

# Sieve Estimation of Panel Data Models with Cross Section Dependence

Liangjun Su<sup>a</sup>, Sainan Jin<sup>a \*</sup>

<sup>a</sup>School of Economics, Singapore Management University, Singapore

July 20, 2010

## Abstract

In this paper we consider the problem of estimating semiparametric panel data models with cross section dependence, where the individual-specific regressors enter the model non-parametrically, and the common factors (both observed and unobserved) enter the model linearly. We consider both heterogeneous and homogenous nonparametric regression relationships when both the time dimension ( $T$ ) and the cross-section dimension ( $n$ ) are large. We propose sieve estimators for the nonparametric regression functions by extending Pesaran's (2006) common correlated effects (CCE) estimator to our semiparametric framework. Asymptotic normal distributions for the proposed estimators are derived and asymptotic variance estimators are provided. Monte Carlo simulations indicate that our estimators perform well in finite samples.

**JEL Classifications:** C13, C14, C33

**Key Words:** Common factor; Cross-section dependence; Heterogeneous regression; Panel data; Sieve estimation

## 1 Introduction

Recently there has been a growing interest in the estimation of panel data models with cross section dependence. See Bai (2009), Coakley, Fuertes, and Smith (2002), Greenaway-McGrevy, Han, and Sul (2008), Harding (2007), Kapetanios and Pesaran (2005), Moon and

---

\*The authors are grateful to the guest editors and two anonymous referees for their many constructive comments on an early version of this paper. They gratefully acknowledge the financial support from the NSFC under grant numbers 70501001 and 70601001. Address Correspondence to: Liangjun Su, School of Economics, Singapore Management University, 90 Stamford Road, Singapore 178903; E-mail: ljsu@smu.edu.sg, Phone: +65 6828 0386.

Weidner (2008), Pesaran (2004, 2006), Pesaran and Tosetti (2007), Phillips and Sul (2003, 2007), among others, for an overview. All of these papers focus on the linear specification of the regression relationship.

In this paper, we consider a semiparametric panel data model with cross section dependence. Let  $y_{it}$  be the observation on the  $i$ th cross section unit at time  $t$  for  $i = 1, \dots, n$ ;  $t = 1, \dots, T$ . We suppose that  $y_{it}$  is generated according to the following semiparametric panel data generating process

$$y_{it} = g_i(x_{it}) + \gamma'_{1i}f_{1t} + e_{it}, \quad (1.1)$$

where  $x_{it} \in \mathcal{X}_i \subset \mathbb{R}^d$  is a vector of observed individual-specific regressors on the  $i$ th cross section unit at time  $t$ ,  $g_i(\cdot) \in \mathcal{G}_i$ ,  $\mathcal{G}_i$  is a specified class of continuous function from  $\mathcal{X}_i$  to  $\mathbb{R}$ ,  $f_{1t}$  is a  $q_1 \times 1$  vector of observed common factors, and  $\gamma_{1i}$ ,  $i = 1, \dots, n$ , are factor loadings. Throughout the paper we assume that  $f_{1t}$  includes the intercept term and impose the condition  $E[g_i(x_{it})] = 0$  in order to identify  $g_i(\cdot)$ .<sup>1</sup>The error term  $e_{it}$  in (1.1) follows the multi-factor structure

$$e_{it} = \gamma'_{2i}f_{2t} + \varepsilon_{it}, \quad (1.2)$$

where  $f_{2t}$  is a  $q_2 \times 1$  vector of unobserved common factors,  $\varepsilon_{it}$  is the individual-specific (idiosyncratic) errors assumed to be independently distributed of  $(x_{it}, f_{1t}, f_{2t})$ , and  $\gamma_{2i}$ ,  $i = 1, \dots, n$ , are factor loadings. We are interested in the estimation of  $g_i(\cdot)$  in the presence of multi-factor error structure. Like Bai (2009), Pesaran (2006), and Moon and Weidner (2008), we focus on the case where both the cross-section dimension ( $n$ ) and the time dimension ( $T$ ) are large unless otherwise stated.

Like Pesaran (2006), the unobserved factors  $f_{2t}$  could be correlated with  $(x_{it}, f_{1t})$ . To allow for such a possibility, we follow Pesaran (2006) and adopt the following fairly general model for the individual-specific regressors,

$$x_{it} = \Gamma'_{1i}f_{1t} + \Gamma'_{2i}f_{2t} + v_{it}, \quad (1.3)$$

where  $\Gamma_{1i}$  and  $\Gamma_{2i}$  are  $q_1 \times d$  and  $q_2 \times d$  factor loading matrices, and  $v_{it}$  is a  $d \times 1$  vector of individual-specific components of  $x_{it}$ .

The model specified in (1.1)-(1.3) is fairly general and includes a variety of panel data models as special cases. First, Pesaran's (2006) model corresponds to the case where  $g_i(x) = \beta'_i x$  for some  $d \times 1$  vector  $\beta_i$  so that model (1.1) becomes  $y_{it} = \beta'_i x_{it} + \gamma'_{1i}f_{1t} + e_{it}$ . Second, it

---

<sup>1</sup>Write  $f_{1t} = (1, f_{1t}^*)'$ . As a referee suggested, one can also allow  $f_{1t}^*$  to enter (1.1) nonparametrically, in which case (1.1) will become  $y_{it} = g_i(x_{it}, f_{1t}^*) + \gamma_{1i} + e_{it}$ , or  $y_{it} = g_i(x_{it}) + h_i(f_{1t}^*) + \gamma_{1i} + e_{it}$  where  $h_i(\cdot)$  is an unknown smooth function. The theory developed below allows some component of  $x_{it}$  in (1.1) not to vary across  $i$ , and thus the former case can be treated as a special case of (1.1), where in (1.1)  $x_{it}$  includes some observable common factors and  $f_{1t} \equiv 1$ . The latter case is a special case of the former case where an additivity structure is imposed.

includes the conventional fixed or random effects models, the models of Bai (2003, 2009), Bai and Ng (2002), and Stock and Watson (2002), where the focus may be different from ours. Third, it includes the usual nonparametric panel data model  $y_{it} = g(x_{it}) + \alpha_i + v_t + \varepsilon_{it}$ , where the individual effects  $\alpha_i$  and the time effects  $v_t$  enter the model additively. See Henderson, Carroll and Li (2008) and Huang (2006) for kernel estimation of such models.

In practice, one may also be interested in estimating a restricted submodel of (1.1)

$$y_{it} = g(x_{it}) + \gamma'_{1i} f_{1t} + e_{it}. \quad (1.4)$$

That is,  $g_i(x) = g(x)$  for all  $i$  in model (1.1). In the case where  $\gamma_{1i} = 0$ , (1.4) can be regarded as a nonparametric extension of Bai's (2009) linear panel data model with multi-factor error structure or a simple extension of Huang's (2006) nonparametric panel data from his single-factor error structure to multiple-factor error structure. We call the regression functions homogeneous when  $g_i(x) = g(x)$  for all  $i$  as in (1.4) and heterogeneous otherwise.

To proceed, it is worth mentioning that the study of the estimation of  $g_i(\cdot)$  is important in several contexts despite the fact that for given  $i$ , it essentially involves only time series regression. First, if we do not want to impose homogeneous regression relationship that  $g_i(\cdot)$  is the same across  $i$ , we can only estimate  $g_i(\cdot)$  or certain averages of these functions. Second, the estimation of  $g_i(\cdot)$  for  $i = 1, \dots, n$ , will serve as a basis for testing the homogeneous regression relationship in (1.4). For example, Jin and Su (2010) consider a test statistic based upon the measure  $\Upsilon \equiv \sum_{i=1}^{n-1} \sum_{j=i+1}^n \int (g_i(x) - g_j(x))^2 w(x) dx$  where  $w(\cdot)$  is a weight function. Third, the presence of unobservable common factor  $f_{2t}$  complicates the analysis of the estimate of  $g_i(\cdot)$  to a great deal. Fourth, the analysis of the estimate of  $g_i(\cdot)$  will facilitate the study of the homogenous regression relationship  $g(\cdot)$ .

In the following we study the sieve estimation of both the heterogeneous regression function  $g_i(\cdot)$  in (1.1) and the homogeneous regression function  $g(\cdot)$  in (1.4) under the assumption that both the error term  $e_{it}$  and the individual specific regressors  $x_{it}$  exhibit multi-factor structures defined in (1.2)-(1.3), respectively. In either case, we can extend the common correlated effect (CCE) estimator of Pesaran (2006) to our semiparametric model. We show that the CCE estimators of both the heterogeneous and homogenous regression functions are consistent as both  $n$  and  $T$  tend to infinity, as long as certain rank condition concerning the mean of  $(\Gamma_{2i}, \gamma_{2i})$  is satisfied. We establish the asymptotic normality of these estimators and propose consistent estimators for their asymptotic variances. A small set of Monte Carlo simulations are conducted to investigate the finite sample performance of our estimators. We find that our estimators perform quite well in finite samples.

The rest of the paper is structured as follows. Section 2 motivates and proposes the sieve-based CCE estimation of heterogeneous and homogenous regression functions. In Section 3 we make some basic assumptions that underlie our asymptotic analysis. Sections 4 and 5 study the asymptotic properties of the estimators for the heterogeneous and homogenous

regression functions, respectively. A small set of Monte Carlo simulation results are reported in Section 6. Final remarks are contained in Section 7. All technical details are relegated to the Appendix.

NOTATION. Throughout the paper we adopt the following notation and conventions. For a real matrix  $A$ , we denote its Euclidean norm as  $\|A\| = [\text{tr}(AA')]^{1/2}$  and its generalized inverse as  $A^-$ . When  $A$  is symmetric, we use  $\lambda_{\min}(A)$  and  $\lambda_{\max}(A)$  to its minimum and maximum eigenvalues, respectively. For any real square matrices  $A$  and  $B$ , we write  $A \leq B$  to signify that  $B - A$  is positive semidefinite (p.s.d.). For a vector  $a \equiv (a_1, \dots, a_T)'$ ,  $\text{diag}(a)$  denotes a diagonal matrix with  $a_i$  as a typical diagonal element.  $I_T$  denotes a  $T \times T$  identity matrix. The operator  $\xrightarrow{p}$  denotes convergence in probability,  $\xrightarrow{a.s.}$  convergence almost surely, and  $\xrightarrow{d}$  convergence in distributions. We use  $(n, T) \rightarrow \infty$  to denote the joint convergence of  $n$  and  $T$  in Section 4, and it denotes the case where  $T$  is either fixed or passing to  $\infty$  as  $n \rightarrow \infty$  in Section 5.

## 2 Motivation and estimation

In this section we first motivate the idea of CCE estimation and then propose sieve-based CCE estimators for the heterogenous and homogeneous regression functions in (1.1) and (1.4), respectively.

### 2.1 Motivation

Let  $\bar{x}_t \equiv n^{-1} \sum_{i=1}^n x_{it}$  and  $\bar{y}_t \equiv n^{-1} \sum_{i=1}^n y_{it}$ . Then (1.1)-(1.3) implies that

$$\begin{pmatrix} \bar{x}_t \\ \bar{y}_t \end{pmatrix} = \begin{pmatrix} \bar{\Gamma}'_1 \\ \bar{\gamma}'_1 \end{pmatrix} f_{1t} + \begin{pmatrix} \bar{\Gamma}'_2 \\ \bar{\gamma}'_2 \end{pmatrix} f_{2t} + \begin{pmatrix} \bar{v}_t \\ \bar{g}_t + \bar{\varepsilon}_t \end{pmatrix}, \quad (2.1)$$

where  $\bar{\Gamma}_1$ ,  $\bar{\Gamma}_2$ ,  $\bar{\gamma}_1$ ,  $\bar{\gamma}_2$ ,  $\bar{v}_t$ , and  $\bar{\varepsilon}_t$  are sample averages of  $\Gamma_{1i}$ ,  $\Gamma_{2i}$ ,  $\gamma_{1i}$ ,  $\gamma_{2i}$ ,  $v_{it}$ , and  $\varepsilon_{it}$  over  $i$ , respectively, and  $\bar{g}_t = n^{-1} \sum_{i=1}^n g_i(x_{it})$ . Let  $\bar{\Gamma}_2^* \equiv (\bar{\Gamma}_2, \bar{\gamma}_2)$ . Following the lead of Pesaran (2006), we can premultiply both sides of (2.1) by  $\bar{\Gamma}_2^*$  and solve for  $f_{2t}$ :

$$f_{2t} = \left( \bar{\Gamma}_2^* \bar{\Gamma}_2^{*'} \right)^{-1} \bar{\Gamma}_2^* \left( \begin{pmatrix} \bar{x}_t \\ \bar{y}_t \end{pmatrix} - \begin{pmatrix} \bar{\Gamma}'_1 \\ \bar{\gamma}'_1 \end{pmatrix} f_{1t} - \begin{pmatrix} \bar{v}_t \\ \bar{g}_t + \bar{\varepsilon}_t \end{pmatrix} \right) \quad (2.2)$$

provided that

$$\text{rank}(\bar{\Gamma}_2^*) = q_2 \leq d + 1 \text{ for sufficiently large } n. \quad (2.3)$$

As  $n \rightarrow \infty$ ,  $\bar{v}_t \xrightarrow{p} 0$ ,  $\bar{\varepsilon}_t \xrightarrow{p} 0$  and  $\bar{g}_t \xrightarrow{p} 0$  for each  $t$  under weak conditions. It follows

$$f_{2t} - \left( \bar{\Gamma}_2^* \bar{\Gamma}_2^{*'} \right)^{-1} \bar{\Gamma}_2^* \left( \begin{pmatrix} \bar{x}_t \\ \bar{y}_t \end{pmatrix} - \begin{pmatrix} \bar{\Gamma}'_1 \\ \bar{\gamma}'_1 \end{pmatrix} f_{1t} \right) \xrightarrow{p} 0 \text{ as } n \rightarrow \infty. \quad (2.4)$$

The last line suggests that we can use  $h_t \equiv (f'_{1t}, \bar{x}'_t, \bar{y}_t)'$  as observable proxies for  $f_{2t}$ . As we shall see later, we can consistently estimate  $g_i(\cdot)$  or  $g(\cdot)$  by augmenting the semiparametric regression of  $y_{it}$  on  $x_{it}$  with  $h_t$ . Following Pesaran (2006), we call such an estimator as the *common correlated effect* (CCE) estimator.

## 2.2 Common correlated effect estimation of $g_i(\cdot)$ and $g(\cdot)$

(1.1) and (1.4) are additive panel data models. We propose to estimate  $g_i(\cdot)$  or  $g(\cdot)$  by sieve methods. For an excellent review on sieve methods, see Chen (2007).

To proceed, let  $\{p_l(x), l = 1, 2, \dots\}$  denote a sequence of known basis functions that can approximate any square-integrable function of  $x$  very well (to be more precise later). Let  $K \equiv K(T)$  (in the estimation of  $g_i(\cdot)$  in (1.1)) or  $K \equiv K(n, T)$  (in the estimation of  $g(\cdot)$  in (1.4)) be some integer such that  $K \rightarrow \infty$  as  $n \rightarrow \infty$  or as  $(n, T) \rightarrow \infty$ . Let  $p^K(x) = (p_1(x), p_2(x), \dots, p_K(x))'$ ,  $p_{it} = p^K(x_{it})$ ,  $p_i = (p_{i1}, p_{i2}, \dots, p_{iT})'$ , and  $P = (p'_1, p'_2, \dots, p'_n)'$ . Obviously we have suppressed the dependence of  $p_{it}$ ,  $p_i$ , and  $P$  on  $K$ ,  $T$ , or  $n$ . In particular,  $p_i$  is a  $T \times K$  matrix and  $P$  is of dimension  $nT \times K$ .

Under fairly weak conditions, we can approximate  $g_i(x)$  in (1.1) very well by  $\alpha'_{g_i} p^K(x)$  for some  $K \times 1$  vector  $\alpha_{g_i}$ , and  $g(x)$  in (1.4) very well by  $\alpha'_g p^K(x)$ . In estimating  $g_i(x)$ , one can allow  $K$  to be  $i$ -dependent and write  $K$  as  $K_i$ . But we keep using the same notation  $K$  in estimating  $g_i(x)$  and  $g(x)$  for notational simplicity.

### Estimation of $g_i(\cdot)$

To estimate  $\alpha_{g_i}$ , we run the regression of  $y_{it}$  on  $p^K(x_{it})$  and  $h_t \equiv (f'_{1t}, \bar{x}'_t, \bar{y}_t)'$

$$y_{it} = \alpha'_{g_i} p^K(x_{it}) + \vartheta'_i h_t + u_{it} \quad (2.5)$$

where  $u_{it}$  is the new error term. Let  $y_i = (y_{i1}, y_{i2}, \dots, y_{iT})'$ ,  $h = (h_1, h_2, \dots, h_T)'$ , and  $u_i = (u_{i1}, u_{i2}, \dots, u_{iT})'$ . We can rewrite (2.5) in vector form

$$y_i = p_i \alpha_{g_i} + h \vartheta_i + u_i \quad (2.6)$$

By the formula for partitioned regression, the estimator of  $\alpha_{g_i}$  in (2.5) or (2.6) is given by

$$\hat{\alpha}_{g_i} = (p'_i m_h p_i)^{-} p'_i m_h y_i, \quad (2.7)$$

where  $m_h \equiv I_T - h(h'h)^{-} h$ , and  $(\cdot)^{-}$  denotes any symmetric generalized inverse. The estimator of  $g_i(x)$  is then given by

$$\hat{g}_i(x) = p^K(x)' \hat{\alpha}_{g_i}. \quad (2.8)$$

We will show that  $\hat{g}_i(x)$  is a consistent estimator of  $g_i(x)$  and establish its asymptotic normality under suitable assumptions.

### Estimation of $g(\cdot)$

If model (1.4) is assumed to be correctly specified in conjunction with (1.2) - (1.3), we can estimate  $g(\cdot)$  by pooling all the data together and obtain the *CCE pooled* estimator. Let  $Y = (y'_1, \dots, y'_n)'$ ,  $U = (u'_1, \dots, u'_n)'$ ,  $\vartheta = (\vartheta'_1, \dots, \vartheta'_n)'$ , and  $H = I_n \otimes h$ . We can rewrite (2.6) in matrix form

$$Y = P\alpha_g + H\vartheta + U \quad (2.9)$$

By the formula for partitioned regression, the estimator of  $\alpha_g$  in (2.9) is given by

$$\hat{\alpha}_g = (P'M_h P)^{-1} P'M_h Y = \left( \sum_{i=1}^n p'_i m_h p_i \right)^{-1} \sum_{i=1}^n p'_i m_h y_i, \quad (2.10)$$

where  $M_h = I_n \otimes m_h$ . The estimator of  $g(x)$  is then given by

$$\hat{g}(x) = p^K(x)' \hat{\alpha}_g. \quad (2.11)$$

We will show later that  $\hat{g}(x)$  is a consistent estimator of  $g(x)$  and establish its asymptotic normality under suitable assumptions.

