Fixed T Estimation of Linear Panel Data Models with Interactive Fixed Effects[*]

Ayden Higgins[†]
University of Cambridge

October 11, 2021

Abstract

This paper studies the estimation of linear panel data models with interactive fixed effects, where one dimension of the panel, typically time, may be fixed. To this end, a novel transformation is introduced that reduces the model to a lower dimension, and, in doing so, relieves the model of incidental parameters in the cross-section. The central result of this paper demonstrates that transforming the model and then applying the principal component (PC) estimator of Bai (2009) delivers $\sqrt{n}$ consistent estimates of regression slope coefficients with T fixed. Moreover, these estimates are shown to be asymptotically unbiased in the presence of cross-sectional dependence, serial dependence, and with the inclusion of dynamic regressors, in stark contrast to the usual case. The large n, large T properties of this approach are also studied, where many of these results carry over to the case in which n is growing sufficiently fast relative to T. Transforming the model also proves to be useful beyond estimation, a point illustrated by showing that with T fixed, the eigenvalue ratio test of Ahn and Horenstein (2013) provides a consistent test for the number of factors when applied to the transformed model.

Keywords: interactive fixed effects, dynamic panels, factor models.
JEL classification: C13, C33, C38.

[*] This work was supported by the European Research Council through the grant ERC-2016-STG-715787-MiMo. I am grateful to Valentina Corradi, Koen Jochmans, Federico Martellosio, João Santos Silva and Martin Weidner for their comments and suggestions.

[†] Email: amh239@cam.ac.uk.

1 Introduction

This paper contributes to the extensive literature on linear panel data models with interactive effects. These models have proven to be very popular since, in many situations, the existence of such structures is well motivated; for example, arising due to unobserved heterogeneity across individuals, or exposure to common shocks. The model studied in this paper assumes that, in a panel with entries indexed $i = 1, ..., n$, $t = 1, ..., T$, outcomes are generated according to

$$\boldsymbol{y}_t = \alpha \boldsymbol{y}_{t-1} + \boldsymbol{X}_t \boldsymbol{\beta} + \boldsymbol{\Lambda}^* \boldsymbol{f}_t^* + \boldsymbol{\varepsilon}_t, \tag{1.1}$$

where $\boldsymbol{y}_t$ and $\boldsymbol{\varepsilon}_t$ are $n \times 1$ vectors of outcomes and error terms, respectively, $\boldsymbol{X}_t$ is an $n \times K$ matrix of exogenous covariates, $\boldsymbol{\Lambda}^*$ is an $n \times R^*$ matrix of time-invariant factor loadings, and $\boldsymbol{f}_t^*$ is an $R^* \times 1$ vector of time-varying factors. It is assumed that both the outcomes and the covariates are observed by the econometrician, while the factors, the loadings, and the error terms are not. The parameter of interest in this model is the $(K+1) \times 1$ vector $\boldsymbol{\theta} := (\alpha, \boldsymbol{\beta}^\top)^\top$, comprised of the scalar autoregressive parameter α and the $K \times 1$ vector $\boldsymbol{\beta}$.

This model can be seen as a generalisation of familiar models of additive effects, such as individual, time or group effects. For example, individual and time effects nest as a special case of (1.1) in which

$$\boldsymbol{\Lambda}^* = \begin{pmatrix} \lambda_1 & 1 \\ \vdots & \vdots \\ \lambda_n & 1 \end{pmatrix}, \quad \boldsymbol{f}_t^* = \begin{pmatrix} 1 \\ f_t \end{pmatrix}, \tag{1.2}$$

that is, where a vector of heterogeneous loadings is interacted with a unit factor, and where a vector of unit loadings is interacted with a time-varying factor. More generally, however, with interactive effects, no restrictions are placed on the factors or the loadings to be multiples of unit vectors, or otherwise, and both are permitted to be fully heterogeneous.

