Interactive Effects Panel Data Models with General Factors and Regressors*

Bin Peng† Liangju Su‡ Joakim Westerlund§

Monash University Tsinghua University Lund University

and

Deakin University

Yanrong Yang¶

Australian National University

February 12, 2023

*Previous versions of the paper were presented at seminars at Aarhus University, Lund University and Michigan State University. The authors would like to thank seminar participants, and in particular Richard Baillie, Nicholas Brown, David Edgerton, Yousef Kaddoura, Yana Petrova, Simon Reese, Tim Vogelsang, Jeffrey Wooldridge, and Morten Ørregaard Nielsen. Su thanks the National Natural Science Foundation of China for financial support under the grant number 72133002, Westerlund thanks the Knut and Alice Wallenberg Foundation for financial support through a Wallenberg Academy Fellowship, and Peng, Westerlund and Yang thank the Australian Research Council Discovery Grants Program for financial support under grant numbers DP210100476 and DP230102250.

†E-mail address: bin.peng@monash.edu.

‡E-mail address: sulj@sem.tsinghua.edu.cn.

§E-mail address: joakim.westerlund@nek.lu.se.

¶E-mail address: yanrong.yang@anu.edu.au.

Proposed running head: Panel Models with General Factors and Regressors

Corresponding author: Bin Peng

Address: Department of Econometrics and Business Statistics

Monash University

Caulfield East, VIC 3145

Australia

E-mail address: bin.peng@monash.edu

Abstract

This paper considers a model with general regressors and unobservable common factors. An estimator based on iterated principal component analysis is proposed, which is shown to be not only asymptotically normal, but also under certain conditions free of the otherwise so common asymptotic incidental parameters bias. Interestingly, the conditions required to achieve unbiasedness become weaker the stronger the trends in the factors, and if the trending is strong enough unbiasedness comes at no cost at all. The approach does not require any knowledge of how many factors there are, or whether they are deterministic or stochastic. The order of integration of the factors is also treated as unknown, as is the order of integration of the regressors, which means that there is no need to pre-test for unit roots, or to decide on which deterministic terms to include in the model.

1 Introduction

The use of panel data with interactive fixed effects in regression analyses has attracted considerable attention in the empirical literature in economics and elsewhere. One of the most common approaches to such models by far is the principal component (PC) approach of Bai (2009). In fact, the PC approach is so common that it has given rise to a separate strand of literature (see Moon and Weidner, 2015, and Ando and Bai, 2017, for overviews). The present paper aims to contribute to this strand, and it does so in at least three ways.

The first contribution of the paper is to consider a general data generating process (DGP) that includes most of the specifications considered previously in the literature as special cases. The only requirement is that suitably normalized sample second moment matrices of the factors and regressors have positive definite limits. This is noteworthy because the existing literature is almost exclusively based on the assumption that both the factors and regressors are stationary. The only exceptions known to us are Bai et al. (2009), and Dong et al. (2021), but they limit the non-stationarity to unit root processes only, which is also not realistic. Indeed, regressors and factors of different order of magnitude are likely to be the rule rather than the exception, especially in economic and financial data, due to differences in persistence over time.

The unrestricted DGP is important in itself but also because it can be accommodated without requiring any knowledge thereof. Hence, not only do we treat the factors and their number as unknown, but we also do not require any knowledge of the order of magnitude of both factors and regressors. An important implication of this is that there is no need to distinguish between deterministic and stochastic factors, or stationary and non-stationary factors. In the existing literature, deterministic factors are often treated as known, and are projected out prior to the application of PC (see, for example, Moon and Weidner, 2015). The problem here is that there is typically great uncertainty over which deterministic terms to include, which raises the issue of model misspecification. The fact that in the present paper deterministic terms are treated as additional factors means that the problem of deciding on which terms to include does not arise. Similarly, while the regressors can be tested for unit roots, and the estimation can be made conditional on the test outcome, this raises the issue of pre-testing bias. In the present paper we do not require any knowledge about the order of integration of the regressors, which means that there is no need for any pre-testing.

Equally as important as the general model formulation and its empirical appeal is the extension of the existing econometric theory, which has not yet ventured much outside the stationary or pure unit root environments. This is our second contribution. The main difficulty here is not the unrestricted specification of the factors and regressors per se, but rather that the order of magnitude of the factors may differ. In particular, the problem is that the nonlinearity of the PC estimator distorts the signal coming from the factors, just as it does in estimation of nonlinear regression models with mixtures of

integrated regressors (see, for example, Park and Phillips, 2000). This is true if both the number and order of the factors are known, and the problem does not become any simpler when these quantities are treated as unknown, as they are here. An additional problem that then arises is that existing studies on the selection of the number of factors all require that the data be stationary (see, for example, Bai and Ng, 2002, and Ahn and Horenstein, 2013), and it is not obvious how one should go about this when the order of magnitude of the factor is unknown.

Intuitively, the factors whose order is largest should dominate the PC estimator. This motivates the use of an iterative estimation procedure in which the factors and their number are estimated in order according to their magnitude with relatively larger factors being estimated first. We begin by prescribing a large number of factors, and estimate the resulting model by PC. The estimated factors only capture the most dominating factors whose order of magnitude is largest. In spite of this, we can show that the estimator is consistent, albeit at a relatively low rate of convergence. The rate is, however, high enough to ensure that the number of dominating factors can be consistently estimated using a version of the eigenvalue ratio approach of Lam and Yao (2012), and Ahn and Horenstein (2013). We then apply PC conditional on the first-step factor estimates, and estimate the second most dominating set of factors. This procedure continues until we cannot identify any more factors. Because of the iterative fashion in which the factors are estimated, we refer to the new estimation procedure as "iterative PC" (IPC), which is shown to be asymptotically (mixed) normal.

Our third contribution is to point out a "blessing" of trending factors. The blessing occurs if the magnitude of the factors is sufficiently large, in which case the otherwise so common asymptotic bias of the PC approach can be completely eliminated without imposing any additional restrictions on the cross-sectional and time series dependencies of the regression errors. This is noteworthy, because the sentiment in the previous literature is that in order to eliminate the asymptotic bias, the errors have to be independent.

The reminder of the paper is organized as follows. We begin by describing the IPC approach. This is done in Section 2. Sections 3 and 4 present the formal assumptions and our main asymptotic results, respectively. Section 6 concludes. In the online appendix, we provide (i) an empirical illustration using as an example the long-run relationship between US house prices and income, (ii) the proofs of our asymptotic results, and (iii) some results of secondary nature.

A word on notation. For any T-rowed full column rank matrix $\mathbf{A}$, we define its projection error matrix as $\mathbf{M}_A = \mathbf{I}_T - \mathbf{A}(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}' = \mathbf{I}_T - \mathbf{P}_A$. If $\mathbf{A}$ is square, $\lambda_{min}(\mathbf{A})$ and $\lambda_{max}(\mathbf{A})$ signify its smallest and largest eigenvalues, respectively, $\operatorname{tr}\mathbf{A}$ signifies its trace, and $\|\mathbf{A}\| = \sqrt{\operatorname{tr}\mathbf{A}'\mathbf{A}}$ and $\|\mathbf{A}\|_2 = \sqrt{\lambda_{max}(\mathbf{A}'\mathbf{A})}$ signify its Frobenius and spectral norms, respectively. We write $\mathbf{A} > 0$ to signify that $\mathbf{A}$ is positive definite. If $\mathbf{B}$ is also a matrix, $\operatorname{diag}(\mathbf{A}, \mathbf{B})$ denotes the block-diagonal matrix that takes $\mathbf{A}$ ($\mathbf{B}$) as the upper left (lower right) block. The symbols $\to_D$, $\to_P$ and $MN(\cdot,\cdot)$ signify convergence in distribution, convergence in probability and a mixed normal distribution, respectively.

We use $N, T \to \infty$ to indicate that the limit has been taken while passing both N and T to infinity. We use w.p.a.1 to denote with probability approaching one. Finally, $\mathbb{I}(A)$ is the indicator function for the event A.

2 The IPC procedure

Consider the stacked $T \times 1$ variable $\mathbf{y}_i = (y_{i,1}, \ldots, y_{i,T})'$, observable for $i = 1, \ldots, N$ cross-sectional units. The DGP that we will consider for this variable is given by

$$\mathbf{y}_i = \mathbf{X}_i \boldsymbol{\beta}^0 + \mathbf{F}^0 \boldsymbol{\gamma}_i^0 + \boldsymbol{\varepsilon}_i, \tag{1}$$

where $\mathbf{X}_i = (\mathbf{x}_{i,1}, \ldots, \mathbf{x}_{i,T})'$ is a $T \times d_x$ matrix of regressors, $\mathbf{F}^0 = (\mathbf{f}_1^0, \ldots, \mathbf{f}_T^0)'$ is a $T \times d_f$ matrix of unobservable common factors with $\boldsymbol{\gamma}_i^0$ being a conformable vector of factor loadings, and $\boldsymbol{\varepsilon}_i = (\varepsilon_{i,1}, \ldots, \varepsilon_{i,T})'$ is a $T \times 1$ vector of idiosyncratic errors. The interactive effects are here given by $\mathbf{F}^0 \boldsymbol{\gamma}_i^0$.