## 3 Basic assumptions

In this section, we provide a set of basic assumptions that are used in the asymptotic analysis.

**Assumption 1.** (i) For each  $i$ , the process  $\{(\varepsilon_{it}, v_{it}) : t \geq 1\}$  is a strictly stationary and  $\alpha$ -mixing process with mixing coefficient  $\alpha_i(j)$  such that  $\sum_{j=1}^{\infty} j^2 \alpha_i(j)^{\eta/(4+\eta)} \leq C_{1i}$  for some  $C_{1i} < \infty$  and  $\eta > 0$ . (ii) The common factors process  $\{(f_{1t}, f_{2t}) : t \geq 1\}$  is a strictly stationary and  $\alpha$ -mixing process with mixing coefficient  $\alpha_0(j)$  such that  $\sum_{j=1}^{\infty} j^2 \alpha_0(j)^{\eta/(4+\eta)} \leq C_2 < \infty$ . (iii)  $(f_{1t}, f_{2t})$  is distributed independently of the individual-specific errors  $\varepsilon_{is}$  and  $v_{is}$  for all  $i, t$ , and  $s$ .  $E[(f'_{1t}, f'_{2t})(f'_{1t}, f'_{2t})']$  is positive definite. (iv) The individual-specific errors  $\varepsilon_{it}$  and  $v_{js}$  are distributed independently for all  $i, j, t$ , and  $s$ . (v) Let  $\varepsilon_i \equiv (\varepsilon_{i1}, \varepsilon_{i2}, \dots, \varepsilon_{iT})'$  and  $v_i \equiv (v_{i1}, v_{i2}, \dots, v_{iT})'$ .  $(\varepsilon_i, v_i)$  are independently distributed across  $i$  with zero means. (vi)  $\sup_{n \geq 1} \max_{1 \leq i \leq n} E|\zeta_i|^{4+\eta} \leq \bar{\mu}_{4+\eta} < \infty$  for  $\zeta_i = \varepsilon_{i1}, v_{i1}, g_i(x_{i1}), f_{11}$ , and  $f_{21}$ . (vii)  $\Psi_{i,T} \equiv \text{Var}(\varepsilon_i)$  has the smallest eigenvalue that is bounded away from zero and bounded largest eigenvalue. (viii)  $n^{-1} \sum_{i=1}^n \sum_{j=1}^{\infty} j^2 \alpha_i(j)^{\eta/(4+\eta)} \leq C_3 < \infty$ . (ix)  $E[g_i(x_{it})] = 0$  for each  $i$ .

Assumptions 1(i)-(ii) specify that the processes  $\{\varepsilon_{it}, v_{it}\}$  and  $\{f_{1t}, f_{2t}\}$  are strictly stationary and  $\alpha$ -mixing with mixing rates decaying to zero sufficiently fast. They imply that  $\alpha_i(j)$ ,  $i = 0, 1, \dots, n$ , are of order  $o(j^{-(3+16/\eta)})$ . The smaller  $\eta$  is, the faster these mixing rates decay to zero. It is worth mentioning that these assumptions are quite weak in the sense that many time series processes satisfy the  $\alpha$ -mixing conditions. Assumptions 1(iii)-(v)

are also made in Pesaran (2006), which greatly facilitate the asymptotic analysis. Assumption 1(vi) specifies the moment conditions on  $\varepsilon_{it}$ ,  $v_{it}$ ,  $g_i(x_{it})$ ,  $f_{1t}$ , and  $f_{2t}$ . A combination of Assumption 1(vi) with Assumptions 1(i)-(ii) reflects the typical trade-off between the mixing coefficients and moments. In addition, notice that we rule out weak cross sectional dependence in  $\{\varepsilon_{it}, v_{it}\}$  in Assumptions 1(iv)-(v) while Bai and Ng (2002) allow for it. Assumption 1(vii) is typically assumed in the literature when the data are independently distributed, see Andrews (1991a). It is automatically satisfied if  $\{\varepsilon_{it}, t \geq 1\}$  is a martingale difference sequence with finite positive variance  $\sigma_i^2$ , in which case  $\Psi_{i,T} = \sigma_i^2 I_T$ . Assumption 1(viii) facilitates the presentation of our results. Assumption 1(ix) is an identification condition because we assume that  $f_{1t}$  contains the intercept term.

**Assumption 2.** (i) The unobserved factor loadings  $\gamma_{2i}$  and  $\Gamma_{2i}$  are independently and identically distributed (IID).  $\gamma_{2i}$  and  $\Gamma_{2i}$  are independent of the individual-specific errors  $\varepsilon_{jt}$  and  $v_{jt}$ , and the common factors  $(f_{1t}, f_{2t})$  for all  $j$  and  $t$ . The  $(4 + \eta)$ -th moment of  $\Gamma_{2i}$  is finite. (ii)  $\Gamma_{1i}$  are either fixed factor loadings that are uniformly bounded or random factor loadings that are IID across  $i$  with finite  $(4 + \eta)$ -th moments and are independent of  $\Gamma_{2j}, \varepsilon_{jt}, v_{jt}, f_{1t}$  and  $f_{2t}$  for all  $j$  and  $t$ . (iii) Let  $\Gamma_2^* = E(\Gamma_{2i}^*)$  where  $\Gamma_{2i}^* \equiv (\Gamma_{2i}, \gamma_{2i})$ .  $\text{rank}(\Gamma_2^*) = q_2 \leq d + 1$ .

Assumption 2(i) imposes restrictions on the loadings for the unobserved factors and they allow for random factor loadings. Our results still hold when these loadings are nonrandom as in Bai (2003). Assumption 2(ii) imposes conditions on the loadings for the observed factors. Assumptions 2(i)-(ii) in conjunction with Assumptions 1(v)-(vi) imply that  $h_t \equiv (f_{1t}', \bar{x}_t', \bar{y}_t)'$  has finite  $(4 + \eta)$ -th moments. Similarly, combining Assumptions 2(i)-(ii) with Assumptions 1(i)-(ii) implies that  $\{x_{it}, t \geq 1\}$  is also an  $\alpha$ -mixing process with the mixing coefficients jointly determined by  $\alpha_i(j)$  and  $\alpha_0(j)$ . The rank condition in Assumption 2(iii) ensures that  $\text{rank}(\bar{\Gamma}_2^*) = q_2 \leq d + 1$  with probability approaching 1 (w.p.a. 1) as  $n \rightarrow \infty$ . To see this, let  $\bar{\varsigma} \equiv \bar{\Gamma}_2^* - \Gamma_2^*$ . Then by Assumption 2(i) and the strong law of large numbers,  $\bar{\varsigma} \xrightarrow{a.s.} 0$  as  $n \rightarrow \infty$ . By Assumption 2(iii) and Theorem 7.2.2. of Wang, Wei, and Qiao (2004), this implies that for sufficiently large  $n$ ,  $\text{rank}(\bar{\Gamma}_2^*) = q_2 \leq d + 1$  a.s.

To estimate the unknown function  $g_i(\cdot)$  well by the sieve methods, we assume that  $g_i(x)$  is smooth in some sense with respect to  $x$ . Let  $\mathcal{X}_i$  denote the support of the individual-specific regressor  $x_{it}$ . Typical approximation and estimation of regression functions require that  $\mathcal{X}_i$  be compact. See Newey (1994, 1995, 1997), Li (2000), Baltagi and Li (2002), and Li, Hsiao and Zinn (2003), among others. To allow for the unboundedness of  $\mathcal{X}_i$  (e.g.,  $\mathcal{X}_i = \mathbb{R}^d$ ), we follow Chen, Hong, and Tamer (2005) (see also Blundell, Chen, and Kristensen (2007)) and use a weighted sup-norm metric defined as

$$\|g_i\|_{\infty, \omega} \equiv \sup_{x \in \mathcal{X}_i} |g_i(x)| \left[1 + \|x\|^2\right]^{-\omega/2} \text{ for some } \omega \geq 0.$$

Clearly, the choice of  $\omega = 0$  leads to the usual sup-norm which is suitable if  $\mathcal{X}_i$  is a bounded

subset of  $\mathbb{R}^d$ .

Recall that a typical smoothness assumption requires that a function  $b : \mathcal{X} \rightarrow \mathbb{R}$  belongs to a Hölder space. Let  $\mathbf{a} \equiv (a_1, \dots, a_d)'$  denote a  $d$ -vector of non-negative integers and  $|\mathbf{a}| \equiv \sum_{k=1}^d a_k$ . For any  $x = (x_1, \dots, x_d)' \in \mathbb{R}^d$ , the  $|\mathbf{a}|$ -th derivative of a function  $b : \mathcal{X} \rightarrow \mathbb{R}$  is denoted as  $\nabla^{\mathbf{a}}b(x) = \partial^{|\mathbf{a}|}b(x) / \partial x_1^{a_1} \dots \partial x_d^{a_d}$ . Let  $\lceil \lambda \rceil$  denote the largest integer that is strictly smaller than  $\lambda$ . The Hölder space  $\Lambda^\lambda(\mathcal{X})$  of order  $\lambda > 0$  is a space of functions  $b : \mathcal{X} \rightarrow \mathbb{R}$  such that the first  $\lceil \lambda \rceil$  derivatives are bounded, and the  $\lceil \lambda \rceil$ -th derivatives are Hölder continuous with the exponent  $\lambda - \lceil \lambda \rceil \in (0, 1]$ . Define the Hölder norm:

$$\|b\|_{\Lambda^\lambda} \equiv \sup_{x \in \mathcal{X}} |b(x)| + \max_{|\mathbf{a}|=\lceil \lambda \rceil} \sup_{x \neq x^*} \frac{|\nabla^{\mathbf{a}}b(x) - \nabla^{\mathbf{a}}b(x^*)|}{(\|x - x^*\|)^{\lambda - \lceil \lambda \rceil}} < \infty.$$

The following definition is adopted from Chen, Hong, and Tamer (2005).

**Definition 1.** Let  $\Lambda^\lambda(\mathcal{X}, \omega) = \{b : \mathcal{X} \rightarrow \mathbb{R} \text{ such that } b(\cdot) [1 + \|\cdot\|^2]^{-\omega/2} \text{ is in } \Lambda^\lambda(\mathcal{X})\}$  denote a weighted Hölder space of functions. A weighted Hölder ball with radius  $c$  is

$$\Lambda_c^\lambda(\mathcal{X}, \omega) \equiv \left\{ b \in \Lambda^\lambda(\mathcal{X}, \omega) : \left\| b(\cdot) [1 + \|\cdot\|^2]^{-\omega/2} \right\|_{\Lambda^\lambda} \leq c < \infty \right\}.$$

A function  $b(\cdot)$  is said to be  $H(\lambda, \omega)$ -smooth on  $\mathcal{X}$  if it belongs to a weighted Hölder ball  $\Lambda_c^\lambda(\mathcal{X}, \omega)$  for some  $\lambda > 0$ ,  $c > 0$  and  $\omega \geq 0$ .

As Chen, Hong, and Tamer (2005) remark, the weighted Hölder ball with  $\omega = 0$  reduces to the standard Hölder ball  $\Lambda^\lambda(\mathcal{X})$  condition, which is usually a sufficient condition when the support  $\mathcal{X}$  of the regressor is a bounded support of  $\mathbb{R}^d$ . When  $\mathcal{X} = \mathbb{R}^d$ , the standard Hölder ball  $\Lambda^\lambda(\mathcal{X})$  may exclude simple functions such as  $b(x) = x$ . Let

$$Q_{ipp} = E[p_{it}p'_{it}], \quad Q_{iph} = E[p_{it}h'_t], \quad Q_{hh} = E[h_t h'_t], \quad \text{and} \quad Q_i = Q_{ipp} - Q_{iph}Q_{hh}^{-1}Q'_{iph}, \quad (3.1)$$

where we have suppressed the dependence of  $Q_{hh}(\equiv Q_{n, hh})$  on  $n$  through  $h_t$ . The  $K \times K$  matrices  $Q_{ipp}$  and  $Q_i$  play an important role in this paper. The conditions in the following assumption are quite similar to those imposed by Chen, Hong, and Tamer (2005).

**Assumption 3.** (i) For each  $i$ ,  $g_i(\cdot)$  is  $H(\lambda_i, \omega_i)$ -smooth on  $\mathcal{X}_i$  for some  $\lambda_i > d/2$ ,  $\omega_i \geq 0$ . (ii) For each  $i$ ,  $\int (1 + \|x\|^2)^{\bar{\omega}_i} dF_i(x) < \infty$  for some  $\bar{\omega}_i > \omega_i + \lambda_i$ , where  $dF_i(x) = f_i(x) dx$ , and  $f_i(x)$  is the probability density function of  $x_{it}$ . (iii) For any  $H(\lambda_i, \omega_i)$ -smooth  $g_i(\cdot)$  on  $\mathcal{X}_i$ , there is a function  $\Pi_{\infty K} g_i \equiv \alpha'_{g_i} p^K(\cdot)$  in the sieve space  $\mathcal{G}_K \equiv \{f(\cdot) = a' p^K(\cdot)\}$  such that  $\|g_i(\cdot) - \Pi_{\infty K} g_i(\cdot)\|_{\infty, \bar{\omega}_i} = O(K^{-\lambda_i/d})$ . (iv) For each  $i$ ,  $\sup_{1 \leq j \leq K} E|p_j(x_{i1})|^{4+\eta} < \infty$  for the same  $\eta$  defined in Assumption 1(i),  $Q_i$  has the smallest eigenvalues bounded away from zero,  $Q_{ipp}$  has bounded largest eigenvalues uniformly in  $K$ , and  $Q_{hh} \equiv Q_{n, hh}$  tends to a positive definite matrix as  $n \rightarrow \infty$ .

Assumption 3(i) imposes a weighted smoothness condition on  $g_i(\cdot)$ . If we are only interested in the consistency of the estimator of  $g_i(\cdot)$ , we can simply require  $\lambda_i > 0$ . The

requirement  $\lambda_i > d/2$  will ensure the CCE estimator of  $g_i(\cdot)$  to achieve the Stone's (1982) optimal rate of convergence. Assumption 3(ii) imposes conditions on the tail behavior of the marginal densities. In fact, the weight function  $(1 + \|x\|^2)^{-\bar{\omega}_i/2}$  can be regarded as an alternative to the trimming function. It is used to deal with unbounded support as in Ai and Chen (2003), and Chen, Hong, and Tamer (2005). Assumption 3(iii) quantifies the approximation error of functions in  $H(\lambda_i, \omega_i)$  by the linear sieve basis functions  $p^K(x)$ . Assumption 3(iv) is a little stronger than what is typically assumed for sieve estimation in the IID framework (e.g., Newey, 1997).

## 4 Asymptotic properties of $\widehat{g}_i(x)$

In this section, we study the asymptotic properties of  $\widehat{g}_i(x)$ . Let  $f_1 = (f_{11}, f_{12}, \dots, f_{1T})'$ ,  $f_2 = (f_{21}, f_{22}, \dots, f_{2T})'$ ,  $\varepsilon_i = (\varepsilon_{i1}, \varepsilon_{i2}, \dots, \varepsilon_{iT})'$ , and  $\mathbf{g}_i = (g_i(x_{i1}), g_i(x_{i2}), \dots, g_i(x_{iT}))'$ . Using (1.1), (1.2), and the approximation for  $g_i(\cdot)$  we have

$$y_i = p_i \alpha_{g_i} + f_1 \gamma_{1i} + f_2 \gamma_{2i} + \varepsilon_i + (\mathbf{g}_i - p_i \alpha_{g_i}). \quad (4.1)$$

Therefore

$$\widehat{\alpha}_{g_i} - \alpha_{g_i} = (p_i' m_h p_i)^{-1} p_i' m_h \varepsilon_i + (p_i' m_h p_i)^{-1} p_i' m_h f_2 \gamma_{2i} + (p_i' m_h p_i)^{-1} p_i' m_h (\mathbf{g}_i - p_i \alpha_{g_i}). \quad (4.2)$$

The first term on the right hand side (r.h.s.) of the above expression is present even if there is no multi-factor error structure and  $g_i(x)$  is linear in  $x$ ; the second term is due to the presence of the unobserved factors in the error term; and the third term is due to the approximation of  $g_i(x)$  by  $p^K(x)' \alpha_{g_i}$ . As shown in the proof of the following theorem, each of the three terms on the r.h.s. of (4.2) contributes to the convergence rate of  $\widehat{g}_i$ .

**Theorem 4.1** (*Convergence rate*) Under Assumptions 1-3, (i)  $T^{-1} \sum_{t=1}^T [\widehat{g}_i(x_{it}) - g_i(x_{it})]^2 = O_p(K/T + K^{-2\lambda_i/d} + K/n)$ , (ii)  $\int_{x \in \mathcal{X}_i} [\widehat{g}_i(x) - g_i(x)]^2 dF_i(x) = O_p(K/T + K^{-2\lambda_i/d} + K/n)$ .

**Remark 1.** The above theorem states the results for both sample mean square error (c.f., Newey, 1994) and integrated mean square error (c.f., Newey, 1997). Let  $\|b\|_2 \equiv \{\int_{\mathcal{X}_i} b(x)^2 dF_i(x)\}^{1/2}$ . If the common factors  $f_{2t}$  were observable, we can run the semiparametric regression of  $y_{it}$  on  $x_{it}$ ,  $f_{1t}$  and  $f_{2t}$ . In this case,  $\|\widehat{g}_i - g_i\|_2 = O_p(\sqrt{K/T} + K^{-\lambda_i/d})$  and the optimal choice of  $K$  would balance the standard deviation ( $\sqrt{K/T}$ ) part and the bias ( $K^{-\lambda_i/d}$ ) part by choosing  $K \propto T^{d/(d+2\lambda_i)}$ . In this case,  $\|\widehat{g}_i - g_i\|_2 = O_p(T^{-\lambda_i/(d+2\lambda_i)})$ , which is the renowned optimal convergence rate of Stone (1982) for nonparametric least-squares regression, see also Theorem 1 of Newey (1997). Here, due to the use of proxies for  $f_{2t}$ , even though the standard deviation part ( $\sqrt{K/T}$ ) in Theorem 4.1 is of the same order as the standard case, the bias part ( $K^{-\lambda_i/d} + \sqrt{K/n}$ ) in Theorem 4.1 has an extra term ( $\sqrt{K/n}$ ).