The main obstacle to consistent estimation of $\boldsymbol{\theta}$ arises in situations where the unobserved interactive effects are somehow correlated with covariates in the model. In this event, an endogeneity problem arises resulting in standard estimators, such as least squares, producing biased estimates. One response to this is to treat the components of the factor term as additional parameters to estimate, known as the fixed effects approach. Doing this has the benefit of allowing for arbitrary correlation between the covariates, the factors and the loadings, in contrast to its main rival, random effects. However, treating both the factors and loadings as fixed effects gives rise to incidental parameters in both dimensions of the panel, which, in turn, may generate significant complications for the estimation of the parameter of interest $\boldsymbol{\theta}$, on account of the incidental parameter problem; see Neyman and Scott (1948). Where both n and T are large this problem can, to some extent, be overcome, and estimation procedures have been developed for large panels which, under a broad array of circumstances, will produce consistent estimates of $\boldsymbol{\theta}$. Yet where the time dimension of the panel is small, the methods that are presently available tend to be restrictive and/or difficult to implement.

The two main approaches currently used for short panels are the common correlated effects estimator introduced in Pesaran (2006), and the quasi-difference approach of Ahn et al. (2013). The first of these assumes that the latent factors in the error term also impact some model covariates, such that the factors can be instrumented by cross-sectional averages. These instruments can then be levered to purge the factor term and, as a result, give rise to an estimator that is consistent with either T fixed, or where both n and T diverge. The drawback to this approach, however, is that it relies crucially on imposing a particular functional form for the relationship between the latent factors and model covariates, which, in many instances, may not be easy to justify. The second method, the quasi-difference approach, takes advantage of the inherent indeterminacy associated with factor models. By normalising the factors and loadings in a certain way, the model can be multiplied by a difference matrix to purge the factor term. The authors apply GMM to estimate both the difference matrix and the slope coefficients, yet since the moment conditions the model yields are highly non-linear, this generates a difficult optimisation problem which, as pointed out by Hayakawa (2016), may not satisfy the identification conditions (which are a necessary precursor to consistency) of their GMM procedure. Indeed the problem is not unique to their approach, and several closely related methods which rely on the same normalisation also suffer this affliction.

Notwithstanding the contributions of these authors, there remains scope for a general and simple to implement method for the estimation of linear panel data models with interactive fixed effects, where the time dimension is small relative to the size of the cross-section, or indeed, may be fixed. The present paper address this by introducing a new estimation approach at the centre of which lies a transformation that relieves the model of incidental parameters in the cross-section. In contrast to other approaches, the objective of this transformation is not to purge the incidental parameters from the model entirely, but rather to transfer those in the cross-section into the time dimension, and, in doing so, facilitate the estimation of the model in situations where T is small. The most appealing aspect of this transformation is its simplicity, since it is constructed directly from the data and applied to the model prior to estimation without introducing any additional parameters. Moreover, it is shown to have remarkably far-reaching consequences, and, in the main result of this paper, it is established that simply transforming the model and then applying a third estimator, the PC estimator of Bai (2009), will produce $\sqrt{n}$ consistent and asymptotically unbiased estimates with T fixed, irrespective of the possible inclusion of dynamic regressors and/or the presence of cross-sectional and serial dependence in the error term. This contrasts sharply with the usual case where, with fixed T, the PC estimator is, in general, both inconsistent and biased outside of exceptional circumstances.

Outline: Section 2 sets out the estimation approach, first introducing a transformation, and then going on to describe why the PC estimator is well suited to the task of estimating the transformed model. Section 3 begins the study of the asymptotic properties of the estimator by establishing consistency, under quite general conditions, and drawing a comparison between the result obtained here and those obtained when both n and T are large. Following this, an asymptotic expansion of the objective function is derived in Section 4, from which the asymptotic distribution of the estimator is then established in Section 5. In order to paint a more complete picture of the estimator, multiple results are provided to cover both the situation in which T is fixed, and where both n and T diverge. Some additional considerations are collected in Section 6, including estimation of the number of factors and alternative approaches to treating the initial condition. Monte Carlo simulations follow in Section 7. Section 8 concludes. Additional discussion and proofs of the results can be found in the appendices.