The factors are divided into groups according to their order of magnitude. There are G groups of size $d_1, \ldots, d_G$, which means that $d_1 + \cdots + d_G = d_f$. Because the grouping is unknown, we may without loss of generality assume that the factors are ordered, such that the first d_1 factors have the highest order of magnitude, the next d_2 factors have the second highest order, and so on. Hence, if we denote by $\mathbf{F}_g^0$ and $\boldsymbol{\gamma}_{g,i}^0$ the $T \times d_g$ matrix of factors and $d_g \times 1$ vector of loadings associated with group g, respectively, then $\mathbf{F}^0 \boldsymbol{\gamma}_i^0 = \sum_{g=1}^G \mathbf{F}_g^0 \boldsymbol{\gamma}_{g,i}^0$, where $\mathbf{F}^0 = (\mathbf{F}_1^0, \ldots, \mathbf{F}_G^0)$ and $\boldsymbol{\gamma}_i^0 = (\boldsymbol{\gamma}_{1,i}^{0\prime}, \ldots, \boldsymbol{\gamma}_{G,i}^{0\prime})'$.

The goal is to infer $\boldsymbol{\beta}^0$. The main difficulty in the estimation process is how to control for $\mathbf{F}^0$. Our proposed IPC estimation procedure consists of three steps. We first initialize the estimation procedure by applying the PC estimator of Bai (2009). However, because the first group of factors dominates all the other groups in terms of order of magnitude, the first-step PC factor estimator will only estimate (the space spanned by) $\mathbf{F}_1^0$. The second step of the procedure therefore involves iteratively applying the PCA conditional on previous factor estimates to estimate all subsequent groups of factors; hence, the "I" in IPC. In the third and final step, we estimate $\boldsymbol{\beta}^0$ conditional on the second-step IPC estimator of $\mathbf{F}^0$ and the first-step PC estimator of $\boldsymbol{\beta}^0$.

Step 1 (Initial estimation)**.** The objective function that we consider is given by

$$\text{SSR}(\boldsymbol{\beta}, \mathbf{F}) = \sum_{i=1}^N (\mathbf{y}_i - \mathbf{X}_i \boldsymbol{\beta})' \mathbf{M}_F (\mathbf{y}_i - \mathbf{X}_i \boldsymbol{\beta}), \tag{2}$$

where $\mathbf{F} \in \mathbb{D}_F$ with $\mathbb{D}_F = \{\mathbf{F}_{T \times d_{max}} : T^{-\delta} \mathbf{F}'\mathbf{F} = \mathbf{I}_{d_{max}}\}$, and $d_{max} \geq d_f$ and $\delta \in [0, \infty)$ are user-specified numbers. As we explain in Remark 1 below, the IPC estimator of $\boldsymbol{\beta}^0$ is invariant to the choice

of δ and the need to select d_{max} is standard. The initial estimator is the minimizer of $\text{SSR}(\boldsymbol{\beta}, \mathbf{F})$;

$$(\widehat{\boldsymbol{\beta}}_0, \widehat{\mathbf{F}}_0) = \underset{(\boldsymbol{\beta},\mathbf{F})\in\mathbb{D}}{\text{argmin}}\ \text{SSR}(\boldsymbol{\beta}, \mathbf{F}), \tag{3}$$

where $\mathbb{D} = \mathbb{R}^{d_x} \times \mathbb{D}_F$. It is useful to note that $\widehat{\boldsymbol{\beta}}_0$ satisfies $\widehat{\boldsymbol{\beta}}_0 = \widehat{\boldsymbol{\beta}}(\widehat{\mathbf{F}}_0)$, where

$$\widehat{\boldsymbol{\beta}}(\mathbf{F}) = \left(\sum_{i=1}^{N} \mathbf{X}_i'\mathbf{M}_F\mathbf{X}_i\right)^{-1} \sum_{i=1}^{N} \mathbf{X}_i'\mathbf{M}_F\mathbf{y}_i. \tag{4}$$

Step 2 (Iterative estimation of factors)**.** As already pointed out, the factors are estimated in order according to magnitude. Therefore, $\widehat{\mathbf{F}}_0$ is estimating $\mathbf{F}_1^0$. Since $d_1 \leq d_f \leq d_{max}$, in general the dimension of $\widehat{\mathbf{F}}_0$ will be larger than that of $\mathbf{F}_1^0$. We therefore begin this step of the estimation procedure by estimating d_1, and for this purpose we employ a version of the ratio of eigenvalue-based estimator considered by, for example, Lam and Yao (2012), and Ahn and Horenstein (2013), which is given by

$$\widehat{d}_1 = \underset{0\leq d\leq d_{max}}{\text{argmin}} \left\{ \frac{\widehat{\lambda}_{1,d+1}}{\widehat{\lambda}_{1,d}} \cdot \mathbb{I}\left(\frac{\widehat{\lambda}_{1,d}}{\widehat{\lambda}_{1,0}} \geq \tau_N\right) + \mathbb{I}\left(\frac{\widehat{\lambda}_{1,d}}{\widehat{\lambda}_{1,0}} < \tau_N\right) \right\}, \tag{5}$$

where $\tau_N = 1/\ln(\max\{\widehat{\lambda}_{1,0}, N\})$, $\widehat{\lambda}_{1,0} = N^{-1}\sum_{i=1}^{N} \|\mathbf{y}_i - \mathbf{X}_i\widehat{\boldsymbol{\beta}}_0\|^2$, and $\widehat{\lambda}_{1,1} \geq \cdots \geq \widehat{\lambda}_{1,d_{max}}$ are the d_{max} largest eigenvalues of the following matrix:

$$\widehat{\boldsymbol{\Sigma}}_1 = \frac{1}{N}\sum_{i=1}^{N} (\mathbf{y}_i - \mathbf{X}_i\widehat{\boldsymbol{\beta}}_0)(\mathbf{y}_i - \mathbf{X}_i\widehat{\boldsymbol{\beta}}_0)'. \tag{6}$$

The threshold τ_N, the "mock" eigenvalue $\widehat{\lambda}_{1,0}$, and the indicator function are there to ensure that the estimator is consistent. The need for these will be explained later. Given $\widehat{d}_1$, we update the estimate of $\mathbf{F}_1^0$ by setting $\widehat{\mathbf{F}}_1$ equal to the first $\widehat{d}_1$ columns of $\widehat{\mathbf{F}}_0$, and estimate $\boldsymbol{\gamma}_{1,i}^0$ by $\widehat{\boldsymbol{\gamma}}_{1,i} = T^{-\delta}\widehat{\mathbf{F}}_1'(\mathbf{y}_i - \mathbf{X}_i\widehat{\boldsymbol{\beta}}_0)$.

The estimation of $\mathbf{F}_2^0, \ldots, \mathbf{F}_G^0$ is analogous to that of $\mathbf{F}_1^0$. The main difference is that we have to condition on all previous estimates. Let us therefore use $\widehat{\mathbf{F}}_{-g} = (\widehat{\mathbf{F}}_1, \ldots, \widehat{\mathbf{F}}_{g-1})$ and $\widehat{\boldsymbol{\gamma}}_{-g,i} = (\widehat{\boldsymbol{\gamma}}_{1,i}', \ldots, \widehat{\boldsymbol{\gamma}}_{g-1,i}')'$ to denote the matrices containing the previously estimated factors and loadings, respectively, when estimating group g. The estimator of d_g is then given by

$$\widehat{d}_g = \underset{0\leq d\leq d_{max}}{\text{argmin}} \left\{ \frac{\widehat{\lambda}_{g,d+1}}{\widehat{\lambda}_{g,d}} \cdot \mathbb{I}\left(\frac{\widehat{\lambda}_{g,d}}{\widehat{\lambda}_{g,0}} \geq \tau_N\right) + \mathbb{I}\left(\frac{\widehat{\lambda}_{g,d}}{\widehat{\lambda}_{g,0}} < \tau_N\right) \right\}, \tag{7}$$

where we update τ_N by letting $\tau_N = 1/\ln(\max\{\widehat{\lambda}_{g,0}, N\})$, $\widehat{\lambda}_{g,0} = N^{-1}\sum_{i=1}^{N} \|\mathbf{M}_{\widehat{F}_{-g}}(\mathbf{y}_i - \mathbf{X}_i\widehat{\boldsymbol{\beta}}_0)\|^2$ and $\widehat{\lambda}_{g,1} \geq \cdots \geq \widehat{\lambda}_{g,d_{max}-\widehat{d}_{g-1}-\cdots-\widehat{d}_1}$ are the $d_{max} - \widehat{d}_{g-1} - \cdots - \widehat{d}_1$ largest eigenvalues of

$$\widehat{\boldsymbol{\Sigma}}_g = \frac{1}{N}\sum_{i=1}^{N} (\mathbf{y}_i - \mathbf{X}_i\widehat{\boldsymbol{\beta}}_0 - \widehat{\mathbf{F}}_{-g}\widehat{\boldsymbol{\gamma}}_{-g,i})(\mathbf{y}_i - \mathbf{X}_i\widehat{\boldsymbol{\beta}}_0 - \widehat{\mathbf{F}}_{-g}\widehat{\boldsymbol{\gamma}}_{-g,i})'. \tag{8}$$

The resulting estimator $\widehat{\mathbf{F}}_g$ of $\mathbf{F}_g^0$ is given by the eigenvectors associated with $\widehat{\lambda}_{g,1}, \ldots, \widehat{\lambda}_{g,\widehat{d}_g}$ and $\widehat{\boldsymbol{\gamma}}_{g,i} = T^{-\delta}\widehat{\mathbf{F}}_g'(\mathbf{y}_i - \mathbf{X}_i\widehat{\boldsymbol{\beta}}_0 - \widehat{\mathbf{F}}_{-g}\widehat{\boldsymbol{\gamma}}_{-g,i})$. New groups of factors are estimated until $\widehat{d}_g = 0$. At this point, we set $\widehat{G} = g - 1$ and define $\widehat{\mathbf{F}} = (\widehat{\mathbf{F}}_1, \ldots, \widehat{\mathbf{F}}_{\widehat{G}})$. This is the IPC estimator of $\mathbf{F}^0$.

