider different locations of the knots and simply use equispaced knots.

The simulation results are displayed in Table 5.1. The integrated mean square errors (IMSE) and the integrated mean absolute errors (IMAE) are calculated over $y \in \mathcal{Y}_c = [-3, 3]$ for each case. The IMSE is computed as[27] $\sum_{j=0}^{121}(0.05)\left\{(1/1000)\sum_{r=1}^{1000}\left(m\left(-3+0.05j\right) - \widehat{m}_r\left(-3+0.05j\right)\right)^2\right\}$, where m is the true nonlinear function and $\widehat{m}_r$ is the estimate in rth replication. The IMAE is similarly obtained by $\sum_{j=0}^{121}(0.05)\left\{(1/1000)\sum_{r=1}^{1000}\left|m\left(-3+0.05j\right) - \widehat{m}_r\left(-3+0.05j\right)\right|\right\}$. 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.

[27] These discrete expressions are borrowed from Ai and Chen (2003).

TABLE 5.1

SIMULATION RESULT[a]

	Cubic Splines				Power Series			
	IMSE		IMAE		IMSE		IMAE	
	original	bias-c	original	bias-c	original	bias-c	original	bias-c
M1	0.9228	0.4842	0.7959	0.5374	0.0450	0.0355	0.1441	0.1375
M2	0.1010	0.0428	0.2472	0.1429	0.0458	0.0415	0.1416	0.1179
M3	0.6174	0.3163	0.6318	0.4220	0.1771	0.0784	0.1733	0.0708
M4	0.1930	0.1134	0.3509	0.2505	0.1341	0.0480	0.1120	0.0451
M5	0.1111	0.0444	0.2681	0.1559	0.1387	0.0427	0.1162	0.0387

[a]Within-group series estimation over 1,000 iterations with (N,T) = (100,50). "original" displays IMSE and IMAE before bias correction; "bias-c" displays IMSE and IMAE after bias correction.

6 Application: Cross-Country Growth Regression

Most of the empirical studies examining cross-country growth equations are based on the assumption that there is a common linear dynamic 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 group-specific 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. We also take the semiparametric approach in this section.

Liu and Stengos (1999) use the pooled cross-country data. As pointed out in Islam (1995), one drawback of the conventional single cross section 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). Moreover, such an approach will reduce the possible variable omission bias in the cross-country regression because unobserved country-specific effects can be captured in the fixed effects. Similarly as in 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:[28]

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

where $y_{i,t}$ is the GDP per capita of country i at year t, the log-difference $\Delta \ln y_{i,t} = \ln y_{i,t} - \ln y_{i,t-1}$ is the growth rate, $s_{i,t}$ is the savings rate. $n_{i,t}$ 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)[29], 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 (2000)[30] for 115 countries in every five years from 1960 to 2000. μ_i is a country-specific fixed effect and $u_{i,t}$ is simply assumed *i.i.d.*; we do not consider cross-country dependence. Recall that in the growth equation (18), when $m\left(\ln y_{i,t-1}\right) = \alpha_1 \ln y_{i,t-1}$, namely

$$\Delta \ln y_{i,t} = \alpha_1 \ln y_{i,t-1} + \alpha_2 \ln s_{i,t} + \alpha_3 \ln\left(n_{i,t} + g + \delta\right) + \alpha_4 \ln h_{i,t} + \mu_i + u_{i,t}, \tag{19}$$

it supports the growth convergence hypothesis if $\alpha_1 < 0$. Analogously, if the slope of $m\left(\cdot\right)$ 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 D.4 in Appendix D. OECD member countries 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[31] 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. In the conventional growth analysis, annual data is not used because they are more likely affected by short-run factors. It is therefore difficult to recover long-run dynamics from high frequency data. Taking

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

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

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

[31] 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, and Slovak Republic.

it into account, we also choose to employ a five year interval, which is also 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). 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 analysis. For the annual data, we use from the year 1960 to 2000 for the entire countries and the non-OECD countries, whereas we use from the year 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. For the analysis with five-year time intervals, savings rates and population growth variables are averaged over each five-year interval.

The estimation results are provided in Tables D.1 to D.3 and Figures D.1 to D.3 in Appendix D. The tables display estimation results both for the linear specification (19) below and for the partial linear specification (18). For the nonparametric part, we use cubic splines with four knots. For the linear regressions (19), the results are close to Islam (1995) and all the estimates for α_1 support the growth convergence hypothesis with 1% significance level. The bias correction, which is proposed in Lee (2005), does not change the results much. For the partial linear regressions (18), 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 (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 (19) except for non-OECD countries.

Figures D.1 to D.3 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 country-specific fixed effects and the other variables – savings rate 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,[32] only the OECD countries reveal the convergence patterns. (See Figure D.2) For the entire 73 countries and for the non-OECD countries, we hardly can find the convergence except for the very upper income range. (See Figure D.1 and D.3)

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-

[32] 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 (19).

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 D.4 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 much lower ranks; but Ireland and Barbados are ranked in the top tier.

7 Concluding Remarks

This paper calls into question the linear autoregressive structure in dynamic panels. In most cases, we do not have prior information on the functional form of the regression model, so we employ nonparametric estimation without imposing any structural assumptions. For the nonparametric estimation, fixed effects are eliminated by the within transformation and series approximation is employed. No instrument variables are required since the endogeneity bias is directly corrected. Based on the stationary β-mixing condition, 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.

Even though we allow for a general functional form in the regression, we still suppose the additive separable structure so that both individual effects and the error term are not included in the unknown function m. Nonseparability can be considered with a cost of more restrictions on the unknown function m, which is required for a proper identification. See, for example, Chesher (2003), Altonji and Matzkin (2005) and references therein for the discussion of nonparametric identification in non-dynamic setup.

Comparing with series approximation, the kernel-based estimation (or the local linear estimation) seems more appealing when we are interested in a local properties of the unknown function. However, it is required that most of the observations $\{y_{i,t}\}$ should be concentrated around a particular interesting point for all i and t; otherwise, we cannot linearly approximate the unknown function with a negligible approximation error for each observation. More precisely, we Taylor expand $m(\cdot)$ around $y \in \mathbb{R}$ to obtain

$$y_{i,t} = m(y_{i,t-1}) + \mu_i + u_{i,t} = m(y) + (y_{i,t-1} - y) m'(y) + \mu_i + v_{i,t},$$

where $m'(y) = dm(y)/dy$, $v_{i,t} = u_{i,t} + \sum_{j=2}^{\infty} \frac{1}{j!} (y_{i,t-1} - y)^j m^{(j)}(y)$, and $m^{(j)}$ is the j-th derivative of m. For each y, we can eliminate the intercept term, $m(y)+\mu_i$, using the first-differencing transformation or the within-transformation. Once we estimate $m'(y)$, we can recover the estimate for $m(y)$ under Assumption

ID or the condition (6). In order to employ the conventional nonparametric analysis as in Ullah and Roy (1998), however, we need that the residual term $r_{i,t}(y) \equiv \sum_{j=2}^{\infty} \frac{1}{j!} (y_{i,t-1} - y)^j m^{(j)}(y)$ disappears fast enough for all i and t as $N, T \to \infty$. It is possible (e.g., $r_{i,t}(y) \leq O_{a.s.}\left(h^2\right)$) when $|y_{i,t-1} - y| \leq O_{a.s.}(h)$ for all i and t with $h = h_{N,T} \to 0$ as $N, T \to \infty$ and $m^{(j)}(y)$'s are uniformly bounded over y and j. In static panel models, such conditions can be easily obtained by imposing a (small) compact support of the explanatory variables. Unfortunately, it is not feasible for the dynamic panel case. A closer investigation is in progress by the author and the statistical properties of kernel-based estimator are expounded in detail in a companion paper.