Notation: Throughout the paper, all vectors and matrices are real unless stated otherwise. For an $n \times 1$ vector $\boldsymbol{a}$ with elements a_i, $||\boldsymbol{a}||_1 := \sum_{i=1}^n |a_i|$, $||\boldsymbol{a}||_2 := \sqrt{\sum_{i=1}^n a_i^2}$, $||\boldsymbol{a}||_\infty := \max_{1\leq i\leq n} |a_i|$. Let $\boldsymbol{A}$ be an $n \times m$ matrix with elements A_{ij}. When $m = n$, and the eigenvalues of $\boldsymbol{A}$ are real, they are denoted $\mu_{\min}(\boldsymbol{A}) := \mu_n(\boldsymbol{A}) \leq ... \leq \mu_1(\boldsymbol{A}) =: \mu_{\max}(\boldsymbol{A})$. The following matrix norms are those induced by their vector counterparts: $||\boldsymbol{A}||_1 := \max_{1\leq j\leq m} \sum_{i=1}^n |A_{ij}|$ which is the maximum absolute column sum of $\boldsymbol{A}$, $||\boldsymbol{A}||_2 := \sqrt{\mu_{\max}(\boldsymbol{A}^\top \boldsymbol{A})}$, and $||\boldsymbol{A}||_\infty := \max_{1\leq i\leq n} \sum_{j=1}^m |A_{ij}|$ which is the maximum absolute row sum of $\boldsymbol{A}$. The Frobenius norm of $\boldsymbol{A}$ is denoted $||\boldsymbol{A}||_F := \sqrt{\sum_{i=1}^n \sum_{j=1}^m A_{ij}^2} = \sqrt{\mathrm{tr}(\boldsymbol{A}^\top \boldsymbol{A})}$. Let $\boldsymbol{P}_{\boldsymbol{A}} := \boldsymbol{A}(\boldsymbol{A}^\top \boldsymbol{A})^+ \boldsymbol{A}^\top$ and $\boldsymbol{M}_{\boldsymbol{A}} := \boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{A}}$, where $\boldsymbol{I}_n$ is the $n \times n$ identity matrix and $+$ denotes the Moore-Penrose generalised inverse. An $n \times 1$ vector of ones is denoted $\boldsymbol{\iota}_n$, and an $n \times m$ matrix of zeros is denoted $\boldsymbol{0}_{n\times m}$. The operation $\mathrm{vec}(\cdot)$ applied to an $n \times m$ matrix $\boldsymbol{A}$ creates an $nm \times 1$ vector $\mathrm{vec}(\boldsymbol{A})$ by stacking the columns of $\boldsymbol{A}$. The operation $\mathrm{diag}(\cdot)$ applied to an $n \times n$ matrix $\boldsymbol{B}$ creates an $n \times n$ diagonal matrix $\mathrm{diag}(\boldsymbol{B})$ which contains the diagonal elements of $\boldsymbol{B}$ along its diagonal, and $\mathrm{diagv}(\boldsymbol{B})$ is used as shorthand for $\mathrm{diag}(\boldsymbol{B})\boldsymbol{\iota}_n$.

For a matrix $\boldsymbol{A}$ which potentially has an increasing dimension, $\boldsymbol{\mathcal{O}}_p(1)$ is used to indicate that $||\boldsymbol{A}||_2 = \mathcal{O}_p(1)$ and similarly $\boldsymbol{o}_p(1)$ signifies that $||\boldsymbol{A}||_2 = o_p(1)$. Throughout, c is used to denote some arbitrary positive constant, with indexation often indicating to which quantity the constant is associated; e.g. c_x or c_λ, and 'w.p.a.1' indicates 'with probability approaching 1'.

2 Estimation Approach

Introduced in Bai (2009), the PC estimator is one of the foremost approaches taken to estimate models with interactive fixed effects, in situations where both n and T are large. This estimator is shown by the author to deliver consistent estimates of regression slope coefficients, and of rotational counterparts to the factors and the loadings, where the number of factors is known, and both n and T diverge. Further results have been provided by Moon and Weidner (2015, 2017) who demonstrate that the estimator remains consistent with the number of factors unknown, but not underestimated, and also with the possible inclusion of predetermined regressors. These authors establish the asymptotic properties of the PC estimator and, in particular, document asymptotic biases that arise in the presence of cross-sectional and serial dependence, and due to inclusion of predetermined regressors. These biases originate from the incidental parameter problem, and although it is sometimes possible to mitigate their impact when both n and T are large, in situations where T is small relative to n, or indeed fixed, they have proven to be more implacable, so much so that use of the PC estimator has been confined almost entirely to the large n, large T setting. Yet, as is shown shortly, it is possible to resolve many of these issues by first transforming the model, and then going on to apply this estimator in the usual way.