Step 3 (Estimation of $\boldsymbol{\beta}^0$)**.** Given $\widehat{\mathbf{F}}$, we compute $\widehat{\boldsymbol{\beta}}_1 = \widehat{\boldsymbol{\beta}}(\widehat{\mathbf{F}})$ using (4). The IPC-based estimator of $\boldsymbol{\beta}^0$ is given by

$$\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}_0 + \left(\sum_{i=1}^{N} \widehat{\mathbf{Z}}_i'\widehat{\mathbf{Z}}_i\right)^{-1} \sum_{i=1}^{N} \mathbf{X}_i'\mathbf{M}_{\widehat{F}}\mathbf{X}_i(\widehat{\boldsymbol{\beta}}_1 - \widehat{\boldsymbol{\beta}}_0), \tag{9}$$

where $\widehat{\mathbf{Z}}_i = \mathbf{M}_{\widehat{F}}\mathbf{X}_i - \sum_{j=1}^{N} \mathbf{M}_{\widehat{F}}\mathbf{X}_j\widehat{a}_{ij}$ with $\widehat{a}_{ij} = \widehat{\boldsymbol{\gamma}}_i'(\widehat{\boldsymbol{\Gamma}}'\widehat{\boldsymbol{\Gamma}})^{-1}\widehat{\boldsymbol{\gamma}}_j$, $\widehat{\boldsymbol{\gamma}}_i = (\widehat{\boldsymbol{\gamma}}_{1,i}', \ldots, \widehat{\boldsymbol{\gamma}}_{\widehat{G},i}')'$ and $\widehat{\boldsymbol{\Gamma}} = (\widehat{\boldsymbol{\gamma}}_1, \ldots, \widehat{\boldsymbol{\gamma}}_N)'$.

Remark 1. In the bulk of the previous literature, the appropriate value of δ to use depends on whether $\mathbf{F}^0$ is stationary or unit root non-stationary (see, for example, Bai, 2004). The assumed knowledge of δ is therefore tantamount to assuming that the order of integration of $\mathbf{F}^0$ is known, which is not needed here. In fact, the IPC procedure is invariant with respect to δ, which can therefore be set arbitrarily. Choosing τ_N is analogous to choosing a suitable penalty in information criteria. The choice is therefore not unique. The main requirement is that τ_N should tend to zero at a slower rate than $\widehat{\lambda}_{g,d}/\widehat{\lambda}_{g,0}$. Extensive Monte Carlo experimentation suggests that $\tau_N = 1/\ln(\max\{\widehat{\lambda}_{g,0}, N\})$ works well in small-samples. The need to specify a maximum d_{max} for the number of factors is standard in the literature (see, for example, Bai and Ng, 2002).

Remark 2. The eigenvalue ratio $\widehat{\lambda}_{g,d+1}/\widehat{\lambda}_{g,d}$ is self-normalizing, which makes it possible to handle factors that are of different order of magnitude. Still, there are two issues. First, since $\widehat{\lambda}_{g,d+1}/\widehat{\lambda}_{g,d}$ is not defined for $d = 0$, we cannot have $d_f = 0$. The use of the mock eigenvalue $\widehat{\lambda}_{g,0}$ allows us to entertain this possibility. Second, the limiting behaviour of $\widehat{\lambda}_{g,d+1}/\widehat{\lambda}_{g,d}$ is unknown for $d > d_g$ (Lam and Yao, 2012). The use of the indicator functions allows us to circumvent this problem. The idea is to look at $\widehat{\lambda}_{g,d}$ only. If this eigenvalue is "small", we take it as a sign of $d > d_g$ and set $\widehat{\lambda}_{g,d+1}/\widehat{\lambda}_{g,d}$ to one. However, because the order of magnitude of $\mathbf{f}_t^0$ is assumed to be unknown, we cannot look at $\widehat{\lambda}_{g,d}$ directly but rather we look at $\widehat{\lambda}_{g,d}/\widehat{\lambda}_{g,0}$, which in contrast to $\widehat{\lambda}_{g,d}$ is self-normalizing.

Remark 3. Intuition suggests to take $\widehat{\boldsymbol{\beta}}_1$, the ordinary least squares (OLS) estimator conditional on $\widehat{\mathbf{F}}$, as the final estimator of $\boldsymbol{\beta}^0$ in Step 3. Interestingly, while consistent, because of the stepwise estimation of the factors, the asymptotic distribution of $\widehat{\boldsymbol{\beta}}_1$ is generally not (mixed) normal and nuisance parameter-free. In Section 5, we use Monte Carlo simulations to evaluate the extent of this non-normality.

3 Assumptions

Assumption 1 is a high-level moment condition concerned mainly with the order of magnitude of $\mathbf{f}_t^0$ and $\mathbf{x}_{i,t}$. The high-level formulation is convenient because it is the moment conditions that drive the distribution theory, and we are not specifically interested here in the various sets of conditions under which they hold. It may be noted, however, that there are a variety of more primitive conditions that lead to Assumption 1 (see Westerlund, 2018, for a discussion).

Assumption 1 (Moments)**.**

(a) There exists a matrix $\mathbf{\Sigma}_X$ such that $\mathbb{E}\|(NT)^{-1}\sum_{i=1}^N \mathbf{D}_T\mathbf{X}_i'\mathbf{M}_{F^0}\mathbf{X}_i\mathbf{D}_T - \mathbf{\Sigma}_X\|^2 = o(1)$, where $\mathbf{D}_T = \text{diag}(T^{-\kappa_1/2},\dots,T^{-\kappa_{d_x}/2})$ with $0 \leq \kappa_j < \infty$ for $j = 1,\dots,d_x$, $\mathbb{E}\|\mathbf{\Sigma}_X\|^2 < \infty$, and $0 < \lambda_{min}(\mathbf{\Sigma}_X) \leq \lambda_{max}(\mathbf{\Sigma}_X) < \infty$ w.p.a.1.

(b) $\|(NT)^{-1}\sum_{i=1}^N \mathbf{D}_T\mathbf{X}_i'\boldsymbol{\varepsilon}_i\| = O_P(1/\min\{\sqrt{N},\sqrt{T}\})$ and $\|\boldsymbol{\varepsilon}\|_2 = O_P(\max\{\sqrt{N},\sqrt{T}\})$ with $\boldsymbol{\varepsilon} = (\boldsymbol{\varepsilon}_1,\dots,\boldsymbol{\varepsilon}_N)$.

(c) There exists a matrix $\mathbf{\Sigma}_{F^0}$ such that $\mathbb{E}\|\mathbf{C}_T\mathbf{F}^{0\prime}\mathbf{F}^0\mathbf{C}_T - \mathbf{\Sigma}_{F^0}\|^2 = o(1)$, where $\mathbf{C}_T = \text{diag}(T^{-\nu_1/2}\mathbf{I}_{d_1},\dots,T^{-\nu_G/2}\mathbf{I}_{d_G})$ with $\nu_1 > \cdots > \nu_G > 1/2$, $\mathbb{E}\|\mathbf{\Sigma}_{F^0}\|^2 < \infty$ and $0 < \lambda_{min}(\mathbf{\Sigma}_{F^0}) \leq \lambda_{max}(\mathbf{\Sigma}_{F^0}) < \infty$ w.p.a.1.

(d) There exists a matrix $\mathbf{\Sigma}_{\Gamma^0}$ such that $\|N^{-1}\mathbf{\Gamma}^{0\prime}\mathbf{\Gamma}^0 - \mathbf{\Sigma}_{\Gamma^0}\| = o_P(1)$ and $\max_{i\geq 1}\mathbb{E}\|\boldsymbol{\gamma}_i^0\|^4 < \infty$, where $\mathbf{\Gamma}^0 = (\boldsymbol{\gamma}_1^0,\dots,\boldsymbol{\gamma}_N^0)'$ is $N \times d_f$ and $0 < \lambda_{min}(\mathbf{\Sigma}_{\Gamma^0}) \leq \lambda_{max}(\mathbf{\Sigma}_{\Gamma^0}) < \infty$.

Let us start with Assumption 1 (c), which is analogous to Assumption F of Westerlund (2018). This condition is very general in that it imposes almost no restrictions on the type of trending behaviour that $\mathbf{f}_t$ may have. The trending can be deterministic and/or stochastic. Either way, the degree of the trending is not restricted. The main requirement is that $\lambda_{min}(\mathbf{\Sigma}_{F^0}) > 0$ w.p.a.1, which implies that the elements of $\mathbf{f}_t$ cannot be asymptotically collinear. A majority of previous PC-based works assume that $T^{-1}\mathbf{F}^{0\prime}\mathbf{F}^0$ converges to positive definite matrix (see, for example, Bai, 2009, and Moon and Weidner, 2015). Notable exceptions include Bai (2004), and Bai et al. (2009), in which $\mathbf{f}_t^0$ is assumed to follow pure unit root process, and Bai and Ng (2004), who allow for a mix of stationary and unit root factors. The fact that $\mathbf{\Sigma}_{F^0}$ is not required to be a constant matrix means that we do not rule out factors that are stochastically integrated. We also do not place any restrictions on the long-run covariance matrix of differences of $\mathbf{f}_t^0$, which means that we permit linear combinations of factors that are of reduced integration order, commonly referred to as "multicointegration". This is similar to the scenario considered by Bai and Ng (2004), except that they restrict the order of integration of $\mathbf{f}_t^0$ to be at most one. The only study that comes close to ours in terms of the generality of the factors is that of Westerlund (2018). However, he assumes that $\mathbf{x}_{i,t}$ has a factor structure that loads on the same set of factors as $y_{i,t}$, which is not required here. Also, unlike Westerlund (2018),

we allow $1/2 < \nu_G < 1$, which means that the signal coming from $\mathbf{f}^0_{G,t}$ is even weaker than under stationarity, as when $\mathbf{f}^0_{G,t}$ is stationary and sparse. Similarly to Lam and Yao (2012), we refer to this type of factors as "signal-weak".