In order to achieve the above optimal rate of convergence, one would require  $K \propto T^{d/(d+2\lambda_i)}$  and that  $n/T \rightarrow c \in (0, \infty]$  as  $(n, T) \rightarrow \infty$ , and the latter holds in conventional panel data models where there are a large number of observations across individuals and a short span over time.

To derive the asymptotic normality of  $\widehat{g}_i(x)$ , we add the following assumption.

**Assumption 4.** (i) Either of the following conditions hold: a) For fixed  $i$ ,  $\{\varepsilon_{it}, \mathcal{F}_{it}\}$  is a martingale difference sequence (MDS), where  $\mathcal{F}_{it} \equiv \sigma(\varepsilon_{is} : 1 \leq s \leq t)$ ; b) the  $\alpha$ -mixing condition in Assumption 1 is strengthened to  $\phi$ -mixing with mixing coefficients  $\alpha_i(j)$  replaced by  $\phi_i(j)$ , and  $\phi(j) \equiv \sup_{n \geq 1} \max_{0 \leq i \leq n} \phi_i(j) = O(j^{-1-\epsilon})$  for some  $\epsilon > 0$ ; c) the  $\alpha$ -mixing condition in Assumption 1 holds, and  $E[p_{it}p'_{it}p_{it}p'_{it}]$  has bounded largest eigenvalue for each  $K$ . (ii) For every  $K \equiv K(T)$ , it satisfies  $K^3/T \rightarrow 0$ ,  $KT/n \rightarrow 0$ , and  $TK^{-2\lambda_i/d} \rightarrow 0$  as  $(n, T) \rightarrow \infty$ .

We prove the asymptotic normality of  $\widehat{g}_i(x)$  under different conditions specified in Assumption 4(i). Under Condition a), a martingale central limit theorem (CLT) is needed, whereas under Condition b) or c) a CLT for double-array mixing processes is called upon. Assumption 4(ii) imposes some additional conditions on the choice of  $K$  and they also restrict the relative size of  $n$  versus  $T$ . The condition  $KT/n \rightarrow 0$  guarantees the proxy error is asymptotically negligible. Such a condition corresponds to Pesaran's (2006) condition  $T/n^2 \rightarrow 0$  when  $g_i$  is linear, and it would not be needed if the common factors  $f_{2t}$  were also observable. The condition  $TK^{-2\lambda_i/d} \rightarrow 0$  ensures that the bias of  $\widehat{g}_i(x)$  due to the sieve approximation error is asymptotically negligible.

Let  $V_{inT} = p^K(x)'(p'_i m_h p_i)^- p'_i m_h \Psi_{i,T} m_h p_i (p'_i m_h p_i)^- p^K(x)$  and  $A_{inT} = V_{inT}^{-1/2}$ , where  $\Psi_{i,T}$  is defined in Assumption 1(vii). The following theorem establishes the asymptotic normality of  $\widehat{g}_i(x)$ .

**Theorem 4.2** (*Asymptotic normality*) *Let  $x \in \mathcal{X}_i$  be given and  $\|p^K(x)\| > c$  for some constant  $c > 0$ . Under Assumptions 1-4,  $A_{inT}[\widehat{g}_i(x) - g_i(x)] \xrightarrow{d} N(0, 1)$ .*

**Remark 2.** The asymptotic normality result in Theorem 4.2 is similar to that obtained in Andrews (1991a) and Newey (1997) who considered the nonparametric series estimation in the IID setup. Even though it is not explicitly revealed,  $\widehat{g}_i(x)$  converges to  $g_i(x)$  at a rate slower than the usual  $\sqrt{T}$ -parametric rate as  $K \rightarrow \infty$ . To see this explicitly, observe that

$$\underline{\lim}_{n, T \rightarrow \infty} TV_{inT} \geq \lambda_{\min}(\Psi_{i,T}) \underline{\lim}_{n, T \rightarrow \infty} [p^K(x)'(p'_i m_h p_i/T)^- p^K(x)].$$

The reverse inequality holds with  $\lambda_{\min}(\Psi_{i,T})$  replaced by  $\lambda_{\max}(\Psi_{i,T})$ . By Lemma A.1,  $\|p'_i m_h p_i/T - Q_i\| = o_p(1)$ , where  $Q_i$  is defined in (3.1). Hence the exact convergence rate of  $\widehat{g}_i(x)$  depends on  $p^K(x)'Q_i^{-1}p^K(x)$ . Since Assumption 3(iv) implies that  $Q_i$  has the smallest eigenvalue

bounded away from zero and bounded largest eigenvalue, we have

$$[\lambda_{\max}(Q_i)]^{-1} \|p^K(x)\|^2 \leq p^K(x)' Q_i^{-1} p^K(x) \leq [\lambda_{\min}(Q_i)]^{-1} \|p^K(x)\|^2.$$

This implies that the convergence rate of  $\widehat{g}_i(x)$  is given by  $\sqrt{T/K}$  as  $\|p^K(x)\|^2 = O(K)$ .

For statistical inference, we need to estimate the variance  $V_{inT}$ . The case where  $\{\varepsilon_{it}, \mathcal{F}_{it}\}$  is an MDS is trivial. So we only consider the case where  $\{\varepsilon_{it}, \mathcal{F}_{it}\}$  is not an MDS. Since  $\phi$ -mixing implies  $\alpha$ -mixing, we focus on the  $\alpha$ -mixing case. Let

$$S_{inT} \equiv T^{-1} p_i' m_h \Psi_{i,T} m_h p_i. \quad (4.3)$$

Then  $TV_{inT} = p^K(x)' (p_i' m_h p_i / T)^{-1} S_{inT} (p_i' m_h p_i / T)^{-1} p^K(x)$ . A crucial step in the estimation of  $V_{inT}$  is to estimate  $S_{inT}$ . Let  $\widehat{\mathbf{g}}_i = (\widehat{g}_i(x_{i1}), \dots, \widehat{g}_i(x_{iT}))'$ ,  $\widehat{e}_i \equiv m_h(y_i - \widehat{\mathbf{g}}_i)$ , and  $\widehat{p}_i = m_h p_i$ . Following the studies on heteroskedasticity and autocorrelation consistent (HAC) estimation of covariance matrices (e.g., White and Domowitz (1984), Newey and West (1987), Andrews (1991b), and Pesaran (2006)), we propose to estimate  $S_{inT}$  by

$$\widehat{S}_{inT} \equiv \widehat{\Lambda}_{inT,0} + \sum_{j=1}^{l_T} w_{Tj} \left( \widehat{\Lambda}_{inT,j} + \widehat{\Lambda}'_{inT,j} \right), \quad (4.4)$$

where  $\widehat{\Lambda}_{inT,j} \equiv T^{-1} \sum_{t=j+1}^T \widehat{p}_{it} \widehat{p}'_{i,t-j} \widehat{e}_{it} \widehat{e}'_{i,t-j}$ ,  $l_T$  is the window size<sup>2</sup>,  $w_{Tj}$  is a weight function such that  $\sup_j |w_{Tj}| \leq c_w < \infty$  and  $\lim_{T \rightarrow \infty} |w_{Tj}| = 1$  for each  $j$ ,  $\widehat{e}_{it}$  is the  $t$ th element of  $\widehat{e}_i$ , and  $\widehat{p}'_{it}$  is the  $t$ th row of  $\widehat{p}_i$ . A typical choice of  $w_{Tj}$  is  $w_{Tj} = 1 - j / (l_T + 1)$  for  $j \leq l_T$  and 0 otherwise.

Let  $\widehat{V}_{inT} \equiv T p^K(x)' (p_i' m_h p_i)^{-1} \widehat{S}_{inT} (p_i' m_h p_i)^{-1} p^K(x)$  and  $\widehat{A}_{inT} \equiv \widehat{V}_{inT}^{-1/2}$ . The following theorem establishes the consistency of  $\widehat{S}_{inT}$ ,  $\widehat{V}_{inT}$ , and  $\widehat{A}_{inT}$  and justifies the replacement of  $A_{inT}$  by  $\widehat{A}_{inT}$  for statistical inference.

**Theorem 4.3** (*Variance estimation*) *Suppose Assumptions 1-3 and 4(ii) hold and  $\|p^K(x)\| > c > 0$ . If (i)  $\sup_j |w_{Tj}| \leq c_w < \infty$  and  $\lim_{T \rightarrow \infty} |w_{Tj}| = 1$  for each  $j$ , (ii)  $(l_T K) (\sqrt{K/T} + K^{-\lambda_i/d} + \sqrt{K/n}) \rightarrow 0$  and  $l_T^3 K^2 / T \rightarrow 0$  as  $(n, T) \rightarrow \infty$ , (iii) there exists some  $\alpha_0 > 0$  such that  $\sum_{j=1}^{\infty} j^{\alpha_0} \alpha_i^{\eta/(2+\eta)}(j) < \infty$ ,  $\sum_{j=1}^{l_T} l_T^{\alpha_0} \alpha_i^{\eta/(2+\eta)}(j) < \infty$  and  $K l_T^{-\alpha_0} \rightarrow 0$  as  $(n, T) \rightarrow \infty$ , then (i)  $\|\widehat{S}_{inT} - S_{inT}\| = o_p(1)$ ; (ii)  $\widehat{V}_{inT} V_{inT}^{-1} \xrightarrow{p} 1$ ,  $\widehat{A}_{inT} A_{inT}^{-1} \xrightarrow{p} 1$ ; (iii)  $\widehat{A}_{inT} [\widehat{g}_i(x) - g_i(x)] \xrightarrow{d} N(0, 1)$ .*

**Remark 3.** The first additional assumption in Theorem 4.3 is standard in the literature on HAC estimation. The second and third additional assumptions impose conditions on the window size ( $l_T$ ), the number of sieve approximation terms ( $K$ ), and the  $\alpha$ -mixing coefficients

<sup>2</sup>In practice,  $l_T$  can be  $i$ -dependent if one is interested in the estimation of  $g_i(x)$  for certain  $i$ . In this case the notation  $l_{iT}$  may be preferred.

on the stochastic processes. These conditions can easily be satisfied for well chosen  $l_T$  and  $K$ . Nevertheless, there is no such a simple rule as requiring  $l_T = o(T^{1/4})$  in Newey and West (1987). The choice of  $l_T$  and  $K$  relies highly on the mixing coefficients. In particular, if the mixing coefficients decay to zero sufficiently fast such that  $\alpha^{\eta/(2+\eta)}(j) = o(j^{-(\alpha_0+1)})$  for some  $\alpha_0 > 0$ , then we have  $\sum_{j=1}^{\infty} j^{\alpha_0} \alpha^{\eta/(2+\eta)}(j) < \infty$ , and  $\sum_{j=1}^{l_T} l_T^{\alpha_0} \alpha^{\eta/(2+\eta)}(j) < \infty$ . For such  $\alpha_0$ , it is easy to see one can choose  $l_T$  and  $K$  such that other conditions on  $l_T$  and  $K$  are simultaneously met.

**Remark 4.** The above results can be extended to study the estimates of various functionals of  $g_i(\cdot)$ , such as the derivatives, average derivatives, or weighted derivatives of  $g_i(x)$ . Let  $\Phi(g_i)$  denote the estimand, where  $\Phi$  is a function from  $\mathcal{G}_i$  to  $\mathbb{R}^d$  and  $\mathcal{G}_i$  is defined after (1.1). We focus on three cases of  $\Phi$ : a)  $\Phi(g_i) = \partial g_i(x)/\partial x$  for some  $x \in \mathcal{X}_i$ ; b)  $\Phi(g_i) = T^{-1} \sum_{t=1}^T \partial g_i(x_{it})/\partial x$ ; c)  $\Phi(g_i) = \int_{\mathcal{X}_i} \partial g_i(x)/\partial x w(x) dx$  for some weight function  $w(\cdot) : \mathcal{X}_i \rightarrow \mathbb{R}$ . Note that we have suppressed the dependence of  $\Phi$  on  $T$  in b) and one could allow the weight function in c) to depend on  $T$ . For each case, we can estimate  $\Phi(g_i)$  by  $\Phi(\hat{g}_i)$ . Since the functional  $\Phi(\cdot)$  is linear here, we have

$$\Phi(\hat{g}_i) = \Phi(p^K(x)' \hat{\alpha}_{g_i}) = \phi^{K'} \hat{\alpha}_{g_i}$$

where  $\phi^K = (\phi_1, \dots, \phi_K)' \in \mathbb{R}^{K \times d}$ , and  $\phi_k = \Phi(p_k(\cdot)) \in \mathbb{R}^d$  for  $k = 1, \dots, K$ . Clearly, corresponding to cases a)-c), the  $k$ th column of  $\phi^{K'}$  is  $\partial p_k(x)/\partial x$ ,  $T^{-1} \sum_{t=1}^T \partial p_k(x_{it})/\partial x$ , and  $\int_{\mathcal{X}_i} \partial p_k(x)/\partial x w(x) dx$ , respectively. Define  $\tilde{V}_{inT} \equiv \phi^{K'}(p'_i m_h p_i/T)^- S_{inT} (p'_i m_h p_i/T)^- \phi^K$ , and  $\tilde{A}_{inT} = \tilde{V}_{inT}^{-1/2}$ . We make the following additional assumption:

**Assumption D.** (i) There exists some specific norm  $|\cdot|_s$  such that  $\|\Phi(g_i)\| \leq C_1 |g_i|_s$  for some  $C_1 < \infty$  and for any  $g_i \in \mathcal{G}_i$ . (ii)  $\sqrt{T} |g_i(\cdot) - \Pi_{\infty K} g_i(\cdot)|_s \rightarrow 0$  as  $T \rightarrow \infty$ . (iii)  $\lambda_{\min}(\phi^{K'} \phi^K) > c_0 > 0$  for some  $c_0$  as  $T \rightarrow \infty$ .

Then under the conditions of Theorem 4.2 and Assumption D, we can modify the proof of that theorem and show that

$$\tilde{A}_{inT} [\Phi(\hat{g}_i) - \Phi(g_i)] \xrightarrow{d} N(0, 1). \quad (4.5)$$

A consistent estimate of  $\tilde{A}_{inT}$  is also available by replacing  $S_{inT}$  in its definition by  $\hat{S}_{inT}$  defined above. Clearly, if  $\mathcal{X}_i$  is compact as in Newey (1997), it is reasonable to define  $|g_i|_s$  as the Sobolev norm of derivative order 1, i.e.,  $|g_i|_s \equiv \|g_i\|_{1, \mathcal{X}_i} \equiv \sup_{x \in \mathcal{X}_i} |g_i(x)| + \sup_{x \in \mathcal{X}_i} \|\partial g_i(x)/\partial x\|$ , and Assumption D(i) is trivially satisfied for all three cases under investigation (see Andrews (1991a)). When  $\mathcal{X}_i$  is not compact, some weighted norm like the weighted Hölder norm may be desired. Assumption D(ii) is a smoothness condition, which places the same role as Assumption 3(iii). Assumption D(iii) is trivially verified if  $d = 1$  and it otherwise implies that the above result is obtained for vectors of estimands whose asymptotic distribution is nonsingular.

## 5 Asymptotic properties of $\widehat{g}(x)$

In this section, we study the asymptotic properties of  $\widehat{g}(x)$  when (1.4) is correctly specified and both the error term  $e_{it}$  and the individual specific regressors  $x_{it}$  exhibit multi-factor structures defined in (1.2)-(1.3), respectively. We assume that Assumption 3 holds with  $g_i(\cdot)$  replaced by  $g(\cdot)$ . The large  $n$  and large  $T$  assumption is relaxed to large  $n$  only. In other words, as  $n \rightarrow \infty$ ,  $T$  can either be fixed or pass to  $\infty$ , and this dual possibility is denoted by continuing writing  $(n, T) \rightarrow \infty$ .

Let  $\boldsymbol{\gamma}_1 = (\gamma'_{11}, \gamma'_{12}, \dots, \gamma'_{1n})'$ ,  $\boldsymbol{\gamma}_2 = (\gamma'_{21}, \gamma'_{22}, \dots, \gamma'_{2n})'$ ,  $\boldsymbol{\varepsilon} = (\varepsilon'_1, \varepsilon'_2, \dots, \varepsilon'_n)'$ , and  $\mathbf{g} = (\mathbf{g}'_1, \mathbf{g}'_2, \dots, \mathbf{g}'_n)'$ , where  $\mathbf{g}_i$  now becomes  $\mathbf{g}_i = (g(x_{i1}), \dots, g(x_{iT}))'$ . Using (1.4), (1.2), and the approximation for  $g(\cdot)$ , we have

$$Y = P\alpha_g + F_1\boldsymbol{\gamma}_1 + F_2\boldsymbol{\gamma}_2 + \boldsymbol{\varepsilon} + (\mathbf{g} - P\alpha_g), \quad (5.1)$$

where  $F_j = I_n \otimes f_j$  for  $j = 1, 2$ . Therefore

$$\widehat{\alpha}_g - \alpha_g = (P'M_hP)^- P'M_h\boldsymbol{\varepsilon} + (P'M_hP)^- P'M_hF_2\boldsymbol{\gamma}_2 + (P'M_hP)^- P'M_h(\mathbf{g} - P\alpha_g). \quad (5.2)$$

As above, the first term on the r.h.s. of (5.2) is present even if there is no multi-factor error structure and  $g(x)$  is linear in  $x$ ; the second term is due to the presence of the unobserved factor in the error term; and the third term is due to the approximation of  $g$  by  $P\alpha_g$ .

To state the main results, we make the following additional assumption.