Several topics need to be explored further. For example, the asymptotic properties of the two stage IV estimator, in comparison with the WG estimator, need to be studied when both N and T are large. Analyzing nonseparable models, especially when the unknown function is not smooth everywhere, is another interesting topic because it could cover many economic models such as (smoothed) discrete choice models. Finally, allowing cross section dependence is relevant in practical implementation. For example, a common factor structure can be assumed as in Phillips and Sul (2004) instead of *i.i.d.* errors; imposing a specific spatial dependence structure using spatial econometrics is another way.

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 3.3. 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 (1991a) 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 $\widehat{\Gamma}_K = (1/NT)\sum_{i=1}^{N}\sum_{t=1}^{T} g_K^0(y_{i,t}) g_K^0(y_{i,t})'$ for every K. Note that Assumption W1-(i) ensures that $\widehat{\Gamma}_K$ is nonsingular almost surely. In what follows, therefore, we simply assume that $\widehat{\Gamma}_K$ is invertible.[33] 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 $\Gamma_K = \mathbb{E}\underline{g}_K(y_{i,t})\underline{g}_K(y_{i,t})'$ and $\widehat{\Gamma}_K = (1/NT)\sum_{i=1}^{N}\sum_{t=1}^{T} g_K^0(y_{i,t}) g_K^0(y_{i,t})'$.*

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 than 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 $\widehat{\theta}_K$.

[33] More precisely, we define an indicator function $\mathbb{I}_{N,T}$ for the smallest eigenvalue of $\widehat{\Gamma}_K$ being away from zero, so $\mathbb{P}(\mathbb{I}_{N,T} = 1) \to 1$ as $N, T \to \infty$. Whenever $\widehat{\Gamma}_K$ appears in the proof, we then need to consider $\mathbb{I}_{N,T}\widehat{\Gamma}_K$ instead of $\widehat{\Gamma}_K$ as in Newey (1997). It only makes the notation more complicated without affecting the asymptotic results. We thus assume $\widehat{\Gamma}_K$ is invertible; in other words, the $NT \times K$ vector $\left(g_K^0(y_{1,0}), \cdots, g_K^0(y_{N,T-1}) \right)'$ is of full column rank K for every K. Since we are considering orthogonal basis functions, in fact, this assumption does not lose generality.

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

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

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

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 of σ^2 and Φ_K. These results justify the bias correction formula in Theorem 3.3.

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 $\sum_{j=1}^{J} |w \left(j, J \right)| \leq C_w J$ 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 $\mathbb{E}\underline{g}_{Kj}\left(y_{i,t}\right)\underline{g}_{Kk}\left(y_{i,t}\right) = \Gamma_{K,jk}$ implies[34]

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

by Assumption W1. Secondly, using Proposition 2.4, under Assumptions E1, E2 and W1[35],

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

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

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

[34] Similarly as in Newey (1997), we can derive the sharper upper bound $\zeta_0^2\left(K\right)K^2/nT$ by assuming $\Gamma_K = I_K$. Letting Γ_K 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 and its largest eigenvalue is also bounded. We, however, do not pursue this sharper bound since the covariance term, $A_2\left(N,T,K\right)$, cannot achieve this sharper bound.

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

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

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. We will use this logic in what follows.

using the Kronecker lemma[36]. 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

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

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

Proof of Lemma A1.3 We decompose

$$\begin{aligned}
\sum_{i=1}^{N} \sum_{t=1}^{T} g_K^0(y_{i,t-1}) g_K^0(y_{i,t-1})' = & \sum_{i=1}^{N} \sum_{t=1}^{T} \underline{g}_K(y_{i,t-1}) \underline{g}_K(y_{i,t-1})' \\
& - \frac{1}{T} \sum_{i=1}^{N} \sum_{t=1}^{T} \underline{g}_K(y_{i,t-1}) \sum_{s=1}^{T} \underline{g}_K(y_{i,s-1})'.
\end{aligned}$$

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 a 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 (Assumption W1-(ii)). The second term converges to zero in probability if $\zeta_0^4(K) K^2/NT \to 0$. This

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

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

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

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 $\zeta_0^4(K)K^2/NT \to 0$. ■

Proof of Lemma A1.4 First note that

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

where

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

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

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

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

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

since $u_{i,t}$ is independent of μ_i. Therefore,

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

and it follows that

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

since $\sum_{\tau=1}^{\infty}\alpha(\tau)^{1-2/r}<\infty$, $\mathbb{E}|u_{i,t}|^r<\infty$ for $r>2$ by assumption E1 ($\mathbb{E}|u_{i,t}|^\nu<\infty$ for $\nu>4$) and $\zeta_0^2(K)K/NT\leq\zeta_0^4(K)K^2/NT\to 0$.

For the second result, we observe

$$\mathbb{E}\left\|\frac{1}{NT^2}\sum_{i=1}^{N}\sum_{t=1}^{T}\underline{g}_K(y_{i,t-1})\sum_{s=1}^{T}u_{i,s}\right\|^2=\frac{1}{NT^4}\sum_{k=1}^{K}\mathbb{E}\left(\sum_{t=1}^{T}\underline{g}_{Kk}(y_{i,t-1})\sum_{s=1}^{T}u_{i,s}\right)^2$$

$$\begin{aligned}&\leq \frac{1}{NT^4}\sum_{k=1}^{K}\mathbb{E}\left(T\zeta_0(K)\sum_{t=1}^{T}u_{i,t}\right)^2\\&= \sigma^2\zeta_0^2(K)K/NT=O\left(\zeta_0^2(K)K/NT\right)\to 0.\ \blacksquare\end{aligned}$$