Assumption 1 (a) is similar to Assumption 1 (c) in that it leaves the trending behaviour of the regressors essentially unrestricted, provided that they are not asymptotically collinear.

The first requirement of Assumption 1 (b) is quite mild and holds if a central limit theorem in only one of the two panel dimensions applies to the normalized sum of $\mathbf{D}_T\mathbf{X}_i'\boldsymbol{\varepsilon}_i$. The second requirement is quite common in the literature, and is expected to hold as long as $\varepsilon_{i,t}$ has zero mean, and weak serial and cross-sectional correlation (see Moon and Weidner, 2015, for a discussion).

Assumption 1 (d) is standard and ensures that each factor has a non-trivial contribution to the variance of $y_{i,t}$ (see, for example, Bai and Ng, 2004, for a discussion).

Assumption 2 (Identification)**.** $\inf_{\mathbf{F}\in\mathbb{D}_F} \lambda_{min}(\mathbf{B}(\mathbf{F})) \geq c_0 > 0$ for all N and T, where $\mathbf{B}(\mathbf{F}) = (NT)^{-1}\sum_{i=1}^N \mathbf{D}_T\mathbf{Z}_i(\mathbf{F})'\mathbf{Z}_i(\mathbf{F})\mathbf{D}_T$ with $\mathbf{Z}_i(\mathbf{F}) = \mathbf{M}_F\mathbf{X}_i - \sum_{j=1}^N \mathbf{M}_F\mathbf{X}_j a_{ij}$ and $a_{ij} = \boldsymbol{\gamma}_i^{0\prime}(\boldsymbol{\Gamma}^{0\prime}\boldsymbol{\Gamma}^0)^{-1}\boldsymbol{\gamma}_j^0$.

Assumption 2 is analogous to Assumption A in Bai (2009), and Assumption NC in Moon and Weidner (2015), and is there to rule out "low-rank" elements in $\mathbf{X}_i$ that are wiped out by the defactoring and demeaning carried out in $\mathbf{Z}_i(\mathbf{F})$ to eliminate the interactive effects. The limitation here is therefore not that we cannot allow for low rank data components, which we do through the interactive effects, but that we cannot identify their effects if included among the observed regressors.

Assumption 3 (Errors)**.**

(a) $\mathbb{E}(\varepsilon_{i,t}) = 0$ and $\mathbb{E}(\boldsymbol{\varepsilon}_i\boldsymbol{\varepsilon}_i') = \boldsymbol{\Sigma}_{\varepsilon,i}$.

(b) Let $\boldsymbol{\varepsilon}_t = (\varepsilon_{1,t},\ldots,\varepsilon_{N,t})'$ in this assumption only. $\{\boldsymbol{\varepsilon}_t : t \geq 1\}$ is strictly stationary and α-mixing such that $\max_{i\geq 1}\mathbb{E}|\varepsilon_{i,1}|^{4+\mu} < \infty$ for some $\mu > 0$ and the mixing coefficient $\alpha(t) = \sup_{A\in\mathcal{F}^0_{-\infty}, B\in\mathcal{F}^\infty_t}|\mathbb{P}(A)\mathbb{P}(B) - \mathbb{P}(AB)|$ satisfies $\sum_{t=1}^\infty \alpha(t)^{\mu/(4+\mu)} < \infty$, where $\mathcal{F}^0_{-\infty}$ and $\mathcal{F}^\infty_t$ are the sigma-algebras generated by $\{\boldsymbol{\varepsilon}_s : s \leq 0\}$ and $\{\boldsymbol{\varepsilon}_s : s \geq t\}$, respectively.

(c) $\sum_{i,j=1}^N\sum_{t,s=1}^T |\mathbb{E}(\varepsilon_{i,t}\varepsilon_{j,s})| = O(NT)$ and $\sum_{i,j=1}^N |\sigma_{\varepsilon,ij}| = O(N)$, where $\sigma_{\varepsilon,ij} = \mathbb{E}(\varepsilon_{i,t}\varepsilon_{j,t})$.

(d) $\varepsilon_{i,t}$ is independent of $\boldsymbol{\gamma}_j^0$, $\mathbf{f}_s^0$ and $\mathbf{x}_{j,s}$ for all i, j, t and s.

Assumption 3 is similar to Assumptions C and D in Bai (2009). Assumptions 3 (b) and (c) ensure that the serial and cross-sectional dependencies of $\varepsilon_{i,t}$ are at most weak. Assumption 3 (d) requires that $\mathbf{x}_{i,t}$ and $\varepsilon_{i,t}$ are independent, which rules out the presence of lagged dependent variables in $\mathbf{x}_{i,t}$. However, $\mathbf{x}_{i,t}$ may still be correlated with the unobserved regression error $\boldsymbol{\gamma}_i^{0\prime}\mathbf{f}_t^0 + \varepsilon_{i,t}$ in (1), as the correlation between $\mathbf{x}_{i,t}$, $\boldsymbol{\gamma}_i^0$ and $\mathbf{f}_t^0$ is not restricted in any way. Hence, $\mathbf{x}_{i,t}$ is actually not required to be strictly exogenous even if we assume it to be independent of $\varepsilon_{i,t}$.

Assumption 4 (Factors and loadings)**.**

(a) $\max_{g\neq h}\|\mathbf{\Gamma}_g^{0\prime}\mathbf{\Gamma}_h^0\| = O_P(N^p)$ and $\max_{g\neq h}\|\mathbf{F}_g^{0\prime}\mathbf{F}_h^0\| = O_P(T^q)$, where $g,h = 1,\dots,G$, $G > 1$, $p < 1$, $q < (\nu_G + \nu_{G-1})/2$ and $\nu_{G-1} \geq 1$.

(b) If $\nu_G < 1$, then $T/N^2 \to c_1 \in [0,\infty)$.

The Assumption 4 (a) condition that $\nu_{G-1} \geq 1$ means that we only allow for one group of weak factors. This can be seen as a form of normalization and is not particularly restrictive. Let us therefore instead consider $\max_{g\neq h}\|\mathbf{F}_g^{0\prime}\mathbf{F}_h^0\| = O_P(T^q)$, which is less restrictive than the exact orthogonality condition typically required in papers on grouped factor structures (see, for example, Ando and Bai, 2017). Even so, we now provide a provide a justification for Assumption 4 (a). As is well known, $\boldsymbol{\gamma}_i^{0\prime}\mathbf{f}_t^0 = \boldsymbol{\gamma}_i^{0\prime}\mathbf{W}^{-1}\mathbf{W}\mathbf{f}_t^0$ for any positive definite rotation matrix $\mathbf{W}$. Now set $\mathbf{W} = (\mathbf{F}^{0\prime}\mathbf{F}^0\mathbf{C}_T^2)^{-1/2}$. This implies $\mathbf{C}_T\mathbf{W}\mathbf{F}^{0\prime}\mathbf{F}^0\mathbf{W}'\mathbf{C}_T = \mathbf{I}_{d_f}$, which means that the rotated factors are exactly orthogonal, and hence that Assumption 4 is satisfied for $\mathbf{F}^0\mathbf{W}'$ with $q = -\infty$. The loading condition, $\max_{g\neq h}\|\mathbf{\Gamma}_g^{0\prime}\mathbf{\Gamma}_h^0\| = O_P(N^p)$, can be justified in the same way.[1] Assumption 4 (b) is not required unless some of the factors are signal-weak.

4 Asymptotic results

Lemma 1 justifies the use of $\widehat{\boldsymbol{\beta}}_0$ in Step 1 of the IPC procedure as an initial estimator of $\boldsymbol{\beta}^0$.

Lemma 1 (Consistency of $\widehat{\boldsymbol{\beta}}_0$)**.** *Under Assumptions 1 and 2, as $N,T \to \infty$,*

$$\min\{\sqrt{N},\sqrt{T}\}\mathbf{D}_T^{-1}(\widehat{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}^0) = O_P(1).$$

According to Lemma 1, Assumptions 1 and 2 are enough to ensure that $\widehat{\boldsymbol{\beta}}_0$ is consistent for $\boldsymbol{\beta}^0$. The rate of convergence is given by $\|\mathbf{D}_T\|/\min\{\sqrt{N},\sqrt{T}\} = \max\{T^{-\kappa_1/2},\dots,T^{-\kappa_{d_x}/2}\}/\min\{\sqrt{N},\sqrt{T}\}$. To put this into perspective, suppose that $\mathbf{x}_{i,t}$ is stationary, such that $\kappa_1 = \cdots = \kappa_{d_x} = 0$. In this case, $\mathbf{D}_T = \mathbf{I}_{d_x}$ and the rate of convergence is given by $1/\min\{\sqrt{N},\sqrt{T}\}$, which is the slowest of the regular rates in pure time series and cross-section regressions. Still, the rate is fast enough for the estimation of the number of factors. This brings us to Step 2 of the estimation procedure.