**Assumption 5.** (i) Let  $\overline{Q}_n \equiv n^{-1} \sum_{i=1}^n Q_i$  where  $Q_i$  is defined in (3.1). For each  $K \equiv K(n, T)$ , the  $K \times K$  matrix  $\overline{Q}_n$  has the smallest eigenvalue that is bounded away from zero and bounded largest eigenvalue as  $n \rightarrow \infty$ . (ii) As  $n \rightarrow \infty$ ,  $T$  is either fixed or tends to  $\infty$ , and  $K^2/n \rightarrow 0$ . (iii) If  $T$  is fixed,  $\|(nT)^{-1} \sum_{i=1}^n \sum_{t=1}^T [p_{it}p'_{it} - E(p_{it}p'_{it})]\| = o_p(K^{-1/2})$  and  $\|(nT)^{-1} \sum_{i=1}^n \sum_{t=1}^T [p_{it}z'_t - E(p_{it}z'_t)]\| = o_p(K^{-1/2})$  where  $z_t \equiv (f'_{1t}, f'_{2t})'$ ; if  $T \rightarrow \infty$ ,  $K^3/T \rightarrow 0$  as  $(n, T) \rightarrow \infty$ . (iv) Let  $\alpha(j) \equiv \sup_{n \geq 1} \max_{0 \leq i \leq n} \alpha_i(j) \cdot \sum_{j=1}^{\infty} j^2 \alpha(j)^{\eta/(4+\eta)} \leq C_4 < \infty$ ;  $0 < \underline{c}_\Psi \leq \min_{1 \leq i \leq n} \lambda_{\min}(\Psi_{i,T}) \leq \max_{1 \leq i \leq n} \lambda_{\max}(\Psi_{i,T}) \leq \overline{c}_\Psi < \infty$ ;  $\max_{1 \leq i \leq n} \sup_{1 \leq j \leq K} E|p_j(x_{i1})|^{4+\eta} < \infty$  for the same  $\eta$  defined in Assumption 1(i).

Assumption 5(i) requires that the average of  $Q_i$  behave properly. If  $x_{it}$  are identically distributed across  $i$ , then  $Q_i$  does not depend on  $i$ . In this case, we can simply write the  $K \times K$  matrices  $Q_i$  and  $\overline{Q}_n$  as  $Q$ . The first part of Assumption 5(iii) is a high level assumption which can be verified easily in the case where  $T \rightarrow \infty$ , and the second part is weak. Assumption 5(iv) strengthens Assumptions 1(vii)-(viii) and 3(iv).

The following theorem establishes the convergence rate of  $\widehat{g}$ .

**Theorem 5.1** (Convergence rate) *Suppose Assumptions 1-2 and 5 hold. Suppose that Assumption 3 holds with  $g_i(\cdot)$  replaced by  $g(\cdot)$ . If model (1.4) holds in conjunction with (1.2)-(1.3), then  $(nT)^{-1} \sum_{i=1}^n \sum_{t=1}^T [\widehat{g}(x_{it}) - g(x_{it})]^2 = O_p(K^{-2\lambda/d} + K/(nT) + K/n^2)$ , where  $\lambda \equiv \min_{1 \leq i \leq n} \lambda_i$ .*

**Remark 5.** In the above theorem, we only establish the convergence rate of the sample mean square error. If  $x_{it}$  are identically distributed over  $i$  with common support  $\mathcal{X}$  and cumulative distribution function  $F(\cdot)$ , then we can also follow the proof of Theorem 4.2(ii) and show that  $\int_{x \in \mathcal{X}} [\hat{g}(x) - g(x)]^2 dF(x) = O_p(K/(nT) + K^{-2\lambda/d} + K/n^2)$ .

Let  $\Psi_{nT} \equiv \text{Var}(\varepsilon)$ . Let  $V_{nT} \equiv p^K(x)' (P'M_h P)^- P'M_h \Psi_{nT} M_h P (P'M_h P)^- p^K(x)$  and  $A_{nT} \equiv V_{nT}^{-1/2}$ . The following theorem establishes the asymptotic normality of  $\hat{g}(x)$ .

**Theorem 5.2** (*Asymptotic normality*) *Let  $x$  be given and  $\|p^K(x)\| > c$  for some constant  $c > 0$ . Suppose the conditions in Theorem 5.1 hold. Suppose that  $\lambda_{\max}(E(\gamma_2 \gamma_2')) = O(r_n)$ ,  $r_n K/n \rightarrow 0$ ,  $KT/n \rightarrow 0$ , and  $nTK^{-2\lambda/d} \rightarrow 0$  as  $(n, T) \rightarrow \infty$ , where  $\lambda \equiv \min_{1 \leq i \leq n} \lambda_i$ . Then  $A_{nT} [\hat{g}(x) - g(x)] \xrightarrow{d} N(0, 1)$ .*

**Remark 6.** Remarks similar to those after Theorem 4.2 hold here. Since  $A_{nT} = O_p(\sqrt{nT/K})$ ,  $\hat{g}(x)$  has a faster convergence rate than  $\hat{g}_i(x)$ . This reflects the benefit by pooling the data together to estimate the homogeneous regression relationship.

**Remark 7.** For statistical inference, we need to estimate  $V_{nT}$ . Let  $S_{nT} \equiv P'M_h \Psi_{nT} M_h P / (nT)$ . Then

$$nT V_{nT} = p^K(x)' \left( \frac{P'M_h P}{nT} \right)^- S_{nT} \left( \frac{P'M_h P}{nT} \right)^- p^K(x).$$

Note that  $S_{nT} = n^{-1} \sum_{i=1}^n S_{inT}$ , where  $S_{inT}$  is defined in (4.3). We can estimate  $S_{nT}$  by  $\hat{S}_{nT} \equiv n^{-1} \sum_{i=1}^n \hat{S}_{inT}$  in the case of diverging  $T$ , and by  $\hat{S}_{nT} \equiv n^{-1} \sum_{i=1}^n \tilde{S}_{inT}$  in the case of fixed  $T$ , where  $\hat{S}_{inT}$  is defined in (4.4),  $\tilde{S}_{inT} \equiv \hat{\Lambda}_{inT,0} + \sum_{j=1}^{T-1} (\hat{\Lambda}_{inT,j} + \hat{\Lambda}'_{inT,j})$ , and  $\hat{e}_i$  used in the definition of  $\hat{S}_{inT}$  and  $\hat{\Lambda}_{inT,j}$  is now defined by  $\hat{e}_i \equiv m_h(y_i - \hat{\mathbf{g}}_i)$  with  $\hat{\mathbf{g}}_i \equiv (\hat{g}(x_{i1}), \dots, \hat{g}(x_{iT}))'$ . With  $\hat{S}_{nT}$ , we propose to estimate  $A_{nT}$  by  $\hat{A}_{nT}$ , where  $\hat{A}_{nT} = \hat{V}_{nT}^{-1/2}$ , and  $\hat{V}_{nT} \equiv nT p^K(x)' (\sum_{i=1}^n p'_i m_h p_i)^- \hat{S}_{nT} (\sum_{i=1}^n p'_i m_h p_i)^- p^K(x)$ . The validity of the replacement of  $A_{nT}$  by  $\hat{A}_{nT}$  in statistical inference follows directly from Theorem 4.3 in the case of diverging  $T$ . When  $T$  is fixed, we can modify the proofs of Lemma A.8 and Theorem 4.3 and show that  $\hat{A}_{nT} [\hat{g}(x) - g(x)] \xrightarrow{d} N(0, 1)$  by requiring only  $n \rightarrow \infty$ . In particular, we realize that the result in Lemma A.8(i) now becomes  $(nT)^{-1} \|\hat{e} - \varepsilon\|^2 = O_p(K/(nT) + K^{-2\lambda_i/d} + K/n^2)$  where  $\hat{e} = (\hat{e}'_1, \dots, \hat{e}'_n)'$ , and the consistent estimation of  $S_{nT}$  will not rely upon large  $T$ .

## 6 Monte Carlo simulations

In this section we conduct a small set of Monte Carlo simulations to evaluate the finite sample performance of our estimators.

## 6.1 Data generating process

We consider the following data generating process (DGP):

$$\begin{aligned} y_{it} &= g_i(x_{it,1}, x_{it,2}) + \gamma_{1i} + \gamma_{2i,1}f_{2t,1} + \gamma_{2i,2}f_{2t,2} + \varepsilon_{it}, \\ g_i(x_{it,1}, x_{it,2}) &= \exp(x_{it,1}) / (\exp(x_{it,1}) + 1) + \delta_i(0.5x_{it,2} - 0.25x_{it,2}^2), \\ x_{it,s} &= \Gamma_{1i,s} + \Gamma_{2i,s1}f_{2t,1} + \Gamma_{2i,s2}f_{2t,2} + v_{it,s}, \quad s = 1, 2, \end{aligned}$$

for  $i = 1, 2, \dots, n$ , and  $t = 1, 2, \dots, T$ . In this DGP, there are two individual-specific regressors ( $x_{it} = (x_{it,1}, x_{it,2})'$ ), one observed common factor ( $f_{1t} = 1$ ), and two unobserved common factors ( $f_{2t} = (f_{2t,1}, f_{2t,2})'$ ). The heterogeneous regression functions are composed of two parts:  $g_{i1}(x_{it,1}) \equiv \exp(x_{it,1}) / (\exp(x_{it,1}) + 1)$ , and  $g_{i2}(x_{it,2}) = \delta_i(0.5x_{it,2} - 0.25x_{it,2}^2)$ . The former is the same across all cross-section units while the latter is heterogeneous unless  $\delta_i$  remains a constant across  $i$ . In the sequel, we will refer to  $\delta_i$  as the heterogeneous interaction parameter. Noting that  $\gamma_{1i}$  is not separately identifiable from  $g_i(\cdot, \cdot)$ , we are interested in the estimation of  $g_i^c(\cdot) = g_i(\cdot) - E[g_i(x_{it})]$  (see Assumption 1(ix)). Such an identification restriction will be imposed in the estimation procedure.

We next specify how to generate the individual-specific errors, unobserved factors, factor loadings, heterogeneous interaction parameter, and other aspects in the DGP.

1. The individual-specific errors of  $y_{it}$  are generated independently of each other as stationary AR(1) processes with zero means and variance  $\sigma_i^2$ :  $\varepsilon_{it} = \rho_{\varepsilon,i}\varepsilon_{i,t-1} + \sigma_i(1 - \rho_{\varepsilon,i}^2)^{1/2}\varrho_{it}$ , where  $\rho_{\varepsilon,i}$  are IIDU[0, 0.95] across  $i$ ,  $\sigma_i^2$  are IIDU[0.5, 1] across  $i$ , and  $\varrho_{it}$  are IIDN(0, 1) across  $i$  and  $t$ . The individual-specific errors  $v_{it,s}$  of  $x_{it,s}$  ( $s = 1, 2$ ) are generated in the same way as  $\varepsilon_{it}$  are generated. For each  $i$ , the three processes  $\{\varepsilon_{it}\}$ ,  $\{v_{it,1}\}$ , and  $\{v_{it,2}\}$  are generated independently of each other.

2. The unobserved common factors  $f_{2t,s}$  ( $s = 1, 2$ ) are generated as independent stationary AR(1) processes with zero means and variances 1:  $f_{2t,s} = 0.5f_{2,t-1,s} + (1 - 0.5^2)^{1/2}\xi_{it,s}$ , where  $\xi_{it,s}$  are IIDN(0, 1) across  $i$  and  $t$ .

3. The factor loadings of the observed common factor  $\gamma_{1i}$  and  $\Gamma_{1i} = (\Gamma_{1i,1}, \Gamma_{1i,2})'$  are generated as follows:  $\gamma_{1i} = 0.5\bar{x}_{i,1} + 0.5\bar{x}_{i,2}$ , and

$$\Gamma_{1i} \sim \text{IIDN} \left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix} \right),$$

where  $\bar{x}_{i,s} = T^{-1} \sum_{t=1}^T x_{it,s}$ ,  $s = 1, 2$ . The factor loadings  $\gamma_{2i} = (\gamma_{2i,1}, \gamma_{2i,2})'$  of the unobserved common factors in the  $y_{it}$  equation are generated in the same way as  $\Gamma_{1i}$  are generated. For the factor loadings  $\Gamma_{2i}$  of the unobserved common factors in the  $x_{it}$  equation, we consider two different cases that we denote by A and B, respectively:  $\text{vec}(\Gamma_{2i}) = (\Gamma_{2i,11}, \Gamma_{2i,12}, \Gamma_{2i,21}, \Gamma_{2i,22})' \sim \text{IIDN}(\Gamma_{2,\tau}, I_4)$ ,  $\tau = A, B$ . In Case A, the rank condition in Assumption 2(iii) is satisfied

and  $\Gamma_{2,A} = (1, 0, 0, 1)'$ . In Case B, the rank condition in Assumption 2 (iii) is not satisfied and  $\Gamma_{2,B} = (1, 1, 0, 0)'$ .

4. The heterogeneous interaction parameters are generated according to two experiment designs: Experiment 1:  $\delta_i \sim \text{IIDU}(0, 1)$ . Experiment 2:  $\delta_i = 0.5$  for each  $i$ . So Experiment 1 corresponds to heterogeneous regression functions whereas Experiment 2 corresponds to homogeneous regression functions.

In the following experiments, we only generate  $\gamma_{2i}$ ,  $\Gamma_{1i}$  and  $\Gamma_{2i}$  for  $i = 1, 2, \dots, n$  once and they are kept fixed across replications. Other variables or parameters are generated independently across replications.

## 6.2 Estimators and evaluation criterion

Under Experiment 1, after generating the data on  $y_{it}$ ,  $x_{it}$ , and  $f_{2t}$  (recall  $f_{1t} = 1$  here), we compute the CCE estimator  $\hat{g}_i(x)$  of  $g_i(x)$  for each  $i$ . In addition, we compute three other estimators: (1) the infeasible estimator  $\hat{g}_i^{(IF)}(x)$  that includes the unobserved common factors  $f_{2t}$  in the sieve regression of  $y_{it}$  on  $(x_{it}, 1)$ , (2) the naive estimator  $\hat{g}_i^{(N)}(x)$  that excludes  $f_{2t}$  in the sieve regression of  $y_{it}$  on  $(x_{it}, 1)$ , and (3) Pesaran's (2006) CCE estimator of  $g_i(x)$  by using the linear specification for  $g_i(x)$ . The infeasible estimator  $\hat{g}_i^{(IF)}(x)$  provides an upper bound to the efficiency of the CCE estimator whereas the naive estimator  $\hat{g}_i^{(N)}(x)$  signifies the bias due to the neglect of unobserved common factors. Similarly, Pesaran's (2006) CCE estimator  $\hat{g}_i^{(P)}(x)$  indicates the bias due to functional form misspecification.

To obtain the first three estimators, we need to choose sieve bases. We have tried both cubic spline polynomials and Hermite polynomials (see Blundell, Chen, and Kristensen, 2007) and found that the results are quantitatively similar. So we will only focus on the case of cubic spline polynomials. Since  $g_i(x_1, x_2)$  has the additive structure and can be written as the sum of  $g_{i1}(x_1)$  and  $g_{i2}(x_2)$ , we approximate either component by the cubic splines:

$$p_s^{J+4} = [1, x_s, x_s^2, x_s^3, (x_s - v_{s1})_+^3, \dots, (x_s - v_{sJ})_+^3]' \quad (6.1)$$

where  $(x_s - v)_+^3 = \max\{(x_s - v)^3, 0\}$ ,  $s = 1, 2$ ,  $j = 1, \dots, J$ . Here  $\{v_{sj}\}_{j=1}^J$  are the knots. In the simulation, for any given number of knots value  $J$ , the knots  $\{v_{sj}\}_{j=1}^J$  are simply chosen as the empirical quantiles of  $x_{it,s}$ , i.e.,  $v_{sj} = j/(J+1)$ -th quantile of  $x_{it,s}$ . Note that the convergence rates of the estimators mainly depend on the time dimension  $T$  and the cross-section dimension does not play any essential roles. To evaluate how the estimators are sensitive to the choice of  $J$ , we will consider choosing  $J = c\lceil T^{1/5} \rceil$  for different values of  $c$ , where  $\lceil a \rceil$  denotes the integer part of  $a$ . In this case, the total number of approximating terms in the sieve base is given by  $K = 6 + 2c\lceil T^{1/5} \rceil$ . (We delete the column of ones in the construction of  $p_i$  to avoid perfect collinearity as  $f_{1t} \equiv 1$ .) We will consider the  $(n, T)$  pairs with  $n, T = 25, 50, 100$ . For evaluation, we first calculate the root mean squared error

(RMSE) for each replication, for example,  $\text{RMSE}(\hat{g}) = \sqrt{\frac{1}{nT} \sum_{i=1}^n \sum_{t=1}^T [\hat{g}_i(x_{it}) - g_i^c(x_{it})]^2}$ , and then obtain the final RMSE by averaging  $\text{RMSE}(\hat{g})$  across replications.

Under Experiment 2, we consider five estimators of  $g(x)$ : our CCE pooled estimator  $\hat{g}(x_{it})$ , the infeasible estimator  $\hat{g}^{(IF)}(x)$ , the naive estimator  $\hat{g}^{(N)}(x)$ , Pesaran's (2006) CCE mean group (CCEMG) and CCE pooled (CCEP) estimators, where  $\hat{g}^{(IF)}(x)$  and  $\hat{g}^{(N)}(x)$  are defined analogously to  $\hat{g}_i^{(IF)}(x)$  and  $\hat{g}_i^{(N)}(x)$ , respectively. The RMSEs of these estimators are defined in the same way as above. We also used the cubic splines to construct the sieve bases. Noting that the cross-section and time dimensions are equally important to the convergence rates so we will consider the  $(n, T)$  pairs with  $n = 25, 50, 100$ , and  $T = 8, 25, 100$ , and choose  $J = c \lfloor (nT)^{1/5} \rfloor$  in the above definition of cubic splines.

In each scenario, the number of replications in the Monte Carlo study is 10,000.

### 6.3 Simulation results

Table 1 reports the results of Experiment 1 for estimating heterogeneous regression functions. In the case where the full rank condition in (2.3) is satisfied, we summarize the main findings from the upper panel of Table 1. (a) The choice of  $J$  (or equivalently  $K$ ) has some effect on the RMSEs of our CCE estimator, the infeasible estimator, and the naive estimators, but the effect is not large. (b) The increase of  $n$  does not help with the estimation of the unknown heterogeneous regression relationships, which is as expected since larger  $n$  implies more heterogeneous relationships and the convergence rate of  $\hat{g}_i(x)$  mainly depends on the time dimension. (c) As  $T$  increases, RMSEs decrease for all estimators. (d) For all the sample sizes under investigation, our CCE estimator significantly outperforms Pesaran's (2006) CCE estimator, and the naive estimator that ignores the multi-factor error structure performs poorly in all cases. (e) Compared with the infeasible estimator, our CCE estimator loses little efficiency. For example, when  $(n, T) = (100, 100)$ , the RMSE of our CCE estimator is about 4% higher than that of the infeasible estimator for all  $J$  under investigation. The lower panel in Table 1 reports the results for the case when the rank condition in (2.3) is not satisfied. Clearly, in the rank deficient case, our CCE estimator is outperformed significantly by the infeasible estimator in terms of RMSE. But it still dominates the naive estimator and Pesaran's (2006) estimator in all cases.