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

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

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

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

Proof of Lemma A1.6 We first define 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 $|Z_{i,t}|\leq\|\rho\|\left\|\Gamma_K^{-1/2}\right\|\left\|\underline{g}(y_{i,t-1})\right\||u_{i,t}|\leq C_1K^{1/2}\zeta_0(K)|u_{i,t}|$ for some constant $C_1>0$ since $\|\rho\|=1$ and by Assumption W1. Thus, for some $r=\nu/2>2$, $\mathbb{E}\left|Z_{i,t}^2\right|^r=\mathbb{E}|Z_{i,t}|^{2r}\leq C_2K^r\zeta_0^{2r}(K)\,\mathbb{E}|u_{i,t}|^{2r}=O\left(K^r\zeta_0^{2r}(K)\right)$ 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

$$\begin{aligned}\mathbb{E}\left\|\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}Z_{i,t}^{2}-1\right\|^{2} &\leq \frac{1}{NT}\left\{\mathbb{E}\left(Z_{i,1}^{2}-1\right)^{2}+2\sum_{\tau=1}^{T-1}\left(1-\frac{\tau}{T}\right)\left|cov\left(Z_{i,1}^{2},Z_{i,\tau+1}^{2}\right)\right|\right\}\\ &\leq \frac{1}{NT}\left\{\mathbb{E}\left|Z_{i,t}^{2}\right|^{2}+16\sum_{\tau=1}^{\infty}\alpha\left(\tau\right)^{1-2/r}\left(\mathbb{E}\left|Z_{i,t}^{2}\right|^{r}\right)^{2/r}\right\}\\ &\leq O\left(K^{2}\zeta_{0}^{4}\left(K\right)/NT\right)\end{aligned}$$

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

$$\begin{aligned}\left|cov\left(Z_{i,1}^{2},Z_{i,\tau+1}^{2}\right)\right| &\leq \mathbb{E}\left|cov\left(Z_{i,1}^{2},Z_{i,\tau+1}^{2}|\mu_{i}\right)\right|\leq 8\alpha\left(\tau\right)^{1-2/r}\left(\mathbb{E}\left[\left|Z_{i,t}^{2}\right|^{r}|\mu_{i}\right]\right)^{2/r}\\ &\leq 8\alpha\left(\tau\right)^{1-2/r}\left(C_{2}K^{r}\zeta_{0}^{2r}\left(K\right)\mathbb{E}\left[\mathbb{E}\left|u_{i,t}\right|^{2r}|\mu_{i}\right]\right)^{2/r}\\ &= 8\alpha\left(\tau\right)^{1-2/r}\left(C_{2}K^{r}\zeta_{0}^{2r}\left(K\right)\mathbb{E}\left|u_{i,t}\right|^{2r}\right)^{2/r}\end{aligned}$$

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). We adopt their idea to derive the asymptotic normality of $\{Z_{i,t}\}$ as follows[37]. 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,

$$\begin{aligned}\mathbb{E}\left(Z_{t}^{2}\mathbf{1}\left\{|Z_{t}|>\epsilon_{1}\sqrt{T}\right\}\right) &= \mathbb{E}\left(Z_{t}^{2}\mathbf{1}\left\{Z_{t}^{2}>\epsilon_{1}^{2}T\right\}\right)\\ &\leq \mathbb{E}\left(Z_{i,t}^{2}\mathbf{1}\left\{Z_{i,t}^{2}>NT\epsilon_{2}\right\}\right)\\ &\leq \left[\mathbb{E}\left(Z_{i,t}^{4}\right)\right]^{1/2}\left[\mathbb{P}\left\{Z_{i,t}^{2}>NT\epsilon_{2}\right\}\right]^{1/2}\\ &\leq \left[\mathbb{E}\left(Z_{i,t}^{4}\right)\right]^{1/2}\left[\mathbb{E}\left(Z_{i,t}^{4}\right)/\left(NT\epsilon_{2}\right)^{2}\right]^{1/2}\\ &= \mathbb{E}\left|Z_{i,t}^{2}\right|^{2}/\left(NT\epsilon_{2}\right)=C_{3}K^{2}\zeta_{0}^{4}\left(K\right)\mathbb{E}\left|u_{i,t}\right|^{4}/NT\to 0\end{aligned}$$

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

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

as $N,T\to\infty$. ■

Proof of Lemma A1.7 Note that

$$\begin{aligned}&\mathbb{E}\left\|\frac{1}{\sqrt{NT^{3}}}\sum_{i=1}^{N}\sum_{t=1}^{T}\underline{g}_{K}\left(y_{i,t-1}\right)\sum_{s=1}^{T}u_{i,s}-\sqrt{\kappa}\Phi_{K}\right\|^{2}\\ \leq\ &\frac{N}{T}\mathbb{E}\left\|\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\underline{g}_{K}\left(y_{i,t-1}\right)\sum_{s=1}^{T}u_{i,s}-\Phi_{K}\right\|^{2}+\left(\sqrt{\frac{N}{T}}-\sqrt{\kappa}\right)^{2}\left\|\Phi_{K}\right\|^{2}.\end{aligned}$$

[37] Alternatively, we can directly apply Theorem 3 of Phillips and Moon (1999) since we already show $\mathbb{E}\left\|\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}Z_{i,t}^{2}-1\right\|^{2}\leq O\left(K^{2}\zeta_{0}^{4}\left(K\right)/NT\right)=o\left(1\right)$.

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

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

by the stationarity. Therefore,

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

By the Kronecker lemma, $B_2\left(N,T,K\right)$ is negligible for large T since Φ_K is defined as $\Phi_K = \sum_{j=1}^{\infty} cov\left(g_K\left(y_{i,t-1}\right),u_{i,t-j}\right)$. Moreover, if we define a $K\times 1$ vector

$$\Psi_K \equiv \mathbb{E}\left(\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\underline{g}_K\left(y_{i,t-1}\right)\sum_{s=1}^{T}u_{i,s}\right)$$

and its kth element as Ψ_{Kk}, then we have

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

Note that (i)

$$\sum_{t=1}^{T}\mathbb{E}\left(\underline{g}_{Kk}\left(y_{i,t-1}\right)\sum_{s=1}^{T}u_{i,s}-\Psi_{Kk}\right)^{2}\leq\sum_{t=1}^{T}\mathbb{E}\left(\underline{g}_{Kk}\left(y_{i,t-1}\right)\sum_{s=1}^{T}u_{i,s}\right)^{2}\leq T^{2}\sigma^{2}\zeta_{0}^{2}\left(K\right);$$

(ii) 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; and (iii) $u_{i,t}$ is *i.i.d.* with $\mathbb{E}\left|u_{i,t}\right|^{4}<\infty$. From (i), (ii) and (iii), it follows that

$$\begin{aligned}B_{1}\left(N,T,K\right)\;\leq\;&\frac{1}{NT^{2}}\sum_{k=1}^{K}T^{2}\sigma^{2}\zeta_{0}^{2}\left(K\right)+\frac{1}{NT^{2}}\sum_{k=1}^{K}\sum_{t=1}^{T-1}\sum_{\tau=1}^{T-t-1}C_{1}\alpha\left(\tau\right)^{1/2}\zeta_{0}^{2}\left(K\right)T\\ &+\frac{1}{NT^{2}}\sum_{k=1}^{K}\sum_{t=2}^{T}\sum_{\tau=1}^{t-1}C_{2}\alpha\left(\tau\right)^{1/2}\zeta_{0}^{2}\left(K\right)T\\ \leq\;&O\left(\zeta_{0}^{2}\left(K\right)K/N\right),\end{aligned}$$