Lemma 2 (Consistency of $(\widehat{d}_1,\dots,\widehat{d}_{G+1})$ and $\widehat{\mathbf{F}}$)**.** *Suppose that Assumptions 1–4 are satisfied. Then, the following results hold as N, $T \to \infty$:*

(a) $\mathbb{P}\big((\widehat{d}_1,\dots,\widehat{d}_{G+1}) = (d_1,\dots,d_{G+1})\big) \to 1$, *where $d_{G+1} = 0$;*

[1] Another way to rationalize Assumption 4 is if $\mathbf{\Gamma}_1^0,\dots,\mathbf{\Gamma}_G^0$ are independent and at most one of them has non-zero mean. Independence is often assumed and we therefore do not justify it here (see, for example, Chudik et al., 2011, and Pesaran, 2006). In order to justify the zero mean assumption, suppose for simplicity that $G = 2$, that $\gamma_{1,i}^0 = 1$ and that $\gamma_{2,i}^0 = \gamma_2^0 + \eta_i$ with $\mathbb{E}(\eta_i) = 0$. Hence, $\boldsymbol{\gamma}_i^{0\prime}\mathbf{f}_t^0 = (\gamma_{1,i}^0,\gamma_{2,i}^0)(f_{1,t}^0,f_{2,t}^0)' = f_{1,t}^0 + (\gamma_2^0+\eta_i)f_{2,t}^0 = (1,\eta_i)(\widetilde{f}_{1,t}^0,f_{2,t}^0)'$, where $\widetilde{f}_{1,t}^0 = (f_{1,t}^0 + \gamma_2^0 f_{2,t}^0)$. Then the zero mean assumption is fulfilled.

(b) $\|\mathbf{P}_{\widehat{F}} - \mathbf{P}_{F^0}\| = o_P(1)$.

The consistency of $(\widehat{d}_1, \ldots, \widehat{d}_G)$ is important for obvious reasons. The consistency of $\widehat{d}_{G+1}$ ensures that the stopping rule of Step 2 is asymptotically valid, which in turn implies that $\mathbb{P}(\widehat{G} = G) \to 1$.

As we alluded to earlier, $\mathbf{F}^0$ and $\boldsymbol{\gamma}_i^0$ are only identified up to a rotation matrix. However, we cannot claim that $\widehat{\mathbf{F}}$ is rotationally consistent for $\mathbf{F}^0$, as the number of rows of both objects is growing with T. We therefore have to resort to alternative consistency concepts. This is where Lemma 2 (b) comes in. It shows that the spaces spanned by $\widehat{\mathbf{F}}$ and $\mathbf{F}^0$ are asymptotically the same.

We have now established that all the estimates of Step 1 and Step 2 are consistent. We therefore move on to investigate the Step-3 IPC estimator $\widehat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}^0$. In Theorem 1 below we provide the asymptotic distribution of this estimator. In order to do so, however, we need to impose another two assumptions.

Assumption 5 (Rates)**.**

(a) $N/T^{\nu_G} \to \rho_1 \in [0, \infty)$;

(b) $T^{2-\nu_G}/N \to \rho_2 \in [0, \infty)$.

Assumption 6 (Asymptotic normality)**.**

$$\frac{1}{\sqrt{NT}} \sum_{i=1}^{N} \mathbf{D}_T \mathbf{Z}_i (\mathbf{F}^0)' \boldsymbol{\varepsilon}_i \to_D MN(\mathbf{0}_{d_x \times 1}, \boldsymbol{\Omega})$$

as N, $T \to \infty$, where $\boldsymbol{\Omega} = \text{plim}_{N,T\to\infty} \frac{1}{NT} \sum_{i=1}^{N} \sum_{j=1}^{N} \mathbf{D}_T \mathbb{E}[\mathbf{Z}_i (\mathbf{F}^0)' \boldsymbol{\varepsilon}_i \boldsymbol{\varepsilon}_j \mathbf{Z}_j (\mathbf{F}^0) | \mathcal{C}] \mathbf{D}_T$ with $\mathcal{C}$ being the sigma-algebra generated by $\mathbf{F}^0$.

If all the factors in $\mathbf{f}_t^0$ are stationary such that $G = 1$ and $\nu_G = \nu_1 = 1$, Assumption 5 requires that $N/T \to \rho_1 = 1/\rho_2 \in (0, \infty)$, which is the same condition as in Bai (2009), and Moon and Weidner (2015). Note also that Assumption 5 rules out the signal-weak case when $\nu_G < 1$.

Assumption 6 is a central limit theorem that is analogous to Assumption E of Bai (2009). The reason for requiring that the asymptotic distribution is mixed normal as opposed to normal is that by doing so we can accommodate stochastically integrated factors (see, for example, Bai et al., 2009). In the absence of such integrated factors, the mixed normal becomes normal. Either way, Assumption 6 ensures that standard normal and chi-squared inference based on $\widehat{\boldsymbol{\beta}}$ is possible.

Theorem 1 (Asymptotic distribution of $\widehat{\boldsymbol{\beta}}$)**.** *Under Assumptions 1–6, as N, $T \to \infty$,*

$$\sqrt{NT} \mathbf{D}_T^{-1} (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0) \to_D MN(\mathbf{B}_0^{-1} (\sqrt{\rho_1} \mathbf{A}_1 + \sqrt{\rho_2} \mathbf{A}_2), \mathbf{B}_0^{-1} \boldsymbol{\Omega} \mathbf{B}_0^{-1}),$$

where

$$\mathbf{B}_0 = \operatorname*{plim}_{N,T\to\infty} \mathbb{E}[\mathbf{B}(\mathbf{F}^0)|\mathcal{C}],$$

$$\mathbf{A}_1 = -\operatorname*{plim}_{N,T\to\infty} \frac{1}{T^{(1-\nu_G)/2}} \sum_{i=1}^{N} \mathbf{D}_T \mathbb{E}[\mathbf{X}_i' \mathbf{M}_{F^0} \boldsymbol{\Sigma}_\varepsilon \mathbf{F}_G^0 (\mathbf{F}_G^{0\prime}\mathbf{F}_G^0)^{-1} (\boldsymbol{\Gamma}_G^{0\prime}\boldsymbol{\Gamma}_G^0)^{-1} \boldsymbol{\gamma}_{G,i}^0 |\mathcal{C}],$$

$$\mathbf{A}_2 = -\operatorname*{plim}_{N,T\to\infty} \frac{1}{T^{(3-\nu_G)/2}} \sum_{i=1}^{N}\sum_{j=1}^{N} \mathbf{D}_T \mathbb{E}[\mathbf{Z}_i(0)' \mathbf{F}_G^0 (\mathbf{F}_G^{0\prime}\mathbf{F}_G^0)^{-1} (\boldsymbol{\Gamma}_G^{0\prime}\boldsymbol{\Gamma}_G^0)^{-1} \boldsymbol{\gamma}_{G,j}^0 \boldsymbol{\varepsilon}_j' \boldsymbol{\varepsilon}_i |\mathcal{C}],$$

with $\boldsymbol{\Sigma}_\varepsilon = N^{-1}\sum_{i=1}^N \boldsymbol{\Sigma}_{\varepsilon,i}$ *and* $\mathbf{Z}_i(0) = \mathbf{X}_i - \sum_{j=1}^N \mathbf{X}_j a_{ij}$.

According to Theorem 1, the asymptotic bias is driven by the factors and loadings of group G, which is intuitive as the factors of this group are smallest in order of magnitude. They therefore dominate the asymptotic bias. By bounding ν_G from below Assumption 5 ensures that $\mathbf{B}_0^{-1}(\sqrt{\rho_1}\mathbf{A}_1 + \sqrt{\rho_2}\mathbf{A}_2)$ is not diverging.

Let us now illustrate the implications of Theorem 1 taking as examples the cases when $\mathbf{f}_t^0$ is stationary and when it is unit root non-stationary. If stationarity holds, such that $G = \nu_1 = 1$ and $\rho_1 = 1/\rho_2 \in (0,\infty)$, the bias in Theorem 1 reduces to $\mathbf{B}_0^{-1}(\sqrt{\rho_1}\mathbf{A}_1 + \rho_1^{-1/2}\mathbf{A}_2)$, which under the additional condition that also $\mathbf{x}_{i,t}$ is stationary is identically the bias reported in Theorem 3 of Bai (2009). It is important to note that while $\sqrt{\rho_1}\mathbf{A}_1$ can be made arbitrarily small (large) by taking ρ_1 to zero (infinity), this will make $\rho_1^{-1/2}\mathbf{A}_2$ divergent (negligible). It follows that unless $\rho_1 \in (0,\infty)$ the bias will diverge, which in turn means that there is no way to make the bias disappear by just manipulating ρ_1, which in practical terms means restricting T/N. Indeed, as pointed out by Bai (2009), the only way to avoid bias under stationarity is to assume that $\varepsilon_{i,t}$ is homoskedastic and serially uncorrelated.[2]

A major point about Theorem 1 is that it showcases the importance of ν_G for the IPC bias. In particular, the theorem makes clear that by allowing $\nu_G > 1$, we can break the above mentioned inverse relationship between ρ_1 and ρ_2, which means that one can be zero without for that matter forcing the other to infinity. In particular, ρ_1 and ρ_2 may both be zero. Let us therefore now consider the case when $\mathbf{f}_t^0$ is unit root non-stationary. In this case, $\nu_G = 2$, implying that $T^{2-\nu_G}/N = 1/N \to 0$ as $N \to \infty$, and hence Assumption 5 (b) is satisfied with $\rho_2 = 0$. The part of the bias that emanates from $\sqrt{\rho_2}\mathbf{B}_0^{-1}\mathbf{A}_2$ is therefore zero. Hence, if we in addition assume that $N/T^{\nu_G} = N/T^2 \to 0$, so that $\rho_1 = 0$, then $\sqrt{\rho_1}\mathbf{B}_0^{-1}\mathbf{A}_1$ is zero too, and hence the bias is gone. This is consistent with Proposition 4 of Bai et al. (2009), which establishes that their version of the regular PC estimator is asymptotically unbiased under exogeneity if $N/T^2 \to 0$.