Table 2 reports the results of Experiment 2 for estimating homogeneous regression functions. The upper panel of Table 2 reports the results for the full rank case. We find that: (a) as in the case of estimating heterogeneous regression functions, the effect of  $J$  on the RMSEs of our CCE pooled estimator, the infeasible estimator, and the naive estimators is not large; (b) as either  $n$  or  $T$  increases, the RMSEs of our CCE pooled estimator, and the infeasible estimators decrease significantly as expected; (c) in comparison with the infeasible estimator, our CCE pooled estimator has little efficiency loss for large  $n$  even when  $T$  is

Table 1: RMSE comparison in Experiment 1 (heterogeneous regression)

Estimator	$n \setminus T$	$J = \lfloor T^{1/5} \rfloor$			$J = 2 \lfloor T^{1/5} \rfloor$			$J = 3 \lfloor T^{1/5} \rfloor$		
		25	50	100	25	50	100	25	50	100
Case A: full rank										
$\widehat{g}_i(x)$	25	1.061	0.736	0.538	1.155	0.795	0.570	1.276	0.858	0.600
	50	0.932	0.646	0.457	1.021	0.707	0.492	1.129	0.769	0.526
	100	0.996	0.674	0.470	1.089	0.736	0.505	1.205	0.798	0.538
$\widehat{g}_i^{(IF)}(x)$	25	1.032	0.705	0.493	1.114	0.762	0.526	1.214	0.822	0.558
	50	0.878	0.616	0.431	0.952	0.673	0.466	1.037	0.731	0.499
	100	0.930	0.646	0.452	1.005	0.704	0.486	1.096	0.762	0.518
$\widehat{g}_i^{(N)}(x)$	25	1.228	1.197	1.158	1.267	1.239	1.181	1.304	1.280	1.204
	50	1.188	1.152	1.101	1.230	1.199	1.126	1.271	1.244	1.152
	100	1.235	1.196	1.142	1.279	1.244	1.169	1.321	1.292	1.196
$\widehat{g}_i^{(P)}(x)$	25	1.622	1.524	1.493						
	50	1.028	0.916	0.862						
	100	1.098	0.963	0.897						
Case B: rank deficient										
$\widehat{g}_i(x)$	25	1.112	0.906	0.796	1.186	0.953	0.818	1.279	1.003	0.840
	50	0.961	0.731	0.587	1.041	0.786	0.616	1.138	0.842	0.645
	100	1.062	0.859	0.745	1.137	0.910	0.770	1.231	0.961	0.795
$\widehat{g}_i^{(IF)}(x)$	25	1.034	0.707	0.494	1.116	0.764	0.527	1.215	0.824	0.558
	50	0.889	0.622	0.435	0.963	0.680	0.470	1.050	0.738	0.503
	100	0.925	0.644	0.450	1.000	0.701	0.484	1.091	0.759	0.516
$\widehat{g}_i^{(N)}(x)$	25	1.245	1.220	1.187	1.281	1.259	1.208	1.316	1.297	1.229
	50	1.220	1.196	1.156	1.258	1.237	1.178	1.294	1.277	1.200
	100	1.263	1.234	1.190	1.302	1.277	1.213	1.341	1.320	1.237
$\widehat{g}_i^{(P)}(x)$	25	1.298	1.242	1.228						
	50	0.898	0.791	0.735						
	100	1.040	0.962	0.927						

Note:  $\widehat{g}_i$ ,  $\widehat{g}_i^{(IF)}$ ,  $\widehat{g}_i^{(N)}$ , and  $\widehat{g}_i^{(P)}$  refer to our CCE estimator, the infeasible estimator, the naive estimator, and Pesaran's (2006) estimator, respectively.

Table 2: RMSE comparison in Experiment 2 (homogeneous regression)

Estimator	$n \setminus T$	$J = \lfloor (nT)^{1/5} \rfloor$			$J = 2 \lfloor (nT)^{1/5} \rfloor$			$J = 3 \lfloor (nT)^{1/5} \rfloor$		
		8	25	100	8	25	100	8	25	100
Case A: full rank										
$\widehat{g}(x)$	25	0.528	0.245	0.143	0.575	0.263	0.152	0.615	0.278	0.161
	50	0.344	0.164	0.095	0.380	0.180	0.104	0.411	0.195	0.111
	100	0.245	0.115	0.065	0.268	0.126	0.072	0.287	0.136	0.078
$\widehat{g}^{(IF)}(x)$	25	0.394	0.205	0.110	0.427	0.224	0.121	0.456	0.240	0.131
	50	0.268	0.139	0.073	0.296	0.156	0.083	0.321	0.170	0.092
	100	0.192	0.101	0.055	0.210	0.112	0.063	0.226	0.122	0.070
$\widehat{g}^{(N)}(x)$	25	0.563	0.524	0.494	0.607	0.548	0.503	0.644	0.569	0.511
	50	0.447	0.410	0.386	0.483	0.429	0.392	0.516	0.446	0.398
	100	0.291	0.241	0.208	0.325	0.261	0.217	0.353	0.277	0.224
$\widehat{g}^{(P1)}(x)$	25	1.548	1.492	1.560						
	50	0.876	0.813	0.842						
	100	0.899	0.885	0.913						
$\widehat{g}^{(P2)}(x)$	25	1.385	1.489	1.560						
	50	0.771	0.810	0.842						
	100	0.838	0.884	0.913						
Case B: rank deficient										
$\widehat{g}(x)$	25	0.535	0.358	0.312	0.581	0.373	0.318	0.623	0.387	0.323
	50	0.333	0.176	0.120	0.372	0.193	0.127	0.406	0.208	0.135
	100	0.245	0.137	0.099	0.273	0.150	0.106	0.297	0.162	0.112
$\widehat{g}^{(IF)}(x)$	25	0.395	0.206	0.110	0.427	0.224	0.121	0.456	0.240	0.131
	50	0.267	0.139	0.074	0.295	0.155	0.083	0.321	0.170	0.092
	100	0.192	0.101	0.055	0.210	0.112	0.063	0.227	0.122	0.069
$\widehat{g}^{(N)}(x)$	25	0.521	0.461	0.412	0.569	0.490	0.424	0.609	0.514	0.434
	50	0.446	0.406	0.378	0.484	0.426	0.386	0.517	0.443	0.392
	100	0.328	0.295	0.275	0.358	0.311	0.282	0.384	0.326	0.288
$\widehat{g}^{(P1)}(x)$	25	1.237	1.104	1.121						
	50	0.734	0.624	0.623						
	100	0.782	0.734	0.739						
$\widehat{g}^{(P2)}(x)$	25	1.081	1.108	1.136						
	50	0.634	0.626	0.628						
	100	0.725	0.732	0.739						

Note:  $\widehat{g}$ ,  $\widehat{g}^{(IF)}$ ,  $\widehat{g}^{(N)}$ ,  $\widehat{g}^{(P1)}$ , and  $\widehat{g}^{(P2)}$  refer to our CCE estimator, the infeasible estimator, the naive estimator, and Pesaran's (2006) CCEMG and CCEP estimators, respectively.

small; (d) Pesaran’s CCEMG and CCEP estimators are significantly outperformed by our CCE pooled estimators, and the RMSEs of these parametric estimators may not decrease at all as either  $n$  or  $T$  increases. The lower panel of Table 2 reports the results for the rank deficient case. We find that: (a) as either  $n$  or  $T$  increases, the RMSEs of our CCE pooled estimator and the naive estimator also decrease but at a smaller rate than the full rank case; (b) the infeasible estimators are largely unaffected by the rank deficiency as expected; (c) Pesaran’s CCEMG and CCEP estimators are significantly outperformed by all nonparametric estimators and their RMSEs may not decrease at all as either  $n$  or  $T$  increases; (d) our CCE pooled estimators have larger efficiency loss relative to the infeasible estimator than the full rank case.

## 6.4 Discussion

In the above subsection we have imposed the additive separability to obtain the nonparametric sieve estimates when we estimate the nonparametric heterogenous or homogeneous regression function. The same restriction was also imposed for the Pesaran’s parametric estimates. In this sense, we think that the comparison of these estimates is fair.

Nevertheless, as a referee remarked, in practice the additivity of the nonparametric relationship may be unknown to us and it is advisable to estimate the nonparametric relationship without imposing additivity. To this goal, we now consider the case where the additivity of  $g_i(x_1, x_2)$  or  $g(x_1, x_2)$  is ignored in our sieve estimation procedure. In this case, we also include terms by interacting the univariate cubic splines, i.e., terms that are formed from the product of elements of  $p_1^{J+4}$  and those of  $p_2^{J+4}$ , where recall  $p_s^{J+4}$  ( $s = 1, 2$ ) is defined in (6.1). After we delete redundant terms and 1 (as  $f_{1t} \equiv 1$ ) in the construction of  $p_i$ , the total number of terms in  $p_i$  is  $K \equiv (J + 3)^2 + 2(J + 3) = (J + 4)^2 - 1$ . Even for as small values as  $J = 3, 4, 5$ , this would result in  $K = 48, 63$ , and 80 terms for the sieve estimation. Therefore one cannot expect to estimate the heterogenous regression function  $g_i(x_1, x_2)$  very well for the values of  $T$  ( $= 25, 50, 100$ ) under study, and we only focus on the estimation of the homogenous regression function  $g(x_1, x_2)$ .

Table 3 reports the results of Experiment 2 for estimating homogeneous regression functions when the additivity of  $g(x_1, x_2)$  is not imposed. The upper and lower panels of Table 3 report the results for the full rank and rank deficient cases, respectively. We find that the results in Table 3 are comparable with the corresponding components in Table 2. The noticeable difference is that when additivity is not imposed, the inclusion of the interaction terms tend to yield estimates with larger variance and thus larger RMSE than the case where additivity is imposed correctly. Despite this, the nonparametric sieve estimates outperform Pesaran’s CCEMG and CCEP estimates (in Table 2) significantly.

Table 3: RMSE comparison in Experiment 2 (homogeneous regression): no additive separability is imposed

Estimator	$n \setminus T$	$J = \lfloor (nT)^{1/5} \rfloor$			$J = 2 \lfloor (nT)^{1/5} \rfloor$			$J = 3 \lfloor (nT)^{1/5} \rfloor$		
		8	25	100	8	25	100	8	25	100
Case A: full rank										
$\widehat{g}(x)$	25	1.102	0.373	0.205	2.368	0.473	0.249	6.212	0.540	0.273
	50	0.728	0.267	0.145	1.333	0.336	0.173	2.042	0.372	0.186
	100	0.460	0.189	0.106	0.659	0.239	0.129	0.846	0.268	0.138
$\widehat{g}^{(IF)}(x)$	25	0.681	0.315	0.172	1.052	0.402	0.217	1.748	0.457	0.240
	50	0.484	0.232	0.125	0.716	0.294	0.154	0.928	0.324	0.166
	100	0.324	0.166	0.097	0.434	0.213	0.119	0.520	0.239	0.129
$\widehat{g}^{(N)}(x)$	25	0.879	0.775	0.678	1.055	0.894	0.733	1.190	0.963	0.764
	50	0.671	0.560	0.475	0.832	0.644	0.505	0.929	0.686	0.519
	100	0.512	0.413	0.328	0.650	0.498	0.367	0.735	0.543	0.383
Case B: rank deficient										
$\widehat{g}(x)$	25	1.034	0.481	0.369	2.031	0.573	0.404	4.698	0.634	0.423
	50	0.707	0.287	0.174	1.351	0.359	0.202	2.176	0.398	0.215
	100	0.479	0.234	0.154	0.686	0.296	0.181	0.873	0.331	0.193
$\widehat{g}^{(IF)}(x)$	25	0.681	0.317	0.173	1.042	0.402	0.216	1.657	0.454	0.238
	50	0.484	0.231	0.125	0.734	0.295	0.155	0.992	0.327	0.168
	100	0.323	0.166	0.097	0.434	0.213	0.119	0.521	0.239	0.129
$\widehat{g}^{(N)}(x)$	25	0.870	0.758	0.650	1.048	0.878	0.703	1.181	0.946	0.733
	50	0.680	0.566	0.475	0.850	0.659	0.509	0.948	0.706	0.526
	100	0.543	0.459	0.389	0.677	0.540	0.424	0.760	0.584	0.440

Note:  $\widehat{g}$ ,  $\widehat{g}^{(IF)}$ , and  $\widehat{g}^{(N)}$  refer to our CCE estimator, the infeasible estimator, and the naive estimator, respectively.

## 7 Concluding remarks

In this paper we propose sieve estimation of semiparametric panel data models with multi-factor error structure. We develop the asymptotic theory under fairly general conditions when both the cross section and time dimensions are large. If only homogenous regression relationships are of interest, the time dimension need not pass to infinity. Our simulation results indicate that the proposed estimators are well behaved for both heterogenous and homogeneous regressions when the rank condition is satisfied. If the rank condition is violated, our CCE pooled estimators deteriorate a little but still outperform both naive and parametric estimators.

Our asymptotic results can be useful in several aspects. First, they serve as a base for testing the linearity of  $g_i(x)$  or  $g(\cdot)$ . Either the generalized likelihood ratio test of Fan, Zhang, and Zhang (2001) or the consistent specification test of Li, Hsiao, and Zinn (2003) can be extended to our framework. Second, one can consider testing the constancy of the nonparametric relationship over individuals in the presence of multi-factor error structure. Possible approaches include but are not limited to those of Baltagi, Hidalgo, and Li (1996) (or Fan and Li (1996)), Stinchcombe and White (1998), and Vilar-Fernández and González-Manteiga (2004). Third, one may consider the estimation of the factor loadings and the unobserved factors as well, say, by extending the Bai and Ng's (2002) and Bai's (2003) procedures to our framework.

In addition, it is worth mentioning the limitation of our model. As a referee noted, the restriction in (1.3) is fundamental to the approach taken here. If it is violated, the proposed CCE estimators are likely not to work any more. In this case, if we are only interested in estimating the homogeneous regression functions  $g(\cdot)$ , we conjecture that one can extend the seminal work of Bai (2009) to our framework by combining the sieve method with the principal component analysis. We leave this for future research.

# Appendix

Let  $C$  signify a generic constant whose exact value may vary from case to case. Let  $\mathcal{D} = \{(x_{it}, f_{1t}, f_{2t}) : i = 1, \dots, n, t = 1, \dots, T\}$ . Let  $E_{\mathcal{D}}(\cdot)$  and  $\text{Var}_{\mathcal{D}}(\cdot)$  denote the conditional expectation and variance given  $\mathcal{D}$ , respectively.

## A Proof of results in section 4

In this appendix, we first state several lemmas that are used in the proof of the main results in Section 4, and then prove Theorems 4.1-4.3. The proof of all lemmas can be found at [http://www.mysmu.edu/faculty/ljsu/Publications/Sieve10\\_supp.pdf](http://www.mysmu.edu/faculty/ljsu/Publications/Sieve10_supp.pdf).

**Lemma A.1** *Suppose Assumptions 1-2 and 3(iv) hold, then (i)  $E\|T^{-1}p'_i p_i - Q_{ipp}\|^2 = O(K^2/T)$ ; (ii)  $E\|T^{-1}p'_i h - Q_{iph}\|^2 = O(K/T)$ ; (iii)  $\|T^{-1}p'_i m_h p_i - Q_i\| = O_p(K/\sqrt{T})$ ; (iv)  $\lambda_{\max}(p'_i m_h p_i/T) = \lambda_{\max}(Q_i) + O_p(K/\sqrt{T}) = O_p(1)$ ; (v)  $\lambda_{\min}(p'_i m_h p_i/T) \geq \lambda_{\min}(Q_i)/2$  with probability approaching 1 (w.p.a.1).*

**Lemma A.2** *Suppose Assumptions 3(i)-(iii) hold, then  $T^{-1}E\|\mathbf{g}_i - p_i a_{g_i}\|^2 = O(K^{-2\lambda_i/d})$ .*

**Lemma A.3** *Suppose Assumptions 1(i), (v), (vi), and (viii) hold, then  $nE[\bar{v}\bar{v}'] \leq CI_T$  and  $nE[\bar{\varepsilon}\bar{\varepsilon}'] \leq CI_T$  for some  $C < \infty$ , where  $\bar{v} = (\bar{v}_1, \bar{v}_2, \dots, \bar{v}_T)'$  and  $\bar{\varepsilon} = (\bar{\varepsilon}_1, \bar{\varepsilon}_2, \dots, \bar{\varepsilon}_T)'$ .*

**Lemma A.4** *Suppose Assumptions 1(iii),(iv) and (vii), 2(i)-(ii), and 3(iv) hold, then  $E\|T^{-1}p'_i m_h \varepsilon_i\|^2 = O(K/T)$ .*

Recall  $h_t \equiv (f'_{1t}, \bar{x}'_t, \bar{y}'_t)'$ . Let  $z_t \equiv (f'_{1t}, f'_{2t})'$ . Then by (2.2), we have

$$h_t = \Gamma' z_t + \bar{v}_t^*, \quad (\text{A.1})$$

where

$$\Gamma_{(q_1+q_2) \times (q_1+d+1)} = \begin{pmatrix} I_{q_1} & \bar{\Gamma}_1 & \bar{\gamma}_1 \\ 0 & \bar{\Gamma}_2 & \bar{\gamma}_2 \end{pmatrix}, \quad \bar{v}_t^*{}'_{1 \times (q_1+d+1) \times 1} = \begin{pmatrix} 0' & \bar{v}_t' & \bar{g}_t + \bar{\varepsilon}_t \end{pmatrix}. \quad (\text{A.2})$$

Let  $z \equiv (z_1, z_2, \dots, z_T)'$ , and  $\bar{v}^* \equiv (\bar{v}_1^*, \bar{v}_2^*, \dots, \bar{v}_T^*)'$ . We can rewrite (A.1) in matrix form:  $h = z\Gamma + \bar{v}^*$ . Let  $\mathbf{h} = \text{diag}(\|h_1\|, \dots, \|h_T\|)$  and  $\mathbf{p}_i = \text{diag}(\|p_{i1}\|, \dots, \|p_{iT}\|)$ . The following two lemmas study the approximation error due to the replacement of  $z\Gamma$  by  $h$ .