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 $\lim_{N,T\to\infty}N/T=\kappa$ with $0<\kappa<\infty$ implies that $\left(\zeta_{0}^{2}\left(K\right)K/N\right)^{2}=\left(\zeta_{0}^{4}\left(K\right)K^{2}/NT\right)\left(T/N\right)\to0$ 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

$$\begin{aligned}&\left\|\frac{1}{\sqrt{NT}}\sum_{i=1}^{N}\sum_{t=1}^{T}g_{K}\left(y_{i,t-1}\right)\left\{m\left(y_{i,t-1}\right)-g_{K}\left(y_{i,t-1}\right)'\theta_{K}\right\}\right\|\\ \leq\;&\left\|\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}g_{K}\left(y_{i,t-1}\right)\right\|\left(C_{m}K^{-\delta}\sqrt{NT}\right).\end{aligned}$$

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 $N, T \to \infty$ since $\|\mathbb{E}g_K(y)\|$ is finite. More precisely,

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

Therefore,

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

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

The second result can be derived similarly since

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

Proof of Lemma A1.9 We have

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

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

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

where the first term is simply $(1/NT)\,\mathbb{E}\left(u_{i,t}^2-\sigma^2\right)^2=(1/NT)\left\{\mathbb{E}u_{i,t}^4-\sigma^4\right\}=O\left(1/NT\right)$ 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

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

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

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

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

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

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

Proof of Lemma A1.10 We first decompose

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

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

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

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

$$\mathbb{P}\left(\left\|\sum_{j=0}^{J}\frac{w\left(j,J\right)}{N\left(T-j\right)}\sum_{i=1}^{n}\sum_{t=1}^{T-j}\left(g_K\left(y_{i,t+j}\right)u_{i,t}-\mathbb{E}g_K\left(y_{i,t+j}\right)u_{i,t}\right)\right\|>\varepsilon\right)$$

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

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

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

for some constant $0<C_\Phi<\infty$ since

$$
\mathbb{E}\left(g_{Kk}\left(y_{i,t+j}\right)u_{i,t}-\mathbb{E}g_{Kk}\left(y_{i,t+j}\right)u_{i,t}\right)^2\leq\mathbb{E}\left(g_{Kk}\left(y_{i,t+j}\right)u_{i,t}\right)^2\leq\zeta_0^2\left(K\right)\mathbb{E}u_{i,t}^2=\zeta_0^2\left(K\right)\sigma^2
$$

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 $J=J\left(T\right)=O\left(T^{1/3}\right)$ since

$$
\begin{aligned}
&\sum_{j=1}^{J}\left(\frac{C_wJ}{\varepsilon}\right)^2\mathbb{E}\left\|\frac{1}{N\left(T-j\right)}\sum_{i=1}^{n}\sum_{t=1}^{T-j}\left(g_K\left(y_{i,t+j}\right)u_{i,t}-\mathbb{E}g_K\left(y_{i,t+j}\right)u_{i,t}\right)\right\|^2\\
\leq\;&\frac{C_w^2C_\Phi}{\varepsilon^2}\cdot\frac{\zeta_0^2\left(K\right)K}{N}\left(\sum_{j=1}^{J}\frac{J^2}{T-j}\right)\\
\leq\;&C_\varepsilon\left(\frac{\zeta_0^2\left(K\right)K}{N}\right)\left(\frac{J^3}{T}\right)\to0
\end{aligned}
$$

for some constant $0<C_\varepsilon<\infty$, as $N,T\to\infty$ with $N/T\to\kappa\in\left(0,\infty\right)$. Note that since N and T are comparable, $\zeta_0^2\left(K\right)K/N$ is close to $\zeta_0^2\left(K\right)K/\sqrt{NT}\to0$ 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

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

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

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

with the same argument on J. ■

A.3 Within-group estimator

Using lemmas in Appendix A.1, we now prove the main results in Section 3.3. 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 $\mathbf{u}=(u_{1,1},\cdots,u_{N,T})'$, $\mathbf{u}^0=\left(u_{1,1}^0,\cdots,u_{N,T}^0\right)'$, $\mathbf{m}=\left(m(y_{1,0}),\cdots,m(y_{N,T-1})\right)'$ 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

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

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,

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

Therefore,

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

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

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

for $\left\|\widehat{\Gamma}_K - \Gamma_K\right\| \to_p 0$ with $\|\Gamma_K\| < \infty$ by Lemma A1.3. It follows that

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

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[38],

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

giving

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

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,

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

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

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

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

where $\widehat{\rho}=g_K\left(y\right)'\widehat{\Gamma}_K^{-1/2}/\sqrt{g_K\left(y\right)'\widehat{\Gamma}_K^{-1}g_K\left(y\right)}$. By construction, $\left\|\widehat{\rho}\right\|=1$. 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

$$\frac{\sqrt{NT}\left(\widehat{m}\left(y\right)-m\left(y\right)+\left(1/T\right)g_K\left(y\right)'\Gamma_K^{-1}\Phi_K\right)}{\sqrt{g_K\left(y\right)'\Gamma_K^{-1}g_K\left(y\right)}}$$

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

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 $\left\|\rho'\Gamma_K^{-1/2}\right\| \le \|\rho\|\left\|\Gamma_K^{-1/2}\right\| < \infty$ from Assumption W1 and

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

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

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

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

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

where the residual term is

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

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

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

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

[Step 3] We finally consider the feasible case[39] 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

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

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

where the residual term is

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

Therefore, using the proof in [Step 2], $\widetilde{\rho}' \Gamma_K^{-1/2} \underline{\mathbf{g}}_K' \mathbf{u} / \sqrt{NT} \to_d \mathcal{N}\left(0, \sigma^2\right)$. 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

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

where the first part converges in distribution to the standard normal as $N, T \to \infty$ by Theorem 3.2. For the second part, we will show that $\left\| (1/T) v(K,N,T)^{-1/2} \left(\widehat{b}_K(y) - b_K(y) \right) \right\| \to_p 0$ as $N, T \to \infty$ to complete the proof. Note that

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

$$\begin{aligned}
\leq & \frac{1}{T} \left| \frac{g_K(y)' \Gamma_K^{-1} g_K(y)}{NT} \right|^{-1/2} \left\| g_K(y)' \widehat{\Gamma}_K^{-1} \widehat{\Phi}_K - g_K(y)' \Gamma_K^{-1} \widehat{\Phi}_K \right\| \\
& + \frac{1}{T} \left| \frac{g_K(y)' \Gamma_K^{-1} g_K(y)}{NT} \right|^{-1/2} \left\| g_K(y)' \Gamma_K^{-1} \widehat{\Phi}_K - g_K(y)' \Gamma_K^{-1} \Phi_K \right\| \\
= & \sqrt{\frac{N}{T}} \left\| \frac{g_K(y)' \left(\widehat{\Gamma}_K^{-1} - \Gamma_K^{-1} \right) \widehat{\Phi}_K}{\sqrt{g_K(y)' \Gamma_K^{-1} g_K(y)}} \right\| + \sqrt{\frac{N}{T}} \left\| \frac{g_K(y)' \Gamma_K^{-1} \left(\widehat{\Phi}_K - \Phi_K \right)}{\sqrt{g_K(y)' \Gamma_K^{-1} g_K(y)}} \right\| \\
= & D_1(N,T,K) + D_2(N,T,K). \qquad \text{(a12)}
\end{aligned}$$