2.1 Transformation

It is useful to begin by re-writing the model in matrix form. Let the $n \times T$ matrix $\boldsymbol{Y} := (\boldsymbol{y}_1, ..., \boldsymbol{y}_T)$, $\boldsymbol{X}_k$ be the $n \times T$ matrix containing observations of the k-th covariate, the $T \times R^*$ matrix $\boldsymbol{F}^* := (\boldsymbol{f}_1^*, ..., \boldsymbol{f}_T^*)^\top$, and $\boldsymbol{S}(\alpha) := \boldsymbol{I}_T - \alpha \boldsymbol{W}$, where $\boldsymbol{W}$ is a $T \times T$ shift matrix with zeros everywhere, except those elements directly above the main diagonal, which take a value of 1. With this notation, the model can be written more succinctly as

$$\boldsymbol{Y}\boldsymbol{S}(\alpha) = \sum_{k=1}^{K} \beta_k \boldsymbol{X}_k + \boldsymbol{y}_0 \boldsymbol{s}^\top(\alpha) + \boldsymbol{\Lambda}^* \boldsymbol{F}^{*\top} + \boldsymbol{\varepsilon}, \tag{2.1}$$

where

$$\boldsymbol{s}(\alpha) \coloneqq \begin{pmatrix} \alpha & \mathbf{0}_{1\times(T-1)} \end{pmatrix}^\top . \tag{2.2}$$

In any dynamic panel model where T is small, special care must be taken with the initial condition $\boldsymbol{y}_0\boldsymbol{s}^\top(\alpha)$ since this may itself be endogenous. In what follows the initial condition is treated as an additional parameter in the model and is absorbed into the factor term.[1] As such, define $\boldsymbol{\Lambda} \coloneqq (\boldsymbol{y}_0, \boldsymbol{\Lambda}^*)$, $\boldsymbol{F}(\alpha) \coloneqq (\boldsymbol{s}(\alpha), \boldsymbol{F}^*)$, and $R \coloneqq R^* + 1$, whereby (2.1) becomes

$$\boldsymbol{Y}\boldsymbol{S}(\alpha) = \sum_{k=1}^{K} \beta_k \boldsymbol{X}_k + \boldsymbol{\Lambda}\boldsymbol{F}^\top + \boldsymbol{\varepsilon}, \tag{2.3}$$

with the dependence of $\boldsymbol{F}$ on α being suppressed. Now, define the $n \times TK$ matrix $\boldsymbol{\mathcal{X}} \coloneqq (\boldsymbol{X}_1, ..., \boldsymbol{X}_K)$ which is assumed to have full column rank. Moreover, assume hereafter that $TK \leq n$.[2] Consider the following group of transformations $\mathcal{G}$, where each element in this group is a bijective mapping from the sample space to itself:

$$\mathcal{G} \coloneqq \{\boldsymbol{Q} \in \mathcal{O}(n) : \boldsymbol{Q}\boldsymbol{\mathcal{X}} = \boldsymbol{\mathcal{X}}\}, \tag{2.4}$$

with $\mathcal{O}(n)$ being the group of $n \times n$ orthogonal matrices. This group $\mathcal{G}$ contains orthogonal transformations that preserve the column space of $\boldsymbol{\mathcal{X}}$. Take some $\boldsymbol{Q} \in \mathcal{G}$. This can be partitioned as

$$\boldsymbol{Q} \eqqcolon \begin{pmatrix} \boldsymbol{Q}_{\mathcal{X}} & \boldsymbol{Q}_{\mathcal{X}^\perp} \end{pmatrix}, \tag{2.5}$$