**Lemma A.5** *Suppose Assumptions 1, 2(i)-(ii) and 3(iv) hold, then (i)  $E\|T^{-1}p'_i h - T^{-1}p'_i z\Gamma\| = O(\sqrt{K/n})$ ; (ii)  $E\|T^{-1}f'_2 h - T^{-1}f'_2 z\Gamma\| = O(1/\sqrt{n})$ ; (iii)  $E\|T^{-1}(h'h - \Gamma'z'z\Gamma)\| = O(1/\sqrt{n})$ . If in addition Assumption 2(iii) holds, then (iv)  $m_b f_2 = 0$  w.p.a.1 as  $n \rightarrow \infty$ , where  $b \equiv z\Gamma$  and  $m_b \equiv I_T - b(b'b)^{-1}b'$ ; (v)  $\|T^{-1}(p'_i m_h f_2 - p'_i m_b f_2)\gamma_{2i}\| = O_p(\sqrt{K/n})$ ; (vi)  $\|T^{-1}(p'_i m_h \varepsilon_i - p'_i m_b \varepsilon_i)\| = O_p(\sqrt{K/nT})$ .*

**Lemma A.6** Suppose Assumptions 1, 2(i)-(ii), and 3(iv) hold, then (i)  $E \|T^{-1}(h - z\Gamma)\| = O(\sqrt{1/nT})$ ,  $E \|T^{-1}\mathbf{h}(h - z\Gamma)\| = O(\sqrt{1/nT})$ , and  $E \|T^{-1}\mathbf{p}_i(h - z\Gamma)\| = O(\sqrt{K/nT})$ ; (ii)  $\|T^{-1}(m_h f_2 - m_b f_2)\| = O_p(\sqrt{1/nT})$ ,  $\|T^{-1}\mathbf{h}(m_h f_2 - m_b f_2)\| = O_p(\sqrt{1/nT})$ , and  $\|T^{-1}\mathbf{p}_i(m_h f_2 - m_b f_2)\| = O_p(\sqrt{K/nT})$ .

The next lemma is used in the proof of Theorem 4.2.

**Lemma A.7** Let  $x \in \mathcal{X}_i$  be given and  $\|p^K(x)\| > c$  for some constant  $c > 0$ . Suppose Assumptions 1, 2, 3(iv), and 4 hold, then  $T_{inT} \equiv A_{inT} p^K(x)' (p'_i m_h p_i)^- p'_i m_h \varepsilon_i \xrightarrow{d} N(0, 1)$ .

Let  $\hat{\varepsilon}_i = (\hat{\varepsilon}_{i1}, \hat{\varepsilon}_{i2}, \dots, \hat{\varepsilon}_{iT})'$ . The following lemma is used in the proof of Theorem 4.3.

**Lemma A.8** Suppose Assumptions 1-4 hold, then (i)  $T^{-1} \|\hat{\varepsilon}_i - \varepsilon_i\|^2 = O_p(K/T + K^{-2\lambda_i/d} + K/n)$ ; (ii)  $T^{-1} \|\mathbf{h}(\hat{\varepsilon}_i - \varepsilon_i)\|^2 = T^{-1} \sum_{t=1}^T \|h_t\|^2 (\hat{\varepsilon}_{it} - \varepsilon_{it})^2 = O_p(K(K/T + K^{-2\lambda_i/d} + K/n))$ ; (iii)  $T^{-1} \|\mathbf{p}_i(\hat{\varepsilon}_i - \varepsilon_i)\|^2 = T^{-1} \sum_{t=1}^T \|p_{it}\|^2 (\hat{\varepsilon}_{it} - \varepsilon_{it})^2 = O_p(K^2(K/T + K^{-2\lambda_i/d} + 1/n))$ .

#### Proof of Theorem 4.1

By (4.2),

$$\begin{aligned} \hat{\alpha}_{g_i} - \alpha_{g_i} &= (p'_i m_h p_i)^- p'_i m_h \varepsilon_i + (p'_i m_h p_i)^- p'_i m_h f_2 \gamma_{2i} + (p'_i m_h p_i)^- p'_i m_h (\mathbf{g}_i - p_i \alpha_{g_i}) \\ &\equiv D_{i1} + D_{i2} + D_{i3}. \end{aligned} \quad (\text{A.3})$$

By Assumption 3(iv), Lemmas A.1(v) and A.4,

$$\begin{aligned} \|D_{i1}\|^2 &= \varepsilon'_i m_h p_i (p'_i m_h p_i)^- (p'_i m_h p_i)^- p'_i m_h \varepsilon_i \leq [\lambda_{\min}(p'_i m_h p_i/T)]^{-2} \|T^{-1} p'_i m_h \varepsilon_i\|^2 \\ &= O_p(1) O_p(K/T) = O_p(K/T). \end{aligned} \quad (\text{A.4})$$

Similarly, by Assumption 3(iv), Lemmas A.1(v) and A.5(iv)-(v),

$$\begin{aligned} \|D_{i2}\|^2 &= \gamma'_{2i} f'_2 m_h p_i (p'_i m_h p_i/T)^- (p'_i m_h p_i/T)^- p'_i m_h f_2 \gamma_{2i} / T^2 \\ &\leq [\lambda_{\min}(p'_i m_h p_i/T)]^{-2} \left\{ \|p'_i m_h f_2 \gamma_{2i}\|^2 / T^2 \right\} \\ &= O_p(1) O_p(K/n) = O_p(K/n). \end{aligned} \quad (\text{A.5})$$

Now, let  $w_i \equiv m_h p_i (p'_i m_h p_i)^- p'_i m_h$ . Noting that  $w_i$  is a projection matrix and thus positive semidefinite (p.s.d.) with  $\lambda_{\max}(w_i) = 1$ , we have by Assumptions 3(iv), Lemmas A.1(v) and A.2,

$$\begin{aligned} \|D_{i3}\|^2 &= (\mathbf{g}_i - p_i \alpha_{g_i})' m_h p_i (p'_i m_h p_i)^- (p'_i m_h p_i)^- p'_i m_h (\mathbf{g}_i - p_i \alpha_{g_i}) \\ &\leq [\lambda_{\min}(p'_i m_h p_i/T)]^{-1} (\mathbf{g}_i - p_i \alpha_{g_i})' w_i (\mathbf{g}_i - p_i \alpha_{g_i}) / T \\ &\leq [\lambda_{\min}(p'_i m_h p_i/T)]^{-1} \lambda_{\max}(w_i) \left\{ \|\mathbf{g}_i - p_i \alpha_{g_i}\|^2 / T \right\} \\ &= O_p(1) O_p(K^{-2\lambda_i/d}) = O_p(K^{-2\lambda_i/d}). \end{aligned} \quad (\text{A.6})$$

By (A.3)-(A.6) and the triangle inequality,

$$\|\widehat{\alpha}_{g_i} - \alpha_{g_i}\| \leq \|D_{i1}\| + \|D_{i2}\| + \|D_{i3}\| = O_p(\sqrt{K/T} + \sqrt{K/n} + K^{-\lambda_i/d}). \quad (\text{A.7})$$

Then by (A.7), Lemmas A.1(i) and A.2, Assumption 3(iv), and the Markov inequality, we have

$$\begin{aligned} & \frac{1}{T} \sum_{t=1}^T [\widehat{g}_i(x_{it}) - g_i(x_{it})]^2 \\ &= \frac{1}{T} \sum_{t=1}^T \{p^K(x_{it})' (\widehat{\alpha}_{g_i} - \alpha_{g_i}) + [p^K(x_{it})' \alpha_{g_i} - g_i(x_{it})]\}^2 \\ &\leq 2(\widehat{\alpha}_{g_i} - \alpha_{g_i})' (T^{-1} p_i' p_i) (\widehat{\alpha}_{g_i} - \alpha_{g_i}) + 2T^{-1} \|\mathbf{g}_i - p_i \alpha_{g_i}\|^2 \\ &\leq 2\lambda_{\max}(T^{-1} p_i' p_i) \|\widehat{\alpha}_{g_i} - \alpha_{g_i}\|^2 + 2T^{-1} \|\mathbf{g}_i - p_i \alpha_{g_i}\|^2 \\ &= O_p(1) O_p(K^{-2\lambda_i/d} + K/T + K/n) + O_p(K^{-2\lambda_i/d}) = O_p(K^{-2\lambda_i/d} + K/T + K/n). \end{aligned}$$

In addition, noting that  $\int_{\mathcal{X}_i} p^K(x) p^K(x)' dF_i(x) = Q_{ipp}$ , by Lemma A.2 we have

$$\begin{aligned} & \int_{\mathcal{X}_i} [\widehat{g}_i(x) - g_i(x)]^2 dF_i(x) \\ &= \int_{\mathcal{X}_i} \{p^K(x)' (\widehat{\alpha}_{g_i} - \alpha_{g_i}) + [p^K(x)' \alpha_{g_i} - g_i(x)]\}^2 dF_i(x) \\ &\leq 2(\widehat{\alpha}_{g_i} - \alpha_{g_i})' Q_{ipp} (\widehat{\alpha}_{g_i} - \alpha_{g_i}) + 2 \int_{\mathcal{X}_i} [p^K(x)' \alpha_{g_i} - g_i(x)]^2 dF_i(x) \\ &\leq 2\lambda_{\max}(Q_{ipp}) \|\widehat{\alpha}_{g_i} - \alpha_{g_i}\|^2 + 2T^{-1} E \|\mathbf{g}_i - p_i \alpha_{g_i}\|^2 \\ &= O_p(K^{-2\lambda_i/d} + K/T + K/n) + O_p(K^{-2\lambda_i/d}) = O_p(K^{-2\lambda_i/d} + K/T + K/n). \blacksquare \end{aligned}$$

#### Proof of Theorem 4.2

Recall  $V_{inT} = p^K(x)' (p_i' m_h p_i)^{-} p_i' m_h \text{Var}(\varepsilon_i) m_h p_i (p_i' m_h p_i)^{-} p^K(x)$  and  $A_{inT} = V_{inT}^{-1/2}$ . By Assumptions 1(vii) and 3(iv), the fact that  $\|p^K(x)\| > c > 0$  for some constant  $c$ , and Lemma A.1,

$$\begin{aligned} TV_{inT} &= T p^K(x)' (p_i' m_h p_i)^{-} p_i' m_h \text{Var}(\varepsilon_i) m_h p_i (p_i' m_h p_i)^{-} p^K(x) \\ &\geq T \lambda_{\min}(\text{Var}(\varepsilon_i)) p^K(x)' (p_i' m_h p_i)^{-} p^K(x) \\ &\geq \lambda_{\min}(\text{Var}(\varepsilon_i)) \|p^K(x)\|^2 / \lambda_{\max}(p_i' m_h p_i / T) > C > 0 \end{aligned} \quad (\text{A.8})$$

for large  $(n, T)$ . Write

$$\begin{aligned} & A_{inT} [\widehat{g}_i(x) - g_i(x)] \\ &= A_{inT} p^K(x)' (\widehat{\alpha}_{g_i} - \alpha_{g_i}) + A_{inT} [p^K(x)' \alpha_{g_i} - g_i(x)] \\ &= A_{inT} p^K(x)' (p_i' m_h p_i)^{-} p_i' m_h \varepsilon_i + A_{inT} p^K(x)' (p_i' m_h p_i)^{-} p_i' m_h f_2 \gamma_{2i} \\ &\quad + A_{inT} p^K(x)' (p_i' m_h p_i)^{-} p_i' m_h (\mathbf{g}_i - p_i \alpha_{g_i}) + A_{inT} [p^K(x)' \alpha_{g_i} - g_i(x)] \\ &\equiv E_{inT,1} + E_{inT,2} + E_{inT,3} + E_{inT,4}. \end{aligned} \quad (\text{A.9})$$

By Lemma A.7,  $E_{inT,1} \xrightarrow{d} N(0, 1)$ . We are left to show that  $E_{inT,s} = o_p(1)$  for  $s = 2, 3, 4$ .

We first show  $E_{inT,2} \xrightarrow{p} 0$  as  $(n, T) \rightarrow \infty$ . By the Cauchy-Schwarz inequality, Assumptions 1(vii), 3(iv) and 4(ii), Lemmas A.1(v) and A.5(iv)-(v),

$$\begin{aligned}
|E_{inT,2}| &= |A_{inT} p^K(x)' (p'_i m_h p_i)^{-} p'_i m_h f_2 \gamma_{2i}| \\
&\leq \left[ A_{inT}^2 p^K(x)' (p'_i m_h p_i)^{-} p^K(x) \right]^{1/2} \left[ \gamma'_{2i} f'_2 m_h p_i (p'_i m_h p_i)^{-} p'_i m_h f_2 \gamma_{2i} \right]^{1/2} \\
&\leq \left[ A_{inT}^2 p^K(x)' (p'_i m_h p_i)^{-} p'_i m_h \text{Var}(\varepsilon_i) m_h p_i (p'_i m_h p_i)^{-} p^K(x) / \lambda_{\min}(\text{Var}(\varepsilon_i)) \right]^{1/2} \\
&\quad \times \left[ \gamma'_{2i} f'_2 m_h p_i (p'_i m_h p_i)^{-} p'_i m_h f_2 \gamma_{2i} \right]^{1/2} \\
&\leq [1/\lambda_{\min}(\text{Var}(\varepsilon_i))]^{1/2} [\lambda_{\min}(p'_i m_h p_i / T)]^{-1/2} \left\{ \|p'_i m_h f_2 \gamma_{2i}\| / \sqrt{T} \right\} \\
&= O(1) O_p(1) O_p(\sqrt{KT/n}) = o_p(1).
\end{aligned}$$

Similarly, by the Cauchy-Schwarz inequality, Assumptions 1(vii), 3(iv) and 4(ii), Lemmas A.1(v) and A.2,

$$\begin{aligned}
|E_{inT,3}| &= |A_{inT} p^K(x)' (p'_i m_h p_i)^{-} p'_i m_h (\mathbf{g}_i - p_i \alpha_{g_i})| \\
&\leq \left[ A_{inT}^2 p^K(x)' (p'_i m_h p_i)^{-} p^K(x) \right]^{1/2} [(\mathbf{g}_i - p_i \alpha_{g_i})' (\mathbf{g}_i - p_i \alpha_{g_i})]^{1/2} \\
&\leq [1/\lambda_{\min}(\text{Var}(\varepsilon_i))]^{1/2} \|\mathbf{g}_i - p_i \alpha_{g_i}\| \\
&\leq O(1) O_p(\sqrt{T} K^{-\lambda_i/d}) = o_p(1).
\end{aligned}$$

Note that  $|E_{inT,4}| = A_{inT} |p^K(x)' \alpha_{g_i} - g_i(x)| = [1/V_{inT}]^{1/2} |p^K(x)' \alpha_{g_i} - g_i(x)|$  for  $(n, T)$  large. In addition, by (A.8), Assumptions 3(iii) and 4(ii),

$$\begin{aligned}
|E_{inT,4}| &\leq CT^{1/2} \left\{ |p^K(x)' \alpha_{g_i} - g_i(x)| \left[ 1 + \|x\|^2 \right]^{-\bar{\omega}_i/2} \right\} \left[ 1 + \|x\|^2 \right]^{\bar{\omega}_i/2} \\
&\leq C \left( \|g_i(\cdot) - \Pi_{\infty K} p^K(\cdot)\|_{\infty, \bar{\omega}_i} \right)^2 \left[ 1 + \|x\|^2 \right]^{\bar{\omega}_i/2} \\
&= O(\sqrt{T} K^{-\lambda_i/d}) = o(1).
\end{aligned}$$

This completes the proof. ■

### Proof of Theorem 4.3

**(i) Proof of Theorem 4.3(i):**  $\|\widehat{S}_{inT} - S_{inT}\| = o_p(1)$ . Recall  $\widehat{p}_i = m_h p_i$  and  $\widehat{p}'_{it}$  denotes the  $t$ th row of  $\widehat{p}_i$ . Let  $\eta_{it} \equiv \widehat{p}_{it} \varepsilon_{it}$  and  $\widehat{\eta}_{it} \equiv \widehat{p}_{it} \widehat{\varepsilon}_{it}$ . Then

$$\begin{aligned}
S_{inT} &= \widehat{p}'_i \text{Var}(\varepsilon_i) \widehat{p}_i = \sum_{s=1}^T \sum_{t=1}^T \widehat{p}'_{is} E(\varepsilon_{is} \varepsilon_{it}) \widehat{p}_{it}, \text{ and} \\
\widehat{S}_{inT} &= T^{-1} \sum_{t=1}^T \widehat{\eta}_{it} \widehat{\eta}'_{it} + T^{-1} \sum_{j=1}^{l_T} w_{Tj} \sum_{t=j+1}^T (\widehat{\eta}_{it} \widehat{\eta}'_{i,t-j} + \widehat{\eta}_{i,t-j} \widehat{\eta}'_{it}).
\end{aligned}$$

Let  $\bar{S}_{inT} = T^{-1} \sum_{t=1}^T \eta_{it} \eta'_{it} + T^{-1} \sum_{j=1}^{l_T} w_{Tj} \sum_{t=j+1}^T (\eta_{it} \eta'_{i,t-j} + \eta_{i,t-j} \eta'_{it})$ . It follows by the triangle inequality and the form of  $\bar{S}_{inT}$  that

$$\begin{aligned}
& \left\| \hat{S}_{inT} - S_{inT} \right\| \\
\leq & \left\| \hat{S}_{inT} - \bar{S}_{inT} \right\| + \left\| \bar{S}_{inT} - S_{inT} \right\| \\
\leq & \left\| \hat{S}_{inT} - \bar{S}_{inT} \right\| + \left\| T^{-1} \sum_{t=1}^T [\eta_{it} \eta'_{it} - E_{\mathcal{D}}(\eta_{it} \eta'_{it})] \right\| \\
& + \left\| T^{-1} \sum_{j=1}^{l_T} w_{Tj} \sum_{t=j+1}^T [(\eta_{it} \eta'_{i,t-j} + \eta_{i,t-j} \eta'_{it}) - E_{\mathcal{D}}(\eta_{it} \eta'_{i,t-j} + \eta_{i,t-j} \eta'_{it})] \right\| \\
& + \left\| T^{-1} \sum_{t=1}^T E_{\mathcal{D}}(\eta_{it} \eta'_{it}) + T^{-1} \sum_{j=1}^{l_T} w_{Tj} \sum_{t=j+1}^T E_{\mathcal{D}}(\eta_{it} \eta'_{i,t-j} + \eta_{i,t-j} \eta'_{it}) - S_{inT} \right\| \\
\equiv & \xi_{iT1} + \xi_{iT2} + \xi_{iT3} + \xi_{iT4}.
\end{aligned}$$

We will show  $\xi_{iT_s} = o_p(1)$ ,  $s = 1, 2, 3, 4$ , in the following four steps.