The second term $D_2(N,T,K)$ is simply $o(1)$ since $N/T \to \kappa < \infty$ and

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

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 $N,T \to \infty$ by Lemma A1.10. For the first term $D_1(N,T,K)$ in (a12), observe that

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

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 $N,T \to \infty$ by Lemma A1.3, the fourth norm also converges to zero in probability as $N,T \to \infty$ by Lemma A1.10, and the fifth norm is bounded by assumption. ■

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

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

and

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

Similarly as Lemma A1.3, $\left\| \widehat{\Sigma} - \Sigma \right\| \to 0$ 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 3.3, and we simply discuss the heuristic idea of the proof here. By combining the second term in (a13) and the formula (a14), we have

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

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 Theorems 3.2 and 3.3 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), and Ai and Chen (2003), we write a generalization of equation $\mathbb{E}\left(\Delta u_{i,t}|z_{i,t}\right)=0$ as

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

where $\rho\left(y_{i,t},y_{i,t-1},y_{i,t-2};m\right)=\Delta u_{i,t}=\Delta y_{i,t}-\Delta m\left(y_{i,t-1}\right)$ and $\Delta m\left(y_{i,t-1}\right)=m\left(y_{i,t-1}\right)-m\left(y_{i,t-2}\right)$. Since we approximate $m\left(y\right)\approx\sum_{k=1}^{K}\theta_{Kk}g_{Kk}\left(y\right)$ or $\Delta m\left(y\right)\approx\sum_{k=1}^{K}\theta_{Kk}\Delta g_{Kk}\left(y\right)$, the first stage series estimator (as one of the nonparametric estimation methods) of $\mathbb{E}\left[\Delta g_{Kk}\left(y\right)|z\right]$ is given by

$$\begin{aligned}
\widehat{\mathbb{E}}\left[\Delta g_{Kk}\left(y\right)|z_{i,t}\right] &\equiv \Delta\widehat{g}_{Kk}\left(z_{i,t}\right) \\
&= \varsigma_J\left(z_{i,t}\right)'\left(\sum_{j=1}^{N}\sum_{s=1}^{T}\varsigma_J\left(z_{j,s}\right)\varsigma_J\left(z_{j,s}\right)'\right)^{-1}\sum_{j=1}^{N}\sum_{s=1}^{T}\varsigma_J\left(z_{j,s}\right)\Delta g_{Kk}\left(y\right),
\end{aligned}$$

where $\varsigma_J\left(z\right)=\left(\varsigma_{J1}\left(z\right),\varsigma_{J2}\left(z\right),\cdots,\varsigma_{JJ}\left(z\right)\right)'$ denotes approximating functions for $\mathbb{E}\left[\Delta g_{Kk}\left(y\right)|z\right]$ for all $k=1,\cdots,K$.[40] It follows that θ_K can be estimated by solving the minimization problem:

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

where Δy_i, $\Delta\widehat{g}_K\left(z_i\right)$ and H are given in Section 4.2. Recall that $T\times T$ matrix H is positive definite. The nonparametric estimate is then obtained by $\widehat{m}\left(y\right)=\sum_{k=1}^{K}\widehat{\theta}_{Kk}g_{Kk}\left(y\right)$ for any $y\in\mathcal{Y}_c$. We assume the following conditions.

Assumption I1 *(i) $\{y_{i,t}\}$ satisfies the stability conditions in Section 2.2. (ii) We only consider estimating m over a nonempty compact subset $\mathcal{Y}_c$ of the support of $\{y_{i,t}\}$.*

The stationarity and mixing condition over t is only necessary when $T\to\infty$. Considering 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 *There is a metric $\left\|\cdot\right\|_c$ such that $\mathcal{M}_c$ $(m\in\mathcal{M}_c)$ is compact under $\left\|\cdot\right\|_c$ over $\mathcal{Y}_c$.*

Assumption I3 *m 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 I4 *Over $y\in\mathcal{Y}_c$ and for any $m\left(y\right)$ satisfying Assumption E2-(i), there exists a series approximation $g_K\left(y\right)'\theta_K$ such that $\left\|m\left(y\right)-g_K\left(y\right)'\theta_K\right\|_c\to 0$ as $K\to\infty$.*

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

[40] For each k, we could define different sets of approximating functions. However, it will not make any difference empirically.

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

Assumption I6 *(i) For any $b(z)$ with $\mathbb{E}(b(z))^2 < \infty$, there exist $\varsigma_J(z)$ and $\varphi \in \mathbb{R}^J$ with $\mathbb{E}\left(b(z) - \varsigma_J(z)'\varphi\right)^2 \to 0$ as $J \to \infty$, where $J/N \to 0$ if T is fixed; $J/NT \to 0$ if T tends to infinity. (ii) For every J, the $J \times J$ variance-covariance matrix of $\varsigma_J(z)$ exists, whose smallest eigenvalue is bounded away from zero and the largest eigenvalue is bounded.*

We provide the consistency result as in Newey and Powell (2003). The proof follows Theorem 4.1 of Newey and Powell (2003) with defining $Q(m) = \mathbb{E}\left(\mathbb{E}[\rho(y_1, y_2, y_3; m)|z]' H \mathbb{E}[\rho(y_1, y_2, y_3; m)|z]\right)$.

Theorem B.1 (Consistency: Newey and Powell (2003, Theorem 4.1)) *If Assumptions I1 to I6, E1 and E2 are satisfied and $N, K \to \infty$, then $\left\|\widehat{\ell}(y) - \ell(y)\right\|_c \to_p 0$ for $y \in \mathcal{Y}_c$.*

Corollary B.2 (Consistency) *Under the same condition of Theorem B.1, if Assumption ID is satisfied, $\|\widehat{m}(y) - m(y)\|_c \to_p 0$ for $y \in \mathcal{Y}_c$ as $N, K \to \infty$.*

Notice that Theorem B.1 and Corollary B.2 hold 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. We leave the limit properties with large N and T under Assumption NT as a topic for future research.