where $\boldsymbol{Q}_{\mathcal{X}}$ is an $n \times TK$ matrix with orthonormal columns such that $\boldsymbol{Q}_{\mathcal{X}}^\top \boldsymbol{Q}_{\mathcal{X}} = \boldsymbol{I}_{TK}$ and $\boldsymbol{Q}_{\mathcal{X}}\boldsymbol{Q}_{\mathcal{X}}^\top = \boldsymbol{P}_{\mathcal{X}}$, and, similarly, $\boldsymbol{Q}_{\mathcal{X}^\perp}$ is an $n \times (n - TK)$ matrix with orthonormal columns such that $\boldsymbol{Q}_{\mathcal{X}^\perp}^\top \boldsymbol{Q}_{\mathcal{X}^\perp} = \boldsymbol{I}_{(n-TK)}$ and $\boldsymbol{Q}_{\mathcal{X}^\perp}\boldsymbol{Q}_{\mathcal{X}^\perp}^\top = \boldsymbol{M}_{\mathcal{X}}$. Simply put, the matrix $\boldsymbol{Q}_{\mathcal{X}}$ projects into the TK-dimensional space spanned by the columns of the matrix $\boldsymbol{\mathcal{X}}$, while $\boldsymbol{Q}_{\mathcal{X}^\perp}$ on the other hand, projects into the space orthogonal to this. A simple way to construct $\boldsymbol{Q}_{\mathcal{X}}$ would be as $\boldsymbol{\mathcal{X}}(\boldsymbol{\mathcal{X}}^\top\boldsymbol{\mathcal{X}})^{-\frac{1}{2}}$, and, with this in hand, the following transformed variables can be defined:

$$\tilde{\boldsymbol{Y}} \coloneqq \boldsymbol{Q}_{\mathcal{X}}^\top \boldsymbol{Y}, \tag{2.6}$$

$$\tilde{\boldsymbol{X}}_k \coloneqq \boldsymbol{Q}_{\mathcal{X}}^\top \boldsymbol{X}_k, \tag{2.7}$$

$$\tilde{\boldsymbol{\Lambda}} \coloneqq \boldsymbol{Q}_{\mathcal{X}}^\top \boldsymbol{\Lambda}, \tag{2.8}$$

$$\tilde{\boldsymbol{\varepsilon}} \coloneqq \boldsymbol{Q}_{\mathcal{X}}^\top \boldsymbol{\varepsilon}, \tag{2.9}$$

[1] An alternative approach which involves dropping the first observation is discussed in Section 6.3.

[2] Many of the results in this paper will carry over naturally to the small n, large T setting by interchanging n and T.

in which case premultiplying (2.3) by $\boldsymbol{Q}_{\mathfrak{X}}^{\top}$ yields the transformed model

$$\tilde{\boldsymbol{Y}}\boldsymbol{S}(\alpha) = \sum_{k=1}^{K} \beta_k \tilde{\boldsymbol{X}}_k + \tilde{\boldsymbol{\Lambda}}\boldsymbol{F}^{\top} + \tilde{\boldsymbol{\varepsilon}}. \tag{2.10}$$

Looking at (2.10) there are three significant consequences of transforming the model through $\boldsymbol{Q}_{\mathfrak{X}}$ that need to be highlighted. First, the resultant matrices $\tilde{\boldsymbol{Y}}$, $\tilde{\boldsymbol{X}}_k$ and $\tilde{\boldsymbol{\Lambda}}\boldsymbol{F}^{\top}$ are of dimension $TK \times T$, since the entirety of the model has been transformed by $\boldsymbol{Q}_{\mathfrak{X}}$ and projected into the TK-dimensional subspace spanned by the columns of the covariates. Hence, the dimension of the factor term $\tilde{\boldsymbol{\Lambda}}\boldsymbol{F}^{\top}$ will no longer depend on n, thereby reliving the model of incidental parameters as $n \to \infty$. Second, the transformation leads to no loss of information in the covariates since, by construction, transforming the model though $\boldsymbol{Q}_{\mathfrak{X}}$ preserves the columns space of $\mathfrak{X}$. Thirdly, since the covariates used in the construction of $\boldsymbol{Q}_{\mathfrak{X}}$ are strictly exogenous, under quite general conditions, including broad cross-sectional and serial dependence, the transformation renders the error term asymptotically negligible (in a precise scene) so long as $T/n \to 0$. Indeed, it is this final property in particular that will prove key to estimation of (2.10) by principal components.

2.2 Estimation by Principle Components

The intuition underlying the PC estimator is that, given the factors and the loadings, the coefficients can be estimated by least squares, and, similarly, given $\boldsymbol{\theta}$, estimating the factors and loadings is a standard principal component problem. Where T is small relative to n, it is this latter step that proves to be challenging; in particular estimating the n-dimensional factor loadings. For this reason it is useful to consider the factor term in isolation in order to demonstrate the key differences that lie between PC estimation of the original model, and of its transformed counterpart.