In general, the larger ν_G is, the less restrictive the condition on T/N has to be for ρ_1 and ρ_2 to

[2]It is easy to see that $\mathbf{A}_1 = \mathbf{0}_{d_x\times 1}$ if $\varepsilon_{i,t}$ is homoskedastic and serially uncorrelated, as $\mathbf{M}_{F^0}\boldsymbol{\Sigma}_\varepsilon\mathbf{F}_G^0 = \mathbf{0}_{T\times d_G}$ if $\boldsymbol{\Sigma}_\varepsilon = \sigma_\varepsilon^2\mathbf{I}_T$. The proof of $\mathbf{A}_2 = \mathbf{0}_{d_x\times 1}$ requires more work and can be found in Bai (2009, Proof of Theorem 2(ii)).

be zero. The intuition for this result is simple. Indeed, while $\mathbf{F}_G^0$ appears twice in the denominator of $\mathbf{A}_1$ and $\mathbf{A}_2$, it only appears once in the numerator. This "unbalancedness" together with $\|\mathbf{F}_G^0\| = O_P(T^{\nu_G/2})$ means that $\mathbf{A}_1$ and $\mathbf{A}_2$ are $O_P(T^{-\nu_G/2})$, and therefore the bias is decreasing in ν_G.

Remark 4. Note that while biased, $\widehat{\boldsymbol{\beta}}$ is still consistent at the best achievable rate.[3] This is in contrast to Lemma 1 and the relatively slow rate of convergence reported there. The reason for this difference is that unlike Theorem 1, which requires that Assumptions 1–6 all hold, Lemma 1 only requires Assumptions 1 and 2, and under these very relaxed conditions the Theorem 1 rate is not attainable. If, however, the conditions of Theorem 1 are met, then $\widehat{\boldsymbol{\beta}}_0$ and $\widehat{\boldsymbol{\beta}}_1$ are consistent at the same rate as $\widehat{\boldsymbol{\beta}}$ (see Lemma B.6 of the online appendix for a formal proof).

Corollary 1 (Unbiased asymptotic distribution)**.** *Suppose that the conditions of Theorem 1 are met and that $\rho_1 = \rho_2 = 0$. Then, as $N, T \to \infty$,*

$$\sqrt{NT}\mathbf{D}_T^{-1}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0) \to_D MN(\mathbf{0}_{d_x \times 1}, \mathbf{B}_0^{-1}\boldsymbol{\Omega}\mathbf{B}_0^{-1}). \tag{10}$$

In Section C of the online appendix, we provide some alternative conditions that ensure $\mathbf{A}_1 = \mathbf{A}_2 = \mathbf{0}_{d_x \times 1}$. If ρ_1, ρ_2, $\mathbf{A}_1$ and $\mathbf{A}_2$ are all different from zero, one possibility is to use bias correction. In the appendix, we explain how the Jackknife approach can be used for this purpose.

Theorem 1 imposes only minimal conditions on the correlation and heteroskedasticity of $\varepsilon_{i,t}$, and is in this sense very general. Such generality is, however, not possible if we also want to ensure consistent estimation of $\boldsymbol{\Omega}$. Let us therefore assume for a moment that $\mathbb{E}(\varepsilon_{i,t}\varepsilon_{j,s}) = 0$ for all $(i,t) \neq (j,s)$, so that $\varepsilon_{i,t}$ is serially and cross-sectionally uncorrelated. In this case,

$$\sum_{i=1}^N \sum_{j=1}^N \mathbb{E}[\mathbf{Z}_i(\mathbf{F}^0)'\boldsymbol{\varepsilon}_i\boldsymbol{\varepsilon}_j\mathbf{Z}_j(\mathbf{F}^0)|\mathcal{C}] = \sum_{i=1}^N \sigma_{\varepsilon,i}^2 \mathbb{E}[\mathbf{Z}_i(\mathbf{F}^0)'\mathbf{Z}_i(\mathbf{F}^0)|\mathcal{C}]. \tag{11}$$

A natural estimator of this matrix is given by $\sum_{i=1}^N \widehat{\sigma}_{\varepsilon,i}^2 \widehat{\mathbf{Z}}_i'\widehat{\mathbf{Z}}_i$, where $\widehat{\sigma}_{\varepsilon,i}^2 = T^{-1}\sum_{t=1}^T \|\mathbf{M}_{\widehat{F}}(\mathbf{y}_i - \mathbf{X}_i\widehat{\boldsymbol{\beta}})\|^2$ and $\widehat{\mathbf{Z}}_i$ is as in the definition of $\widehat{\boldsymbol{\beta}}$. It is not difficult to show that under the conditions of Theorem 1,

$$\left\| \frac{1}{NT}\sum_{i=1}^N \widehat{\sigma}_{\varepsilon,i}^2 \mathbf{D}_T\widehat{\mathbf{Z}}_i'\widehat{\mathbf{Z}}_i\mathbf{D}_T - \boldsymbol{\Omega} \right\| = o_P(1). \tag{12}$$

Of course, in this paper we do not assume knowledge of the order of the regressors, which in practice means that the appropriate normalization matrix $\mathbf{D}_T$ to use is unknown. This is not a problem, however, as the usual Wald and t-test statistics are self-normalizing. As an illustration, consider testing the null hypothesis of $H_0 : \mathbf{R}\boldsymbol{\beta}^0 = \mathbf{r}$, where $\mathbf{R}$ is a $r_0 \times d_x$ matrix of rank $r_0 \leq d_x$ and $\mathbf{r}$ is a

[3] For example, if $\mathbf{x}_{i,t}$ is stationary, such that $\mathbf{D}_T = \mathbf{I}_{d_x}$, the rate of convergence is given by $1/\sqrt{NT}$, which is the same as in Bai (2009). If, on the other hand, $\mathbf{x}_{i,t}$ is unit root non-stationary, such that $\mathbf{D}_T = T^{-1/2}\mathbf{I}_{d_x}$, then the rate of convergence is given by $1/\sqrt{N}T$, just as in Bai et al. (2009).

$r_0 \times 1$ vector. The Wald test statistic for testing this hypothesis is given by

$$W_{\widehat{\boldsymbol{\beta}}} = (\mathbf{R}\widehat{\boldsymbol{\beta}} - \mathbf{r})' \left[\mathbf{R} \left(\sum_{i=1}^{N} \widehat{\mathbf{Z}}_i' \widehat{\mathbf{Z}}_i \right)^{-1} \sum_{i=1}^{N} \widehat{\sigma}_{\varepsilon,i}^2 \widehat{\mathbf{Z}}_i' \widehat{\mathbf{Z}}_i \left(\sum_{i=1}^{N} \widehat{\mathbf{Z}}_i' \widehat{\mathbf{Z}}_i \right)^{-1} \mathbf{R} \right]^{-1} (\mathbf{R}\widehat{\boldsymbol{\beta}} - \mathbf{r}), \tag{13}$$

which has a limiting chi-squared distribution with r_0 degrees of freedom under H_0, as is clear from

$$\begin{aligned} W_{\widehat{\boldsymbol{\beta}}} = [\mathbf{R}\mathbf{D}_T\sqrt{NT}\mathbf{D}_T^{-1}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)]' \Bigg[\mathbf{R}\mathbf{D}_T \left(\frac{1}{NT} \sum_{i=1}^{N} \mathbf{D}_T \widehat{\mathbf{Z}}_i' \widehat{\mathbf{Z}}_i \mathbf{D}_T \right)^{-1} \frac{1}{NT} \sum_{i=1}^{N} \widehat{\sigma}_{\varepsilon,i}^2 \mathbf{D}_T \widehat{\mathbf{Z}}_i' \widehat{\mathbf{Z}}_i \mathbf{D}_T \\ \times \left(\frac{1}{NT} \sum_{i=1}^{N} \mathbf{D}_T \widehat{\mathbf{Z}}_i' \widehat{\mathbf{Z}}_i \mathbf{D}_T \right)^{-1} \mathbf{D}_T \mathbf{R} \Bigg]^{-1} \mathbf{R}\mathbf{D}_T\sqrt{NT}\mathbf{D}_T^{-1}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0) \to_D \chi^2(r_0). \end{aligned} \tag{14}$$

In the next section, we use Monte Carlo simulations as a means to evaluate the accuracy of this last result in small samples.

The above results are for the case when $\varepsilon_{i,t}$ is serially and cross-sectionally uncorrelated. If $\varepsilon_{i,t}$ is serially and/or cross-sectionally correlated, we recommend following Bai (2009), who discuss the issue of consistent covariance matrix estimation at length. The same arguments can be applied without change in current context.