Appendix C: Simulation Results

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

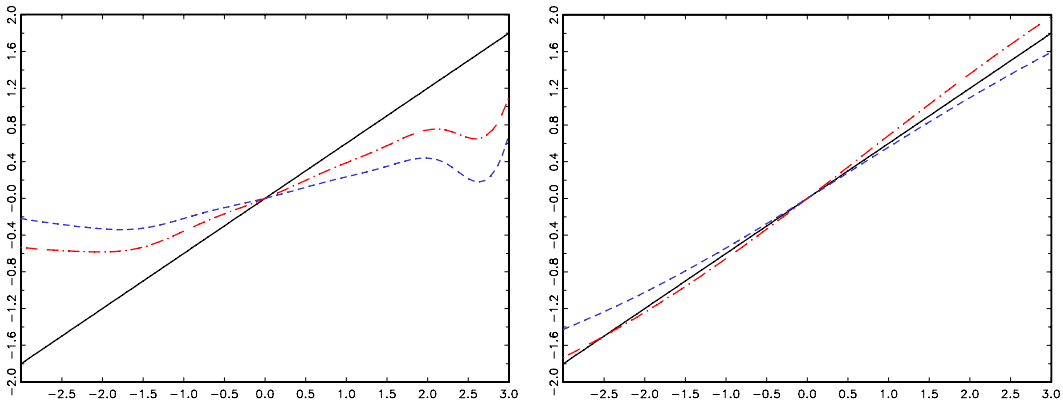

FIGURE C.1 : Nonparametric estimation - Cubic splines (left, 4 knots) *v.s.* Power series (right, 4th polynomial).[41]

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

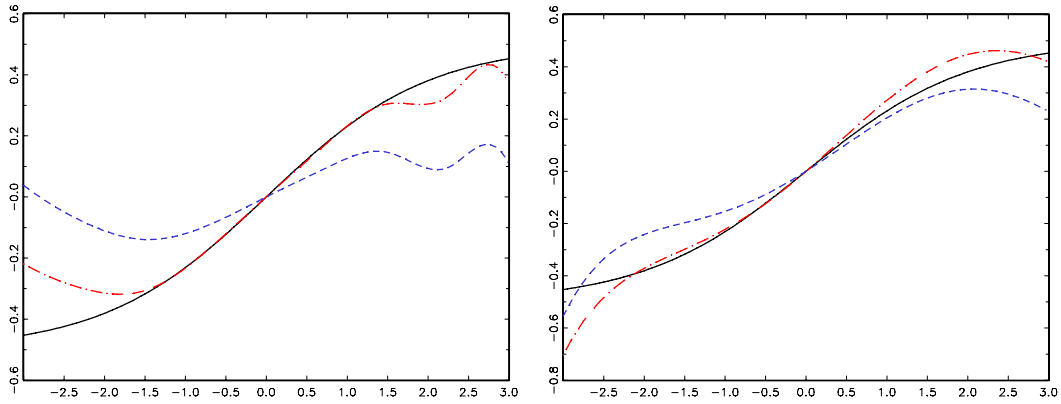

FIGURE C.2 : Nonparametric estimation - Cubic splines (left, 4 knots) *v.s.* Power series (right, 4th polynomial).

[41] For each graph in FIGURE C.1 to C.5, solid line is the true; dotted (- - -) line is series estimate before bias correction; dashed ($- \cdot -\cdot$) line is series estimate after bias correction. Samples of $(N, T) = (100, 50)$ data points are used and the estimate values are averaged over 1000 replications.

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

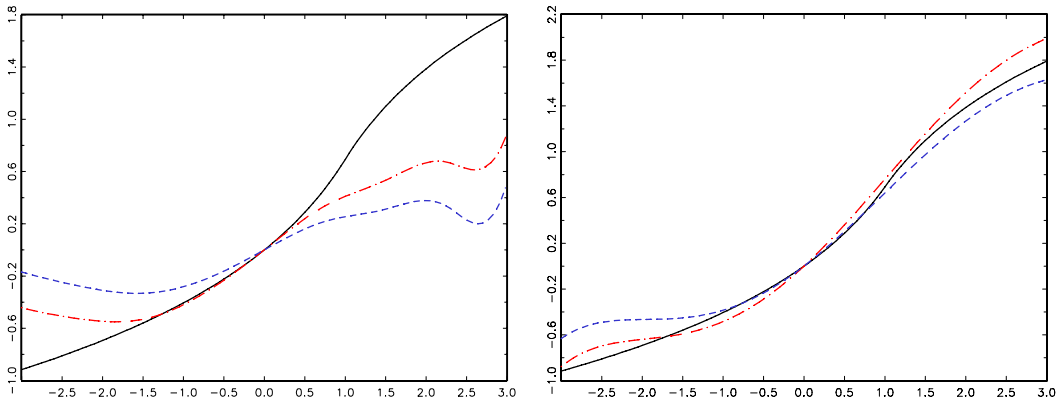

FIGURE C.3 : Nonparametric estimation - Cubic splines (left, 4 knots) *v.s.* Power series (right, 4th polynomial).

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

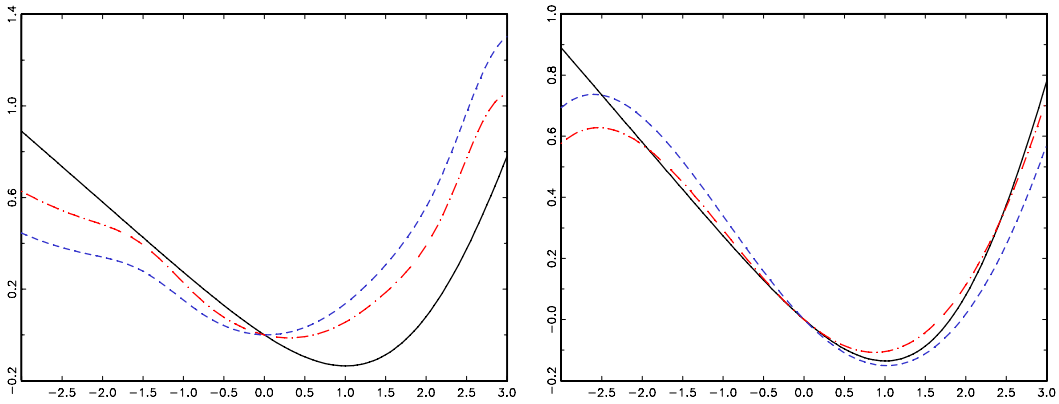

FIGURE C.4 : Nonparametric estimation - Cubic splines (left, 4 knots) *v.s.* Power series (right, 4th polynomial).

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

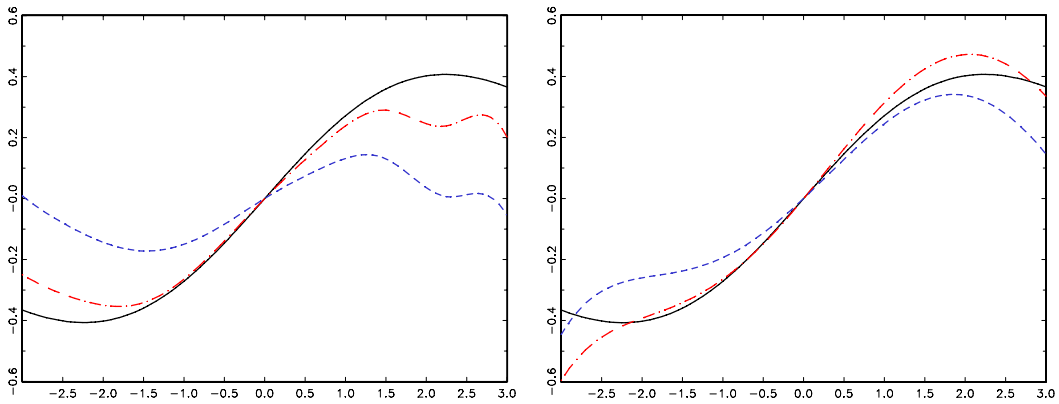

FIGURE C.5 : Nonparametric estimation - Cubic splines (left, 4 knots) *v.s.* Power series (right, 4th polynomial).