**Step 1:**  $\xi_{iT1} = \|\hat{S}_{inT} - \bar{S}_{inT}\| = o_p(1)$ . By the triangle inequality and the forms of  $\hat{S}_{inT}$  and  $\bar{S}_{inT}$ ,

$$\begin{aligned}
\xi_{iT1} & \leq \left\| T^{-1} \sum_{t=1}^T (\hat{\eta}_{it} \hat{\eta}'_{it} - \eta_{it} \eta'_{it}) \right\| + 2 \left\| T^{-1} \sum_{j=1}^{l_T} w_{Tj} \sum_{t=j+1}^T (\hat{\eta}_{it} \hat{\eta}'_{i,t-j} - \eta_{it} \eta'_{i,t-j}) \right\| \\
& \equiv \xi_{iT1,a} + 2\xi_{iT1,b}.
\end{aligned} \tag{A.10}$$

By the triangle inequality,

$$\begin{aligned}
\xi_{iT1,a} & \leq \left\| T^{-1} \sum_{t=1}^T (\hat{\eta}_{it} - \eta_{it}) (\hat{\eta}_{it} - \eta_{it})' \right\| + 2 \left\| T^{-1} \sum_{t=1}^T (\hat{\eta}_{it} - \eta_{it}) \eta'_{it} \right\| \\
& \equiv \xi_{iT1,a1} + 2\xi_{iT1,a2}.
\end{aligned} \tag{A.11}$$

Observe that  $\hat{\eta}_{it} - \eta_{it} = \hat{p}_{it} (\hat{e}_{it} - \varepsilon_{it}) = p'_{it} m_{h,t} (\hat{e}_{it} - \varepsilon_{it})$ , where  $m_{h,t}$  denotes the  $t$ -th column of  $m_h$ . Let  $m_{h,ts}$  denote the  $(t, s)$ th element of  $m_h$ . Then

$$\begin{aligned}
\sum_{s=1}^T \|p_{is} m_{h,ts}\| & = \sum_{s=1}^T \left\| p_{is} \left( \mathbf{1}(s=t) - h'_t (h'h)^{-1} h_s \right) \right\| \\
& \leq \|p_{it}\| + T^{-1} \sum_{s=1}^T \left\| h'_t (h'h/T)^{-1} h_s p'_{is} \right\| \leq \|p_{it}\| + \alpha_{2T} \|h_t\|, \tag{A.12}
\end{aligned}$$

where  $\mathbf{1}(\cdot)$  is the usual indicator function and

$$\alpha_{2T} \equiv \left\| (h'h/T)^{-1} \right\| \left\| T^{-1} \sum_{s=1}^T \|p_{is} h'_s\| \right\| = O_p(\sqrt{K}). \tag{A.13}$$

It follows from the symmetry of  $m_h$ , the triangle inequality, and (A.12) that

$$\|p'_i m_{h,\cdot t}\| = \left\| \sum_{s=1}^T p_{is} m_{h,ts} \right\| \leq \sum_{s=1}^T \|p_{is} m_{h,ts}\| = \|p_{it}\| + \alpha_{2T} \|h_t\|. \quad (\text{A.14})$$

By the triangle and Cauchy-Schwarz inequalities, (A.12), and Lemmas A.8 (ii)-(iii),

$$\begin{aligned} \xi_{iT1,a1} &= \left\| T^{-1} \sum_{t=1}^T p'_i m_{h,\cdot t} (\widehat{e}_{it} - \varepsilon_{it})^2 m'_{h,\cdot t} p_i \right\| \leq T^{-1} \sum_{t=1}^T \|p'_i m_{h,\cdot t}\|^2 (\widehat{e}_{it} - \varepsilon_{it})^2 \\ &\leq T^{-1} \sum_{t=1}^T \left[ 2 \|p_{it}\|^2 + 2\alpha_{2T}^2 \|h_t\|^2 \right] (\widehat{e}_{it} - \varepsilon_{it})^2 \\ &= 2T^{-1} \|\mathbf{p}_i (\widehat{e}_i - \varepsilon_i)\|^2 + 2\alpha_{2T}^2 T^{-1} \|\mathbf{h} (\widehat{e}_i - \varepsilon_i)\|^2 \\ &= O_p \left( K^2 (K/T + K^{-2\lambda_i/d} + K/n) \right) + O_p(K) O_p \left( K (K/T + K^{-2\lambda_i/d} + K/n) \right) \\ &= O_p \left( K^3/T + K^{-2\lambda_i/d+2} + K^3/n \right). \end{aligned} \quad (\text{A.15})$$

Similarly, by the triangle inequality, (A.14), the Cauchy-Schwarz inequality, and Lemma A.8,

$$\begin{aligned} &\xi_{iT1,a2} \\ &= \left\| T^{-1} \sum_{t=1}^T p'_i m_{h,\cdot t} (\widehat{e}_{it} - \varepsilon_{it}) \varepsilon_{it} m'_{h,\cdot t} p_i \right\| \leq T^{-1} \sum_{t=1}^T \|p'_i m_{h,\cdot t}\|^2 |(\widehat{e}_{it} - \varepsilon_{it}) \varepsilon_{it}| \\ &\leq T^{-1} \sum_{t=1}^T \left[ 2 \|p_{it}\|^2 + 2\alpha_{2T}^2 \|h_t\|^2 \right] |(\widehat{e}_{it} - \varepsilon_{it}) \varepsilon_{it}| \\ &= 2T^{-1} \sum_{t=1}^T \|p_{it}\|^2 |(\widehat{e}_{it} - \varepsilon_{it}) \varepsilon_{it}| + 2\alpha_{2T}^2 T^{-1} \sum_{t=1}^T \|h_t\|^2 |(\widehat{e}_{it} - \varepsilon_{it}) \varepsilon_{it}| \\ &\leq 2 \left\{ T^{-1} \sum_{t=1}^T (\widehat{e}_{it} - \varepsilon_{it})^2 \right\}^{1/2} \left\{ T^{-1} \sum_{t=1}^T \|p_{it}\|^4 \varepsilon_{it}^2 \right\}^{1/2} \\ &\quad + 2\alpha_{2T}^2 \left\{ T^{-1} \sum_{t=1}^T (\widehat{e}_{it} - \varepsilon_{it})^2 \right\}^{1/2} \left\{ T^{-1} \sum_{t=1}^T \|h_t\|^4 \varepsilon_{it}^2 \right\}^{1/2} \\ &= O_p(\sqrt{K/T} + K^{-\lambda_i/d} + \sqrt{K/n}) O_p(\sqrt{K}) + O_p(K) O_p(\sqrt{K/T} + K^{-\lambda_i/d} + \sqrt{K/n}) O_p(1) \\ &= O_p \left( K(\sqrt{K/T} + K^{-\lambda_i/d} + \sqrt{K/n}) \right). \end{aligned} \quad (\text{A.16})$$

Hence by (A.11), (A.15) and (A.16),

$$\xi_{iT1,a} = O_p \left( K(\sqrt{K/T} + K^{-\lambda_i/d} + \sqrt{K/n}) \right) = o_p(1). \quad (\text{A.17})$$

By the uniform boundedness of  $w_{Tj}$ , arguments similar to those in the analysis of  $\xi_{iT1,a}$  and the Cauchy-Schwarz inequality, we can show

$$\xi_{iT1,b} = O_p(l_T K) O_p \left( \sqrt{K/T} + K^{-\lambda_i/d} + \sqrt{K/n} \right) = o_p(1). \quad (\text{A.18})$$

Combining (A.10), (A.17) and (A.18) yields  $\xi_{iT1} = o_p(1)$ .

**Step 2:**  $\xi_{iT2} = \|T^{-1} \sum_{t=1}^T [\eta_{it}\eta'_{it} - E_{\mathcal{D}}(\eta_{it}\eta'_{it})]\| = o_p(1)$ . Let  $p_{i,\cdot j}$  denote the  $j$ th column of the  $T \times K$  matrix  $p_i$ . Then

$$E[\xi_{iT2}]^2 = E \left\| T^{-1} \sum_{t=1}^T p'_{i,\cdot t} m_{h,\cdot t} m'_{h,\cdot t} p_i [\varepsilon_{it}^2 - E(\varepsilon_{it}^2)] \right\|^2 = \sum_{j=1}^K \sum_{k=1}^K E[\varpi_{jk}]^2,$$

where  $\varpi_{jk} \equiv T^{-1} \sum_{t=1}^T p'_{i,\cdot j} m_{h,\cdot t} m'_{h,\cdot t} p_{i,\cdot k} [\varepsilon_{it}^2 - E(\varepsilon_{it}^2)]$ . Observe that  $E(\varpi_{jk}) = 0$ . Using the Davydov inequality we can show that  $E(\varpi_{jk}^2) = O(T^{-1})$ . Hence  $E[\xi_{iT2}]^2 = O(K^2/T)$ . It follows that  $\xi_{iT2} = O_p(K/\sqrt{T}) = o_p(1)$ .

**Step 3:**  $\xi_{iT3} = T^{-1} \left\| \sum_{j=1}^{l_T} w_{Tj} \sum_{t=j+1}^T [(\eta_{it}\eta'_{i,t-j} + \eta_{i,t-j}\eta'_{it}) - E_{\mathcal{D}}(\eta_{it}\eta'_{i,t-j} + \eta_{i,t-j}\eta'_{it})] \right\| = o_p(1)$ . Let  $\chi_{i,tj} = (\eta_{it}\eta'_{i,t-j} + \eta_{i,t-j}\eta'_{it}) - E_{\mathcal{D}}(\eta_{it}\eta'_{i,t-j} + \eta_{i,t-j}\eta'_{it})$ . Then  $\xi_{iT3} = T^{-1} \left\| \sum_{j=1}^{l_T} w_{Tj} \sum_{t=j+1}^T \chi_{i,tj} \right\|$ . By the arguments used in the proof of Lemma A.1(i), we can show  $E \left\| T^{-1} \sum_{t=j+1}^T \chi_{i,tj} \right\|^2 = O_p(K^2/T)$ . Because  $\max_j |w_{Tj}| \leq c_w$ , for any  $\varepsilon > 0$ ,

$$\begin{aligned} & P(\|\xi_{iT3}\| > \varepsilon) \\ &= P\left(T^{-1} \left\| \sum_{j=1}^{l_T} w_{Tj} \sum_{t=j+1}^T \chi_{i,tj} \right\| > \varepsilon\right) \leq P\left(T^{-1} \sum_{j=1}^{l_T} |w_{Tj}| \left\| \sum_{t=j+1}^T \chi_{i,tj} \right\| > \varepsilon\right) \\ &\leq \sum_{j=1}^{l_T} P\left(T^{-1} \left\| \sum_{t=j+1}^T \chi_{i,tj} \right\| > \varepsilon / (c_w l_T)\right) \leq \frac{(c_w l_T)^2}{\varepsilon^2} \sum_{j=1}^{l_T} E \left\| T^{-1} \sum_{t=j+1}^T \chi_{i,tj} \right\|^2 \\ &= O(l_T^3 K^2/T) = o(1). \end{aligned}$$

That is,  $\xi_{iT3} = o_p(1)$ .

**Step 4:**  $\xi_{iT4} = \|T^{-1} \sum_{t=1}^T E_{\mathcal{D}}(\eta_{it}\eta'_{it}) + T^{-1} \sum_{j=1}^{l_T} w_{Tj} \sum_{t=j+1}^T E_{\mathcal{D}}(\eta_{it}\eta'_{i,t-j} + \eta_{i,t-j}\eta'_{it}) - S_{inT}\| = o_p(1)$ . By the triangle inequality

$$\begin{aligned} \xi_{iT4} &\leq \left\| T^{-1} \sum_{j=1}^{l_T} |w_{Tj} - 1| \sum_{t=j+1}^T E_{\mathcal{D}}(\eta_{it}\eta'_{i,t-j} + \eta_{i,t-j}\eta'_{it}) \right\| \\ &\quad + \left\| T^{-1} \sum_{j=l_T+1}^{T-1} \sum_{t=j+1}^T E_{\mathcal{D}}(\eta_{it}\eta'_{i,t-j} + \eta_{i,t-j}\eta'_{it}) \right\| \equiv \xi_{iT4,a} + \xi_{iT4,b}. \end{aligned}$$

First, by Assumption 1, the additional assumption in Theorem 4.3, the Davydov inequality, and (A.14)

$$\begin{aligned} E|\xi_{iT4,b}| &= 2E \left\| T^{-1} \sum_{j=l_T+1}^{T-1} \sum_{t=j+1}^T p'_{i,\cdot t} m_{h,\cdot t} m'_{h,\cdot t-j} p_i E[\varepsilon_{it}\varepsilon_{i,t-j}] \right\| \\ &\leq 2T^{-1} E \|p'_{i,\cdot 1} m_{h,\cdot 1}\|^2 \sum_{j=l_T+1}^{T-1} \sum_{t=j+1}^T |E[\varepsilon_{it}\varepsilon_{i,t-j}]| \leq 2CK \sum_{j=l_T+1}^{T-1} \alpha_i^{\eta/(2+\eta)}(j) \left\{ E|\varepsilon_{it}|^{2+\eta} \right\}^{2/(2+\eta)} \\ &\leq 2C \{Kl_T^{-\alpha_0}\} \sum_{j=l_T+1}^{\infty} j^{\alpha_0} \alpha_i^{\eta/(2+\eta)}(j) \left\{ E|\varepsilon_{it}|^{2+\eta} \right\}^{2/(2+\eta)} \rightarrow 0. \end{aligned}$$

where the last line follows because  $Kl_T^{-\alpha_0} = o(1)$  and  $\sum_{j=l_T+1}^{\infty} j^{\alpha_0} \alpha_i^{\eta/(2+\eta)}(j) < \infty$ . Now, by the triangle inequality,  $\xi_{iT4,a} \leq T^{-1} \sum_{j=1}^{l_T} |w_{Tj} - 1| \|\sum_{t=j+1}^T E_{\mathcal{D}} (\eta_{it} \eta'_{i,t-j} + \eta_{i,t-j} \eta'_{it})\|$ . By the Davydov inequality,  $E \|\sum_{t=j+1}^T E_{\mathcal{D}} (\eta_{it} \eta'_{i,t-j} + \eta_{i,t-j} \eta'_{it})\| \leq CK \alpha^{\eta/(2+\eta)}(j)$ . By the additional assumption in Theorem 4.3,

$$K \sum_{j=1}^{l_T} \alpha_i^{\eta/(2+\eta)}(j) \leq \{Kl_T^{-\alpha_0}\} \sum_{j=1}^{l_T} l_T^{\alpha_0} \alpha_i^{\eta/(2+\eta)}(j) < \infty.$$

Since  $\lim_{T \rightarrow \infty} w_{Tj} = 1$  for each  $j$ , it follows from the dominated convergence theorem that

$$E |\xi_{iT4,a}| \leq \sum_{j=1}^{l_T} |w_{Tj} - 1| E \left\| T^{-1} \sum_{t=j+1}^T E_{\mathcal{D}} (\eta_{it} \eta'_{i,t-j} + \eta_{i,t-j} \eta'_{it}) \right\| \rightarrow 0.$$

Consequently,  $\xi_{iT4} = o_p(1)$ . Combining Steps 1-4 yields  $\|\widehat{S}_{inT} - S_{inT}\| = o_p(1)$ .

**(ii) Proof of Theorem 4.3(ii):**  $\widehat{V}_{inT} V_{inT}^{-1} \xrightarrow{p} 1$ ,  $\widehat{A}_{inT} A_{inT}^{-1} \xrightarrow{p} 1$ . By Assumption 3(iv), Lemma A.1(v), and the result in Theorem 4.3(i),

$$\begin{aligned} \left| T \left( \widehat{V}_{inT} - V_{inT} \right) \right| &= \left| p^K(x)' (p'_i m_h p_i / T)^{-} \left( \widehat{S}_{inT} - S_{inT} \right) (p'_i m_h p_i / T)^{-} p^K(x) \right| \\ &\leq \|p^K(x)\|^2 [\lambda_{\min}(p'_i m_h p_i / T)]^{-2} \|\widehat{S}_{inT} - S_{inT}\| = \|p^K(x)\|^2 o_p(1). \end{aligned}$$

Similarly, by Assumptions 1(vii) and 3(iv) and Lemma A.1(iv),

$$\begin{aligned} T^{-1} A_{inT}^2 &= T^{-1} \left\{ p^K(x)' (p'_i m_h p_i)^{-} p'_i m_h \Psi_{i,T} m_h p_i (p'_i m_h p_i)^{-} p^K(x) \right\}^{-1} \\ &\leq [\lambda_{\max}(\Psi_{i,T})]^{-1} \left\{ p^K(x)' (p'_i m_h p_i / T)^{-} p^K(x) \right\}^{-1} \\ &\leq [\lambda_{\max}(\Psi_{i,T})]^{-1} \lambda_{\max}(p'_i m_h p_i / T) \|p^K(x)\|^{-2} = \|p^K(x)\|^{-2} O_p(1). \end{aligned}$$

It follows that

$$\begin{aligned} \left| \widehat{V}_{inT} V_{inT}^{-1} - 1 \right| &= \left| A_{inT}^2 \widehat{V}_{inT} - A_{inT}^2 V_{inT} \right| \\ &= (T^{-1} A_{inT}^2) \left| T \left( \widehat{V}_{inT} - V_{inT} \right) \right| \leq \|p^K(x)\|^{-2} O_p(1) \|p^K(x)\|^2 o_p(1) = o_p(1). \end{aligned}$$

The second conclusion follows from this result and the Slutsky theorem. Alternatively, we have  $|\widehat{A}_{inT}^2 A_{inT}^{-2} - 1| = |\widehat{V}_{inT}^{-1} V_{inT} - 1| = |\widehat{V}_{inT}^{-1} V_{inT}| |\widehat{V}_{inT} V_{inT}^{-1} - 1| \xrightarrow{p} 1 \times 0 = 0$ .