Assume that $\boldsymbol{\theta}$ is observed and define $\dot{\boldsymbol{Y}} := \boldsymbol{Y}\boldsymbol{S}(\alpha) - \sum_{k=1}^{K} \beta_k \boldsymbol{X}_k = \boldsymbol{\Lambda}\boldsymbol{F}^{\top} + \boldsymbol{\varepsilon}$ which has a pure factor structure. Let $\check{\boldsymbol{\Lambda}}$ and $\check{\boldsymbol{F}}$ be $n \times R$ and $T \times R$ matrices, respectively, which satisfy $\check{\boldsymbol{\Lambda}}\check{\boldsymbol{F}}^{\top} = \boldsymbol{\Lambda}\boldsymbol{F}^{\top}$, $\frac{1}{n}\check{\boldsymbol{\Lambda}}^{\top}\check{\boldsymbol{\Lambda}} = \boldsymbol{I}_R$ and $\check{\boldsymbol{F}}^{\top}\check{\boldsymbol{F}}$ being diagonal.[3] Consider the problem of

[3] It is straightforward to see that such matrices exist. For example, by the singular value decomposition, decompose $\boldsymbol{\Lambda}\boldsymbol{F}^{\top} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^{\top}$. Let $\check{\boldsymbol{\Lambda}}$ be the R columns of $\sqrt{n}\boldsymbol{U}$ associated with the nonzero singular values and $\check{\boldsymbol{F}}^{\top}$ be the corresponding R rows of $\boldsymbol{S}\boldsymbol{V}^{\top}/\sqrt{n}$. As the columns of $\boldsymbol{U}$ and $\boldsymbol{V}$ are orthogonal, and $\boldsymbol{S}$ is diagonal, it follows that $\check{\boldsymbol{\Lambda}}^{\top}\check{\boldsymbol{\Lambda}}/n = \boldsymbol{I}_R$, $\check{\boldsymbol{F}}^{\top}\check{\boldsymbol{F}}$ is diagonal and $\check{\boldsymbol{\Lambda}}\check{\boldsymbol{F}}^{\top} = \boldsymbol{\Lambda}\boldsymbol{F}^{\top}$.

trying to estimate the loadings from the variance of $\dot{\boldsymbol{Y}}$.

$$\begin{aligned}\frac{1}{nT}\dot{\boldsymbol{Y}}\dot{\boldsymbol{Y}}^\top\check{\boldsymbol{\Lambda}} &= \frac{1}{nT}(\boldsymbol{\Lambda}\boldsymbol{F}^\top + \boldsymbol{\varepsilon})(\boldsymbol{\Lambda}\boldsymbol{F}^\top + \boldsymbol{\varepsilon})^\top\check{\boldsymbol{\Lambda}} \\ &= \frac{1}{nT}\boldsymbol{\Lambda}\boldsymbol{F}^\top\boldsymbol{F}\boldsymbol{\Lambda}^\top\check{\boldsymbol{\Lambda}} + \frac{1}{nT}\boldsymbol{\Lambda}\boldsymbol{F}^\top\boldsymbol{\varepsilon}^\top\check{\boldsymbol{\Lambda}} + \frac{1}{nT}\boldsymbol{\varepsilon}\boldsymbol{F}\boldsymbol{\Lambda}^\top\check{\boldsymbol{\Lambda}} + \frac{1}{nT}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top\check{\boldsymbol{\Lambda}}. \end{aligned} \tag{2.11}$$

With suitable conditions on the errors, the factors, and the loadings, as $n \to \infty$,

$$\frac{1}{nT}\dot{\boldsymbol{Y}}\dot{\boldsymbol{Y}}^\top\check{\boldsymbol{\Lambda}} = \frac{1}{nT}\check{\boldsymbol{\Lambda}}\check{\boldsymbol{F}}^\top\check{\boldsymbol{F}} + \frac{1}{nT}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top\check{\boldsymbol{\Lambda}} + \boldsymbol{o}_p(1). \tag{2.12}$$