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 D.1 : 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 D.2 : 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 D.3 : 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 D.4 : 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."

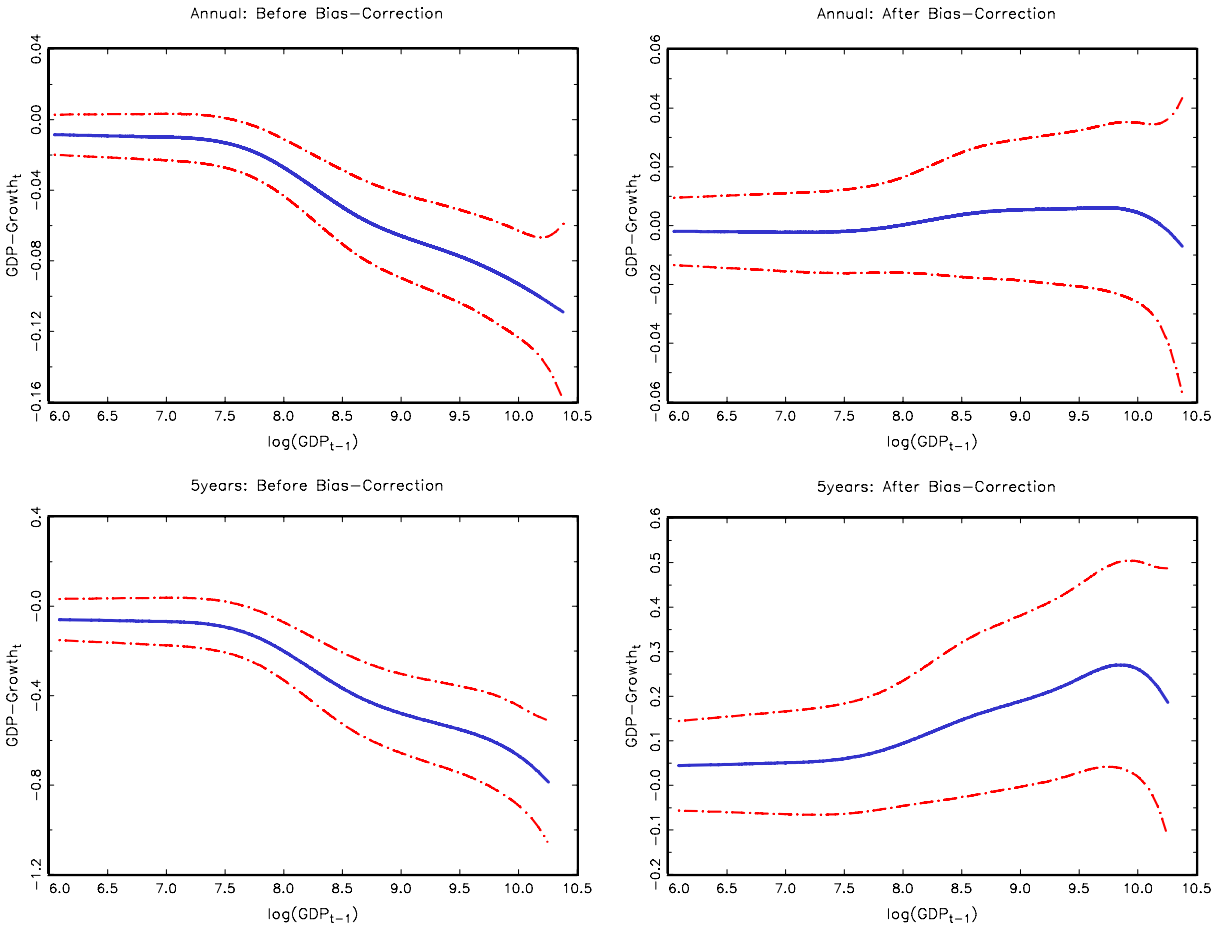

FIGURE D.1 : GDP growth versus $\log(GDP_{t-1})$ of all 73 countries. The vertical axis represents the GDP growth after controlling country-specific fixed effects, saving rates, population growth rate, depreciation rate, technical growth rate and human capital (for bottom two graphs only). Top two graphs 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 95% confidence regions.