**(iii) Proof of Theorem 4.3(iii):**  $\widehat{A}_{inT} [\widehat{g}_i(x) - g_i(x)] \xrightarrow{d} N(0, 1)$ . Write  $\widehat{A}_{inT} (\widehat{g}_i(x) - g_i(x)) = (\widehat{A}_{inT} A_{inT}^{-1}) A_{inT} (\widehat{g}_i(x) - g_i(x))$ . The result follows from the results in Theorem 4.2 and 4.3(ii), and the Slutsky theorem. ■

## B Proof of results in section 5

In this appendix, we first state some lemmas that are used in the proof of the main results in Section 5 and then prove Theorems 5.1-5.2. The proof of all lemmas can be found at [http://www.mysmu.edu/faculty/ljsu/Publications/Sieve10\\_supp.pdf](http://www.mysmu.edu/faculty/ljsu/Publications/Sieve10_supp.pdf).

**Lemma B.1** *Suppose the conditions in Lemma A.1 and Assumption 5(i) hold, then (i)  $\|(nT)^{-1}P'M_hP - \bar{Q}_n\| = o_p(K^{-1/2})$ ; (ii)  $\lambda_{\max}(P'M_hP/(nT)) = \lambda_{\max}(\bar{Q}_n) + o_p(K^{-1/2}) = O_p(1)$ ; (iii)  $\lambda_{\min}(P'M_hP/(nT)) \geq \lambda_{\min}(\bar{Q}_n)/2$  w.p.a.1.*

**Lemma B.2** *Suppose Assumptions 3(i)-(iii) hold with  $g_i(\cdot)$  replaced by  $g(\cdot)$ , then  $(nT)^{-1}E\|\mathbf{g} - P\alpha_g\|^2 = O(K^{-2\lambda/d})$  where  $\mathbf{g} = (\mathbf{g}'_1, \dots, \mathbf{g}'_n)'$ ,  $\mathbf{g}_i = (g(x_{i1}), \dots, g(x_{iT}))'$ , and  $\lambda = \min_{1 \leq i \leq n} \lambda_i$ .*

**Lemma B.3** *Suppose Assumptions 1(iii), (iv) and (vii), 2(i)-(ii), and 3(iv) hold, then  $E\|(NT)^{-1}P'M_h\varepsilon\|^2 = O(K/(nT))$ .*

**Lemma B.4** *Let  $M_b = I_n \otimes m_b$ . Suppose Assumptions 1, 2, and 3(iv) hold and  $\lambda_{\max}(E(\gamma_2\gamma_2')) = O(r_n)$ , then  $\|(P'M_hF_2 - P'M_bF_2)\gamma_2\|/\sqrt{nT} = O_p(\sqrt{r_nK/n} + \sqrt{KT/n})$ .*

**Lemma B.5** *Suppose the conditions in Theorem 5.2 hold, then  $T_{nT} \equiv A_{nT}p^K(x)'(P'M_hP)^{-1}P'M_h\varepsilon \xrightarrow{d} N(0, 1)$ .*

### Proof of Theorem 5.1

The proof is analogous to that of Theorem 4.1(i) and we only sketch it. By (4.2),  $\hat{\alpha}_g - \alpha_g = (P'M_hP)^{-1}P'M_h\varepsilon + (P'M_hP)^{-1}P'M_hF_2\gamma_2 + (P'M_hP)^{-1}P'M_h(\mathbf{g} - P\alpha_g) \equiv D_1 + D_2 + D_3$ . First, by Assumption 5(i), Lemmas B.1(iii) and B.3,

$$\begin{aligned} \|D_1\|^2 &= \varepsilon'M_hP(P'M_hP)^{-1}(P'M_hP)^{-1}P'M_h\varepsilon \\ &\leq [\lambda_{\min}(P'M_hP/(nT))]^{-2} \left\{ \|P'M_h\varepsilon\|^2 / (nT)^2 \right\} = O_p(K/(nT)). \end{aligned}$$

Similarly, by Assumption 5(i) and Lemmas B.1(iii), B.4 and A.5(iv),

$$\begin{aligned} \|D_2\|^2 &= \gamma_2'F_2'M_hP(P'M_hP/(nT))^{-1}(P'M_hP/(nT))^{-1}P'M_hF_2\gamma_2/(nT)^2 \\ &\leq [\lambda_{\min}(P'M_hP/(nT))]^{-2} \gamma_2'F_2'M_hPP'M_hF_2\gamma_2/(nT)^2 \\ &\leq [\lambda_{\min}(P'M_hP/(nT))]^{-2} \left\{ \|P'M_hF_2\gamma_2\|^2 / (n^2T^2) \right\} \\ &= O_p(1)O_p(1)O_p(r_nK/(n^2T) + K/n^2) \\ &= O_p(K/(nT) + K/n^2) \text{ as } r_n \ll n. \end{aligned}$$

Let  $W \equiv M_h P (P' M_h P)^- P' M_h$ . Then  $W$  is symmetric and idempotent and thus  $\lambda_{\max}(W) = 1$ . It follows from Assumption 5(i) and Lemmas B.1(iii) and B.2 that

$$\begin{aligned} \|D_3\|^2 &= (g - P\alpha_g)' M_h P (P' M_h P / (nT))^- (P' M_h P)^- P' M_h (g - P\alpha_g) / (nT) \\ &\leq [\lambda_{\min}(P' M_h P / (nT))]^{-1} (g - P\alpha_g)' M_h P (P' M_h P)^- P' M_h (g - P\alpha_g) / (nT) \\ &\leq [\lambda_{\min}(P' M_h P / (nT))]^{-1} \lambda_{\max}(W) \left\{ \|g - P\alpha_g\|^2 / (nT) \right\} \\ &= O_p(1) O_p(K^{-2\lambda/d}) = O_p(K^{-2\lambda/d}). \end{aligned}$$

By the triangle inequality,

$$\|\hat{\alpha}_g - \alpha_g\| \leq \|D_1\| + \|D_2\| + \|D_3\| = O_p(\sqrt{K/(nT)} + \sqrt{K/n^2} + K^{-\lambda/d}). \quad (\text{B.1})$$

By the Chebyshev inequality, it is straightforward to show that  $\|(nT)^{-1} P' P - n^{-1} \sum_{i=1}^n Q_{ipp}\| = o_p(1)$ . Noting that  $\lambda_{\max}(A + B) \leq \lambda_{\max}(A) + \lambda_{\max}(B)$  for any two real symmetric square matrices  $A$  and  $B$ , we have  $\lambda_{\max}\left((nT)^{-1} P' P\right) \xrightarrow{p} \lambda_{\max}\left(n^{-1} \sum_{i=1}^n Q_{ipp}\right) \leq n^{-1} \sum_{i=1}^n \lambda_{\max}(Q_{ipp}) = O(1)$ . Then by (B.1), Lemma B.2, Assumption 3(iv) with  $g_i$  replaced by  $g$ , and the Markov inequality, we have

$$\begin{aligned} &\frac{1}{nT} \sum_{i=1}^n \sum_{t=1}^T [\hat{g}(x_{it}) - g(x_{it})]^2 \\ &= \frac{1}{nT} \sum_{i=1}^n \sum_{t=1}^T \left\{ p^K(x_{it})' (\hat{\alpha}_g - \alpha_g) + [p^K(x_{it})' \alpha_g - g(x_{it})] \right\}^2 \\ &\leq 2(\hat{\alpha}_g - \alpha_g)' \left( (nT)^{-1} P' P \right) (\hat{\alpha}_g - \alpha_g) + 2(nT)^{-1} \|g - P\alpha_g\|^2 \\ &\leq 2\lambda_{\max}\left( (nT)^{-1} P' P \right) \|\hat{\alpha}_g - \alpha_g\|^2 + 2(nT)^{-1} \|g - P\alpha_g\|^2 \\ &= O_p(1) O_p(K^{-2\lambda/d} + K/(nT) + K/n^2) + O_p(K^{-2\lambda/d}) = O_p(K^{-2\lambda/d} + K/(nT) + K/n^2). \blacksquare \end{aligned}$$

### Proof of Theorem 5.2

Recall  $V_{nT} = p^K(x)' (P' M_h P)^- P' M_h \Psi_{nT} M_h P (P' M_h P)^- p^K(x)$  and  $A_{nT} = V_{nT}^{-1/2}$ , where  $\Psi_{nT} = E(\varepsilon\varepsilon')$ . Write

$$\begin{aligned} &A_{nT} [\hat{g}(x) - g(x)] \\ &= A_{nT} p^K(x)' (\hat{\alpha}_g - \alpha_g) + A_{nT} [p^K(x)' \alpha_g - g(x)] \\ &= A_{nT} p^K(x)' (P' M_h P)^- P' M_h \varepsilon + A_{nT} p^K(x)' (P' M_h P)^- P' M_h F_2 \gamma_2 \\ &\quad + A_{nT} p^K(x)' (P' M_h P)^- P' M_h (g - P\alpha_g) + A_{nT} [p^K(x)' \alpha_g - g(x)] \\ &\equiv E_{1nT} + E_{2nT} + E_{3nT} + E_{4nT}. \end{aligned}$$

By Lemma B.5,  $E_{1nT} \xrightarrow{d} N(0, 1)$ . It remains to show that  $E_{snT} = o_p(1)$  for  $s = 2, 3, 4$ .

We first show  $E_{2nT} \xrightarrow{p} 0$  as  $(n, T) \rightarrow \infty$ . By the Cauchy-Schwarz inequality, Assumptions 1(v), (vii), and 5(i), and Lemmas B.1(iii), B.4 and A.5(iv)

$$\begin{aligned}
|E_{2nT}| &= |A_{nT} p^K(x)' (P' M_h P)^- P' M_h F_2 \gamma_2| \\
&\leq \left[ A_{nT}^2 p^K(x)' (P' M_h P)^- p^K(x) \right]^{1/2} \left[ \gamma_2' F_2' M_h P (P' M_h P)^- P' M_h F_2 \gamma_2 \right]^{1/2} \\
&\leq [1/\lambda_{\min}(\Psi_{nT})]^{1/2} [\lambda_{\min}(P' M_h P/nT)]^{-1/2} \left\{ \|P' M_h F_2 \gamma_2\| / \sqrt{nT} \right\} \\
&= O_p(1) O_p(1) O_p(1) O_p(\sqrt{r_n K/n} + \sqrt{KT/n}) = o_p(1).
\end{aligned}$$

Next, by the Cauchy-Schwarz inequality, Assumptions 1(v) and (vii), and Lemma B.2,

$$\begin{aligned}
|E_{3nT}| &= |A_{nT} p^K(x)' (P' M_h P)^- P' M_h (\mathbf{g} - P\alpha_g)| \\
&\leq \left[ A_{nT}^2 p^K(x)' (P' M_h P)^- P' M_h M_h P (P' M_h P)^- p^K(x) \right]^{1/2} [(\mathbf{g} - P\alpha_g)' (\mathbf{g} - P\alpha_g)]^{1/2} \\
&\leq \left[ A_{nT}^2 p^K(x)' (P' M_h P)^- P' M_h \Psi_{nT} M_h P (P' M_h P)^- p^K(x) / \lambda_{\min}(\Psi_{nT}) \right]^{1/2} \|\mathbf{g} - P\alpha_g\| \\
&\leq [1/\lambda_{\min}(\Psi_{nT})]^{1/2} \|\mathbf{g} - P\alpha_g\| = O_p(1) O_p(\sqrt{nT} K^{-\lambda/d}) = o_p(1).
\end{aligned}$$

Now, write  $|E_{4nT}| = A_{nT} |p^K(x)' \alpha_g - g(x)| = [1/V_{nT}]^{1/2} |p^K(x)' \alpha_g - g(x)|$  for  $(n, T)$  large. By Assumptions 1(v), (vii) and 5(i), Lemma B.1(ii), and the fact that  $\|p^K(x)\| > c > 0$  for some constant  $c$ ,

$$\begin{aligned}
nTV_{nT} &= nTp^K(x)' (P' M_h P)^- P' M_h \Psi_{nT} M_h P (P' M_h P)^- p^K(x) \\
&\geq \lambda_{\min}(\Psi_{nT}) [p^K(x)' (P' M_h P/(nT))^- p^K(x)] \\
&\geq \lambda_{\min}(\Psi_{nT}) [\lambda_{\max}(P' M_h P/(nT))]^{-1} \|p^K(x)\|^2 \geq C > 0 \text{ w.p.a.1.}
\end{aligned}$$

Hence,  $|E_{4nT}| \leq C(nT)^{1/2} |p^K(x)' \alpha_g - g(x)| = O((nT)^{1/2} K^{-\lambda/d}) = o(1)$ . This completes the proof. ■

## References

- Ai, C., Chen, X., 2003. Efficient estimation of models with conditional moment restrictions containing unknown functions. *Econometrica* 71, 1795-1843.
- Andrews, D. W. K., 1991a. Asymptotic normality of series estimators for nonparametric and semi-parametric regression models. *Econometrica* 59, 307-345.
- Andrews, D. W. K., 1991b. Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica* 59, 817-858.
- Bai, J., 2003. Inferential theory for factor models of large dimension. *Econometrica* 71, 135-171.
- Bai, J., 2009. Panel data models with interactive fixed effects. *Econometrica* 77, 1229-1279.
- Bai, J., Ng, S., 2002. Determining the number of factors in approximate factor models. *Econometrica* 70, 191-221.

- Baltagi, B. H., Li, D. 2002. Series estimation of partially linear panel data models with fixed effects. *Annals of Economics and Finance* 3, 103-116.
- Baltagi, B. H., Hidalgo, J., Li, Q., 1996. A nonparametric test for poolability using panel data. *Journal of Econometrics* 75, 345-367.
- Blundell, R., Chen, X., Kristensen, D., 2007. Semi-nonparametric IV estimation of shape-invariant Engel curves. *Econometrica* 75, 1613-1669.
- Chen, X., 2007. Large sample sieve estimation of semi-nonparametric models. In J. Heckman and E. Leamer (eds), *Handbook of Econometrics*, Vol. 6, pp. 5549-5632, North Holland, Amsterdam.
- Chen, X., Hong, H., Tamer, E., 2005. Measurement error models with auxiliary data. *Review of Economic Studies* 72, 343-366.
- Coakley, J., Fuertes, A., Smith, R., 2002. A principal components approach to cross-section dependence in panels. Working paper, Birkbeck College, Univ. of London.
- Fan, J., Zhang, C., Zhang, J., 2001. Generalized likelihood ratio test statistic and Wilks phenomenon. *Annals of Statistics* 29, 153-193.
- Fan, Y., Li, Q., 1996. Consistent model specification tests: omitted variables and semiparametric functional forms. *Econometrica* 64, 865-890.
- Greenaway-McGrevy, R., Han, C., Sul, D., 2008. Estimating and testing idiosyncratic equations using cross-section dependent panel data. Working paper, Dept. of Economics, Univ. of Auckland.
- Harding, M. C., 2007. Structural estimation of high-dimensional factor models: uncovering the effect of global factors on the US economy. Mimeo, Dept. of Economics, MIT.
- Henderson, D. J., Carroll, R. J., Li, Q., 2008. Nonparametric estimation and testing of fixed effects panel data models. *Journal of Econometrics* 144, 257-275.
- Huang, X., 2006. Nonparametric estimation in large panel with cross-section dependence, Department of Economics & Finance, Kennesaw State University.
- Jin, S., L. Su, 2010. A nonparametric poolability test for panel data models with cross section dependence. Working paper, School of Economics, Singapore Management University.
- Kapetanios, G., Pesaran, M. H., 2005. Alternative approaches to estimation and inference in large multifactor panels: small sample results with an application to modelling of asset returns. Working paper, Cambridge University.
- Li, Q., 2000. Efficient estimation of additive partially linear models. *International Economic Review* 41, 1073-1092.
- Li, Q., Hsiao, C., Zinn, J., 2003. Consistent specification tests for semiparametric/nonparametric models based on series estimation methods. *Journal of Econometrics* 112, 295-325.
- Moon, H. R., Weidner, M., 2008. Asymptotic analysis of the quasi-MLE of panel regression models with interactive fixed effects. Working paper, Dept. of Economics, Univ. of Southern California.

- Newey, W. K., 1994. Series estimation of regression functionals. *Econometric Theory* 10, 1-28.
- Newey, W. K., 1995. Convergence rates for series estimators. In: Maddala, G. S., Phillips, P. C. B., Srinivasan, T. N. (Eds.), *Advances in Econometrics and Quantitative Economics: Essays in Honor of C. R. Rao*. Blackwell, Cambridge, USA, pp. 254-275.
- Newey, W. K., 1997. Convergence rates and asymptotic normality for series estimators. *Journal of Econometrics* 79, 147-168.
- Newey, W. K., West, K. D., 1987. A simple positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica* 55, 703-708.
- Pesaran, M. H., 2004. General diagnostic tests for cross section dependence in panels. Working paper no. 1229. Cambridge University.
- Pesaran, M. H., 2006. Estimation and inference in large heterogenous panels with multifactor error. *Econometrica* 74, 967-1012.
- Pesaran, M. H., Tosetti, E., 2007. Large panels with common factors and spatial correlations. Working paper no. 2103, Cambridge University.
- Phillips, P. C. B., Sul, D., 2003. Dynamic panel estimation and homogeneity testing under cross sectional dependence. *The Econometrics Journal* 6, 217-259.
- Phillips, P. C. B., Sul, D., 2007. Bias in dynamic panel estimation with fixed effects, incidental trends and cross section dependence. *Journal of Econometrics* 137, 162-188.
- Stinchcombe, M. B., White, H., 1998. Consistent specification testing with nuisance parameters present only under the alternative. *Econometric Theory* 14, 295-324.
- Stock, J., Watson, M. W., 2002. Macroeconomic forecasting using diffusion indexes. *Journal of Business & Economic Statistics* 20, 147-162.
- Stone, C. J., 1982. Optimal global rates of convergence for nonparametric regression. *Annals of Statistics* 10, 1040-1053.
- Vilar-Fernández, J. M., González-Manteiga W., 2004. Nonparametric comparison of curves with dependent errors. *Statistics* 38, 81-99.
- Wang, G., Wei, Y., Qiao, S., 2004. *Generalized Inverse: Theory and Computations*. Science Press, New York.
- White, H., Domowitz, I., 1984. Nonlinear regression with dependent observations. *Econometrica* 52, 143-161.