Given that $\frac{1}{n}\check{\boldsymbol{\Lambda}}^\top\check{\boldsymbol{\Lambda}} = \boldsymbol{I}_R$ and $\check{\boldsymbol{F}}^\top\check{\boldsymbol{F}}$ is diagonal, then, without the term $\frac{1}{nT}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top\check{\boldsymbol{\Lambda}}$, $\check{\boldsymbol{\Lambda}}$ would be an eigenvector of $\frac{1}{nT}\dot{\boldsymbol{Y}}\dot{\boldsymbol{Y}}^\top$ asymptotically, and would provide a rotational counterpart to the factor loadings $\boldsymbol{\Lambda}$. Where both n and T are large, several authors have shown that, in spite of this distortionary term, estimating the loadings in the manner above is still possible in certain circumstances. For example, under the condition $||\boldsymbol{\varepsilon}||_2 = \mathcal{O}_p(\sqrt{\max\{n, T\}})$ employed in Moon and Weidner (2015), dependence in the error term is sufficiently limited that $\frac{1}{nT}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top\check{\boldsymbol{\Lambda}} = \boldsymbol{o}_p(1)$ as $n, T \to \infty$. Alternatively, where $\frac{1}{nT}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top \xrightarrow{p} \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}$, it may be possible to jointly estimate $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}$, and then a rotation of $\boldsymbol{\Lambda}$ as an eigenvector of $\frac{1}{nT}\dot{\boldsymbol{Y}}\dot{\boldsymbol{Y}}^\top - \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}$. Nonetheless, in either case it is only under the most exceptional of circumstances that the distortions caused by $\frac{1}{nT}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top\check{\boldsymbol{\Lambda}}$ do not affect the estimation of the parameter $\boldsymbol{\theta}$, and, moreover, neither case generally applies to the situation where T is fixed.

Now consider, on the other hand, the transformed model. Let $\check{\tilde{\boldsymbol{\Lambda}}}$ denote an analogue of $\check{\boldsymbol{\Lambda}}$. With $\frac{1}{nT}||\tilde{\boldsymbol{\varepsilon}}\boldsymbol{F}\tilde{\boldsymbol{\Lambda}}^\top\check{\tilde{\boldsymbol{\Lambda}}}||_2 = \mathcal{O}_p(1)$, one arrives at a similar expression to (2.12)

$$\frac{1}{nT}\dot{\tilde{\boldsymbol{Y}}}\dot{\tilde{\boldsymbol{Y}}}^\top\check{\tilde{\boldsymbol{\Lambda}}} = \frac{1}{nT}\check{\tilde{\boldsymbol{\Lambda}}}\check{\boldsymbol{F}}^\top\check{\boldsymbol{F}} + \frac{1}{nT}\tilde{\boldsymbol{\varepsilon}}\tilde{\boldsymbol{\varepsilon}}^\top\check{\tilde{\boldsymbol{\Lambda}}} + \boldsymbol{o}_p(1). \tag{2.13}$$

Yet now, since the regressors used to construct $\boldsymbol{Q}_{\mathcal{X}}$ are strictly exogenous, under quite general conditions $\frac{1}{nT}||\tilde{\boldsymbol{\varepsilon}}\tilde{\boldsymbol{\varepsilon}}^\top||_2 = \frac{1}{nT}||\boldsymbol{\varepsilon}^\top\boldsymbol{P}_{\mathcal{X}}\boldsymbol{\varepsilon}||_2 = \mathcal{O}_p(1)$, even with fixed T. As a consequence, asymptotically, $\check{\tilde{\boldsymbol{\Lambda}}}$ will be an eigenvector of $\frac{1}{nT}\dot{\tilde{\boldsymbol{Y}}}\dot{\tilde{\boldsymbol{Y}}}^\top$ and thus it is possible to estimate the space spanned by $\tilde{\boldsymbol{\Lambda}}$ with fixed T, where this was not possible for $\boldsymbol{\Lambda}$. This, heuristically, is why applying the PC estimator to the transformed model is able to deliver consistent estimates with T fixed.