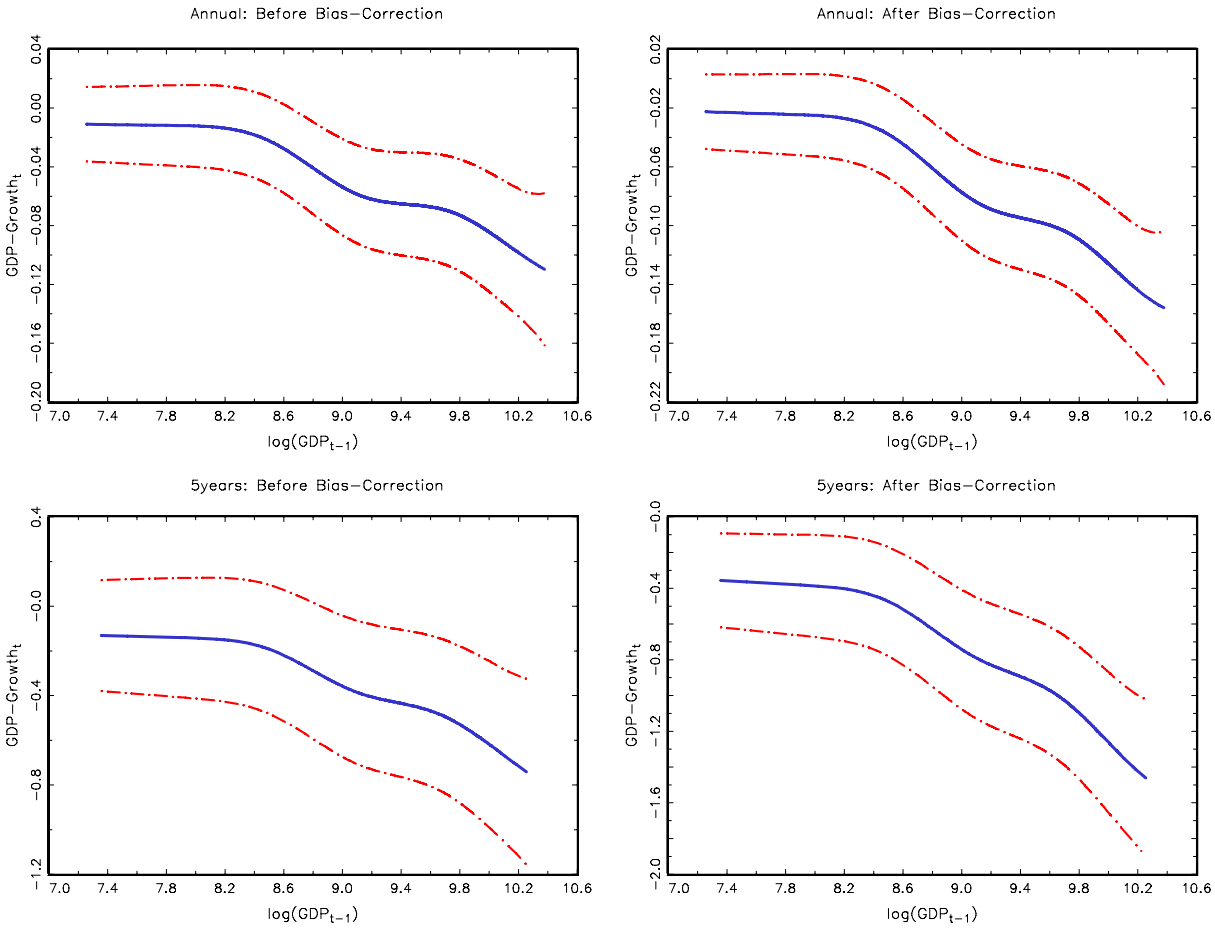

FIGURE D.2 : GDP growth versus $\log(GDP_{t-1})$ of the 24 OECD countries. The vertical axis represents the GDP growth after controlling country-specific fixed effects, saving rates, population growth rate, depreciation rate, technical growth rate and human capital (for bottom two graphs only). Top two 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 95% confidence regions.

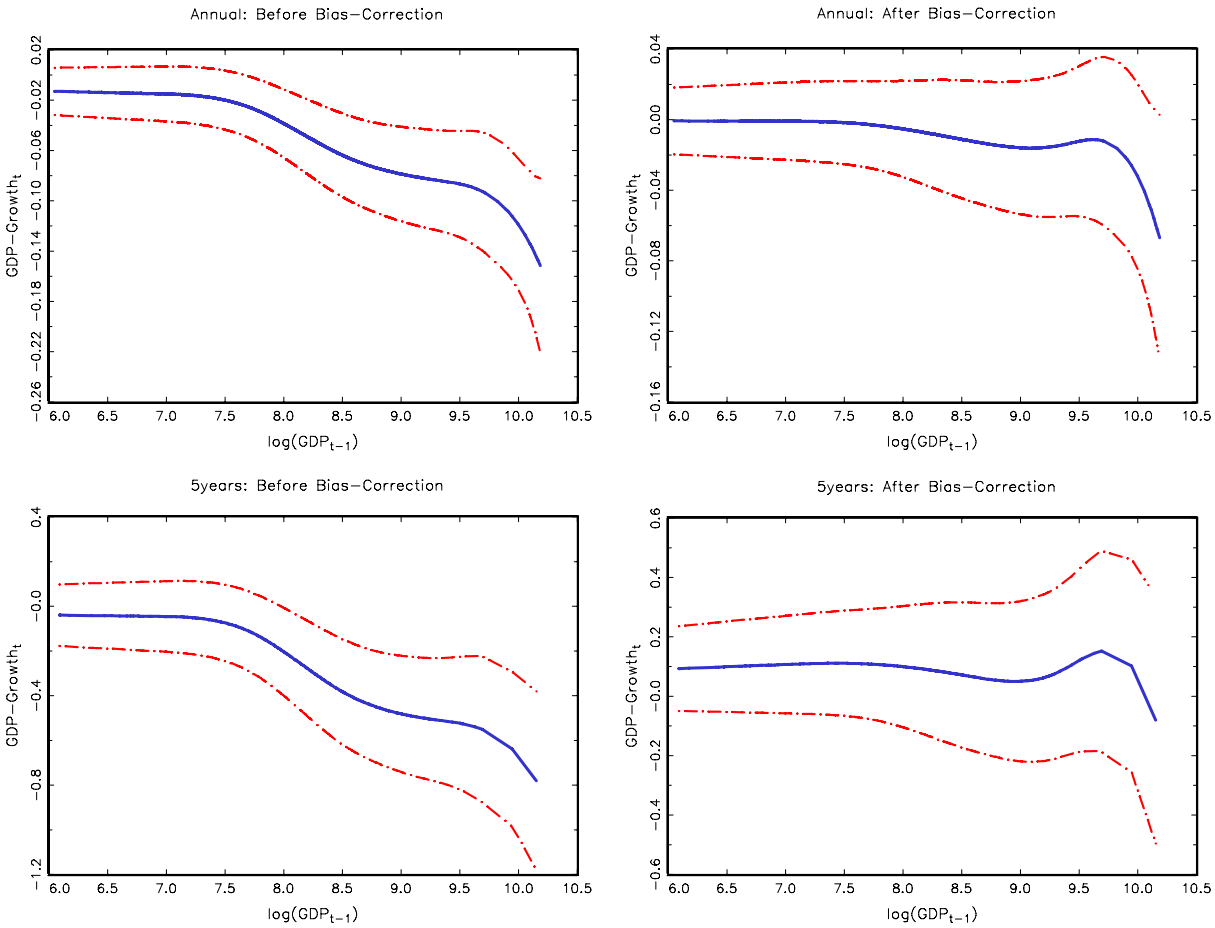

FIGURE D.3 : GDP growth versus $\log(GDP_{t-1})$ of 49 non-OECD countries. The vertical axis represents the GDP growth after controlling country-specific fixed effects, saving rates, population growth rate, depreciation rate, technical growth rate and human capital (for bottom two graphs only). Top two 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 95% confidence regions.

References

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

Altonji, J., and R.L. Matzkin (2005). Cross section and panel data estimators for nonseparable models with endogenous regressors, *Econometrica*, 73, 1053-1102.

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

Andrews, D.W.K. (1991b). 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, *Advances in Economics and Econometrics: Theory and Applications - Eighth World Congress,* Volum II, M. Dewatripont, L.P. Hansen, and S.J. Turnovsky (eds.), Cambridge: Cambridge University Press.

Chesher, A. (2003). Identification in nonseparable models, *Econometrica*, 71, 1405-1441.

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., C. Zhang, and J. Zhang (2001). Generalized likelihood test statistic and Wilks phenomenon, *The Annals of Statistics*, 29, 153-193.

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

Fisher, R.A. (1925). *Statistical Methods for Research Workers*, Edinburgh: Oliver abd Boyd.

Florens, J.-P. (2003). Inverse problems and structural econometrics: the example of instrumental variables, *Advances in Economics and Econometrics: Theory and Applications - Eighth World Congress,* Volum II, M. Dewatripont, L.P. Hansen, and S.J. Turnovsky (eds.), Cambridge: Cambridge University Press.

Gallant, A.R., and D.W. Nychka (1987). Semi-nonparametric maximum likelihood estimation, *Econometrica*, 55, 363-390.

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 (2005). Nonparametric methods for inference in the presence of instrumental variables, *Annals of Statistics*, 33, 2904–2929.

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.

Tikhonov, A., A. Goncharsky, V. Stepanov, and A. Yagola (1995). *Numerical Methods for the Solution of Ill-Posed Problems: Mathematics and Its Applications*, New York: Springer.

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.

Wilson, E.B., and M.M. Hilferty (1931). The distribution of chi-square, *Proceedings of the National Academy of Sciences of the United Staties of America*, 17, 684-688.

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