5 Monte Carlo results

This section reports the results obtained from a small-scale Monte Carlo simulation exercise. The DGP considered for this purpose is given by a restricted version of (1) that sets $d_x = 2$, $\boldsymbol{\beta}^0 = \mathbf{1}_{2\times 1}$, $\varepsilon_{i,t} \sim N(0,1)$ and $N, T \in \{40, 80, 160, 320\}$. We further set $d_f = 3$ and generate the elements of $\boldsymbol{\gamma}_i = (\gamma_{1,i}^0, \gamma_{2,i}^0, \gamma_{3,i}^0)'$ as $\gamma_{1,i}^0 \sim N(1,1)$, $\gamma_{2,i}^0 \sim N(0,1)$ and $\gamma_{3,i}^0 \sim N(0,1)$. The elements of $\mathbf{f}_t^0 = (f_{1,t}^0, f_{2,t}^0, f_{3,t}^0)'$ are generated as $f_{1,t}^0 = t$, $f_{2,t}^0 = \mu_t$ and $f_{3,t}^0 = c_t$, where $\mu_t = \mu_{t-1} + \xi_t$, $\mu_0 = 0$, $\xi_t \sim N(0, 1/4)$ and $c_t = \sin(8\pi t/T)$. Hence, in this DGP, the common component is a random walk with drift and cycle. Also, $d_1 = d_2 = d_3 = 1$ and $(\nu_1, \nu_2, \nu_3) = (3, 2, 1)$. Let us denote by $x_{j,i,t}$ the j-th element of $\mathbf{x}_{i,t}$. The following specification makes $x_{j,i,t}$ correlated with the common component of $y_{i,t}$:

$$x_{j,i,t} = \frac{1}{d_x} \left(\sum_{j=1}^{d_f} |\gamma_{j,i}^0| + |\xi_t| + |c_t| \right) + \left(\frac{t}{4} \right)^{(j-1)/4} + v_{j,i,t}, \tag{15}$$

where $v_{j,i,t}$ is the i-th element of the $N \times 1$ vector $\mathbf{v}_{j,t} = (v_{j,1,t}, \ldots, v_{j,N,t})'$, which we generate as $\mathbf{v}_{j,t} = 0.5\mathbf{v}_{j,t-1} + \boldsymbol{\omega}_{j,t}$, where $\boldsymbol{\omega}_{j,t} \sim N(\mathbf{0}_{N\times 1}, \boldsymbol{\Sigma}_\omega)$ and $\boldsymbol{\Sigma}_\omega$ has $0.5^{|m-n|}$ in row m and column n. Thus, $v_{j,i,t}$ is weakly correlated across both i and t.

For each combination of N and T, we report the correct selection frequency for $(\widehat{d}_1, \ldots, \widehat{d}_{\widehat{G}})$ when seen as an estimator of $(d_1, \ldots, d_G)$ and for $\widehat{d}_g$ individually for each group g. Hence, while the former frequency captures the accuracy of the estimation of both $(d_1, \ldots, d_G)$ and G, the latter frequency

only captures the accuracy of the estimation of each d_g. We also report the root mean squared error (RMSE) of $\widehat{\boldsymbol{\beta}}$ and $\mathbf{P}_{\widehat{F}}$, as measured by the square root of the average of $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\|^2$ and $\|\mathbf{P}_{\widehat{F}} - \mathbf{P}_{F^0}\|^2$, respectively, over the replications. The RMSE of $\widehat{\boldsymbol{\beta}}$ is compared to that of $\widehat{\boldsymbol{\beta}}_0$, $\widehat{\boldsymbol{\beta}}_1$ and the infeasible OLS estimator of $\boldsymbol{\beta}^0$ based on taking $\mathbf{F}^0$ as known, $\widehat{\boldsymbol{\beta}}(\mathbf{F}^0)$. Some results on the Wald test at the 5% level based on $\mathbf{R} = \mathbf{I}_{d_x}$ and $\mathbf{r} = \boldsymbol{\beta}^0$ are also reported. In particular, we report the size of $W_{\widehat{\boldsymbol{\beta}}}$, $W_{\widehat{\boldsymbol{\beta}}_0}$ and $W_{\widehat{\boldsymbol{\beta}}_1}$, which are computed in an obvious fashion by replacing $(\widehat{\boldsymbol{\beta}}, \widehat{\mathbf{F}})$ with $(\widehat{\boldsymbol{\beta}}_0, \widehat{\mathbf{F}}_0)$ and $(\widehat{\boldsymbol{\beta}}_1, \widehat{\mathbf{F}})$, respectively, and $W_{\widehat{\boldsymbol{\beta}}(\mathbf{F}^0)}$, which is calculated in the same way as $W_{\widehat{\boldsymbol{\beta}}}$ but with $\mathbf{M}_{F^0}\mathbf{X}_i$ in place of $\widehat{\mathbf{Z}}_i$. The critical values are taken from $\chi^2(d_x)$. As pointed out in Section 2, the IPC estimator is invariant to δ. The results reported here are based on $\delta = 1$. We follow the bulk of the previous literature (see, for example, Ahn and Horenstein, 2013, Bai and Ng, 2002, and Moon and Weidner, 2015) and set $d_{max} = 10$, which led to the same results as some of the other values we tried. The number of replications is 1,000.

Insert Tables 1 and 2 about here

We begin by considering the results reported in Table 1 for the estimated common component. The correct selection frequency of each $\widehat{d}_g$ suggest that accuracy is decreasing in g, which is partly expected because the signal strength of the factors, as measured by ν_g, is decreasing in g too. Also, the sequential nature of the IPC estimation procedure implies that the error coming from the estimation of d_{g-1} will tend to be imported into the estimation of d_g, and therefore the procedure will be more accurate in the beginning. We also see that the accuracy of $(\widehat{d}_1, \ldots, \widehat{d}_{\widehat{G}})$ is almost identical to that of $\widehat{d}_G$, suggesting that the accuracy of $\widehat{G}$ is driven by the accuracy of the group whose factors has the weakest signal.

Looking next at the RMSE results reported in Table 2 for estimating $\boldsymbol{\beta}^0$, we see that there is a clear improvement as the sample size increases. The best overall performance is generally obtained when taking $\mathbf{F}^0$ as known, which is in accordance with our priori expectations. However, the improvement is not very large and it decreases with increases in N and T. The reason for this is the accuracy of the estimated factors, which according to the RMSE of $\mathbf{P}_{\widehat{F}}$ reported Table 1 is high and increasing in N and T. The second best performance is obtained by using $\widehat{\boldsymbol{\beta}}$, followed by $\widehat{\boldsymbol{\beta}}_1$, and then $\widehat{\boldsymbol{\beta}}_0$.

Insert Figure 1 about here

If our asymptotic theory is correct, while the rejection frequency of $W_{\widehat{\boldsymbol{\beta}}}$ and $W_{\widehat{\boldsymbol{\beta}}(\mathbf{F}^0)}$ should converge to the nominal level 5% as the sample size increases, that of $W_{\widehat{\boldsymbol{\beta}}_0}$ and $W_{\widehat{\boldsymbol{\beta}}_1}$ should not, and this is exactly what we see in Table 2. Note in particular how the size of $W_{\widehat{\boldsymbol{\beta}}_0}$ and $W_{\widehat{\boldsymbol{\beta}}_1}$ is not only nonconvergent but that it is in fact increasing in N and T. In order to illustrate these results, in Figure 1 we plot kernel smoothed versions of the empirical densities of the first element of all four estimators considered (after centering by $\boldsymbol{\beta}^0$ and scaling by $\sqrt{NT}$) as well as the normal density.

The first thing to note is that the densities of $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\beta}}(\mathbf{F}^0)$ approach the normal one as the sample size increases, and they are both unbiased. The densities of $\widehat{\boldsymbol{\beta}}_1$ and $\widehat{\boldsymbol{\beta}}_0$ are by contrast biased and occasionally even bimodal. Consistent with the size results reported in Table 2 we see that there is no improvement as N and T increase, but that the non-normality instead tends to get worse in larger samples.

6 Conclusion

The PC approach of Bai (2009) has attracted considerable interest in recent years so that it has given rise to a separate PC literature. A key assumption in this literature is that both the unknown factors and regressors are stationary, which is rarely the case in practice. In the present paper, we relax this assumption by considering a very general DGP in which the factors and regressors are essentially unrestricted. In spite of this generality, the proposed IPC estimator can be applied without any input from the practitioner, except for the maximum number of factors to be considered. The fact that in IPC there is no need to distinguish between deterministic and stochastic factors means that the usual problem in applied work of deciding on which deterministic terms to include in the model does not arise, as these are estimated along with the other factors of the model. There is also no need to pre-test the regressors for unit roots, which is otherwise standard practice when using procedures that do not require the data to be stationary. In other words, the proposed IPC is not only very general but also extremely user-friendly. It should therefore be a valuable addition to the already existing menu of techniques for panel regression models with interactive effects.

References

Ahn, S. C. and Horenstein, A. R. (2013), 'Eigenvalue ratio test for the number of factors', *Econometrica* **81**, 1203–1227.

Ando, T. and Bai, J. (2017), 'Clustering huge number of financial time series: A panel data approach with high-dimensional predictors and factor structures', *Journal of the American Statistical Association* **112**, 1182–1198.

Bai, J. (2004), 'Estimating cross-section common stochastic trends in nonstationary panel data', *Journal of Econometrics* **122**, 137–183.

Bai, J. (2009), 'Panel data models with interactive fixed effects', *Econometrica* **77**, 1229–1279.

Bai, J., Kao, C. and Ng, S. (2009), 'Panel cointegration with global stochastic trends', *Journal of Econometrics* **149**, 82–99.

Bai, J. and Ng, S. (2002), 'Determining the number of factors in approximate factor models', *Econometrica* **70**, 191–221.

Bai, J. and Ng, S. (2004), 'A Panic attack on unit roots and cointegration', *Econometrica* **72**, 1127–1177.

Chudik, A., Pesaran, M. H. and Tosetti, E. (2011), 'Weak and strong cross section dependence and estimation of large panels', *Econometrics Journal* **14**, 45–90.

Dong, C., Gao, J. and Peng, B. (2021), 'Varying-coefficient panel data models with nonstationarity and partially observed factor structure', *Journal of Business and Economic Statistics* **39**, 700–711.

Lam, C. and Yao, Q. (2012), 'Factor modeling for high-dimensional time series: Inference for the number of factors', *Annals of Statistics* **40**, 694–726.

Moon, H. R. and Weidner, M. (2015), 'Linear regression for panel with unknown number of factors as interactive fixed effects', *Econometrica* **83**, 1543–1579.

Park, J. Y. and Phillips, P. C. B. (2000), 'Nonstationary binary choice', *Econometrica* **68**, 1249–1280.

Pesaran, M. H. (2006), 'Estimation and inference in large heterogeneous panels with a multifactor error structure', *Econometrica* **74**, 967–1012.

Westerlund, J. (2018), 'CCE in panels with general unknown factors', *Econometrics Journal* **21**, 264–276.

Table 1: Monte Carlo results for the estimated common component.

		Correct selection frequency				RMSE
N	T	$\widehat{d}_1, ..., \widehat{d}_{\widehat{G}}$	$\widehat{d}_1$	$\widehat{d}_2$	$\widehat{d}_3$	$\mathbf{P}_{\widehat{F}}$
40	40	0.341	1.000	0.406	0.341	0.9453
	80	0.628	1.000	0.646	0.628	0.5599
	160	0.835	1.000	0.837	0.835	0.3715
	320	0.954	1.000	0.954	0.954	0.3520
80	40	0.348	1.000	0.378	0.348	0.9452
	80	0.661	1.000	0.668	0.661	0.4523
	160	0.867	1.000	0.867	0.867	0.2634
	320	0.958	1.000	0.958	0.958	0.2401
160	40	0.329	1.000	0.337	0.329	0.6401
	80	0.684	1.000	0.686	0.684	0.4359
	160	0.864	1.000	0.864	0.864	0.1761
	320	0.960	1.000	0.960	0.960	0.1696
320	40	0.220	1.000	0.225	0.220	0.4043
	80	0.603	1.000	0.603	0.603	0.2083
	160	0.882	1.000	0.882	0.882	0.1326
	320	0.988	1.000	0.988	0.988	0.1195

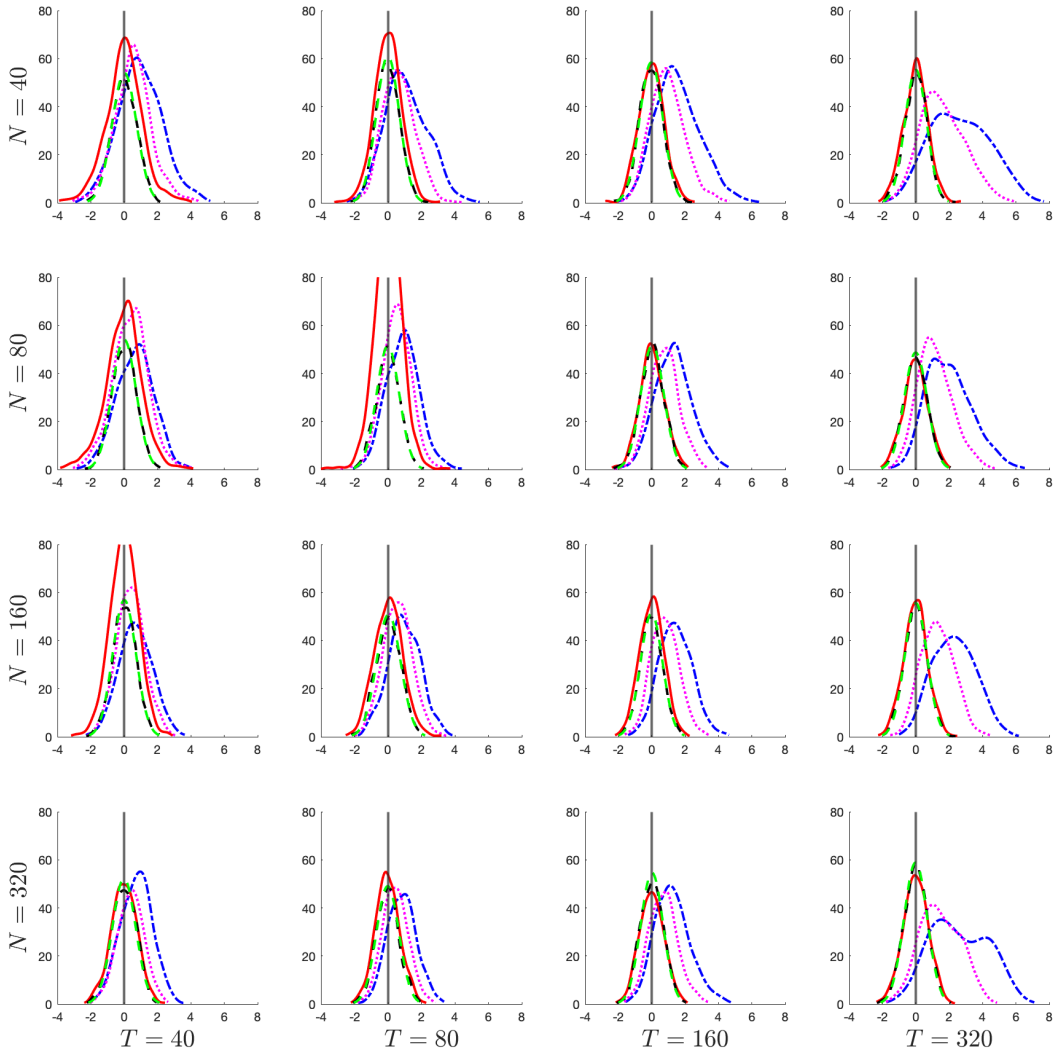
Notes: The correct selection frequencies are for the estimated factor groups, which is Step 2 of the IPC procedure. The results for $\widehat{d}_1, \ldots, \widehat{d}_{\widehat{G}}$ treat both the groups and their number, G, as unknown, while the results for $\widehat{d}_1$, $\widehat{d}_2$ and $\widehat{d}_3$ take G as given. The reported RMSE results for $\mathbf{P}_{\widehat{F}}$ refer to the square root of the average of $\|\mathbf{P}_{\widehat{F}} - \mathbf{P}_{F^0}\|^2$ across the Monte Carlo replications.

Table 2: Monte Carlo results for the estimated slopes.

		RMSE				5% size			
Estimator	$N\backslash T$	40	80	160	320	40	80	160	320
$\widehat{\boldsymbol{\beta}}_0$	40	0.0573	0.0423	0.0380	0.0410	0.6800	0.7280	0.8410	0.9160
	80	0.0335	0.0234	0.0212	0.0220	0.5990	0.6550	0.7940	0.9050
	160	0.0219	0.0169	0.0158	0.0169	0.5820	0.3500	0.8680	0.9560
	320	0.0148	0.0109	0.0105	0.0142	0.5680	0.6230	0.7750	0.9040
$\widehat{\boldsymbol{\beta}}_1$	40	0.0435	0.0296	0.0252	0.0262	0.3330	0.4320	0.6080	0.7520
	80	0.0286	0.0177	0.0144	0.0138	0.2520	0.3130	0.4720	0.6830
	160	0.0172	0.0127	0.0104	0.0102	0.2300	0.7260	0.5610	0.7560
	320	0.0113	0.0079	0.0069	0.0084	0.2330	0.2770	0.4580	0.7060
$\widehat{\boldsymbol{\beta}}$	40	0.0383	0.0212	0.0129	0.0091	0.1320	0.0950	0.0580	0.0680
	80	0.0280	0.0146	0.0092	0.0062	0.1500	0.0670	0.0650	0.0550
	160	0.0164	0.0105	0.0064	0.0043	0.0960	0.0750	0.0690	0.0430
	320	0.0099	0.0064	0.0044	0.0032	0.0720	0.0480	0.0420	0.0660
$\widehat{\boldsymbol{\beta}}(\mathbf{F}^0)$	40	0.0254	0.0171	0.0114	0.0081	0.0550	0.0610	0.0440	0.0410
	80	0.0186	0.0117	0.0084	0.0057	0.0750	0.0470	0.0560	0.0410
	160	0.0127	0.0086	0.0059	0.0041	0.0560	0.0600	0.0530	0.0490
	320	0.0087	0.0059	0.0041	0.0030	0.0640	0.0420	0.0520	0.0660

Notes: The RMSE of $\widehat{\boldsymbol{\beta}}$ refers to the square root of the average of $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\|^2$ across the Monte Carlo replications. The RMSEs of $\widehat{\boldsymbol{\beta}}_0$, $\widehat{\boldsymbol{\beta}}_1$ and $\widehat{\boldsymbol{\beta}}(\mathbf{F}^0)$ are constructed in an analogous fashion, where $\widehat{\boldsymbol{\beta}}_0$ is the initial Step 1 IPC estimator, $\widehat{\boldsymbol{\beta}}_1$ is the regular PC estimator based on the IPC estimator of the factors, $\widehat{\mathbf{F}}$, and $\widehat{\boldsymbol{\beta}}(\mathbf{F}^0)$ is the infeasible OLS estimator based on the true factors. The 5% size results are for the Wald test associated with each estimator.

Figure 1: Simulated densities for the estimated slopes.



Notes: The plotted curves are kernel smoothed empirical densities of the first element of each of $\sqrt{NT}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)$ (red solid line), $\sqrt{NT}(\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}^0)$ (magenta dotted line), $\sqrt{NT}(\widehat{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}^0)$ (blue dash-dotted line) and $\sqrt{NT}(\widehat{\boldsymbol{\beta}}(\mathbf{F}^0) - \boldsymbol{\beta}^0)$ (black dashed line). The normal density (green dashed line) is also included for reference.