2.3 Objective Function

Following Moon and Weidner (2015), the transformed model (2.10) can be estimated by minimising the following least squares objective function:

$$\mathcal{Q}(\boldsymbol{\theta}, \tilde{\boldsymbol{\Lambda}}, \boldsymbol{F}) := \frac{1}{nT}\mathrm{tr}\left(\left(\tilde{\boldsymbol{Y}}\boldsymbol{S}(\alpha) - \sum_{k=1}^{K}\beta_k\tilde{\boldsymbol{X}}_k - \tilde{\boldsymbol{\Lambda}}\boldsymbol{F}^\top\right)^\top\left(\tilde{\boldsymbol{Y}}\boldsymbol{S}(\alpha) - \sum_{k=1}^{K}\beta_k\tilde{\boldsymbol{X}}_k - \tilde{\boldsymbol{\Lambda}}\boldsymbol{F}^\top\right)\right). \tag{2.14}$$

Both the factors and the transformed loadings can be concentrated out of (2.14), in which case one arrives at an objective function involving $\boldsymbol{\theta}$ alone,

$$\mathcal{Q}(\boldsymbol{\theta}) := \frac{1}{nT}\sum_{r=R+1}^{T}\mu_r\left(\left(\tilde{\boldsymbol{Y}}\boldsymbol{S}(\alpha) - \sum_{k=1}^{K}\beta_k\tilde{\boldsymbol{X}}_k\right)^\top\left(\tilde{\boldsymbol{Y}}\boldsymbol{S}(\alpha) - \sum_{k=1}^{K}\beta_k\tilde{\boldsymbol{X}}_k\right)\right), \tag{2.15}$$

that is, the profile objective function now involves the sum of the $(T-R)$ smallest eigenvalues of the right-hand matrix.[4] Using this, the estimator $\hat{\boldsymbol{\theta}}$ can then be defined as

$$\hat{\boldsymbol{\theta}} := \underset{\boldsymbol{\theta}\in\Theta}{\arg\min}\,\mathcal{Q}(\boldsymbol{\theta}). \tag{2.16}$$

Remark 1. Since the PC estimator draws on the factor model literature in several important ways, it is unsurprising that, in order to simultaneously estimate the n-dimensional factor loadings, and T-dimensional factors, it is usually necessary that both n and T diverge. This can be related to results on pure factor models, where only exceptionally will principal component methods produce consistent estimates of both factors and loadings with fixed T; see, for instance, Theorem 4 of Bai (2003).

Remark 2. Reducing the dimension of the factor term may relieve the model of incidental parameters in the cross-section, but the effect of these parameters does not disappear entirely. Their effect is still present through $\tilde{\boldsymbol{\Lambda}}$, the part of the factor loadings that remains, which manifests itself as an additional incidental parameter in the time dimension.

Remark 3. This paper focuses on the case where, with the exception of lagged outcomes, the regressors are strictly exogenous, as in Bai (2009) and Moon and Weidner (2015). If instead some of the covariates $\boldsymbol{X}_k$ are endogenous, but valid instruments for these are available, then those instruments can substitute for $\boldsymbol{X}_k$ in the construction of $\boldsymbol{\mathfrak{X}}$.

[4] See Appendix A.1 for details.

Remark 4. When estimating the original model, the least squares objective function can be interpreted as the negative of a quasi-maximum likelihood function that uses the standard normal distribution. The objective function (2.14) can then be interpreted as a marginal quasi-likelihood which uses only a part of this.

3 Consistency

This section studies the asymptotic properties of the estimator defined in (2.16), beginning by deriving a general consistency result. As in Moon and Weidner (2015), throughout the following, both $\mathbf{\Lambda}$ and $\boldsymbol{F}$ are treated as fixed parameters in estimation and the superscript 0 is now introduced to distinguish true parameter values. Moreover, let $\boldsymbol{S} := \boldsymbol{S}(\alpha^0)$, $\boldsymbol{G} := \boldsymbol{S}^{-1}\boldsymbol{W}$, and $\mathcal{C}$ denote $\sigma(\boldsymbol{X}_1, ..., \boldsymbol{X}_K)$, that is, the sub-algebra generated by the exogenous covariates. The following assumptions are made.

Assumption MD (Model)**.**

(i) The parameter vector $\boldsymbol{\theta}^0$ lies in the interior of Θ, where Θ is a compact subset of $\mathbb{R}^{K+1}$ in which $|\alpha| < 1$.

(ii) The elements of the matrices $\boldsymbol{X}_1, ..., \boldsymbol{X}_K$, $\mathbf{\Lambda}^0$ and $\boldsymbol{F}^0$ have uniformly bounded fourth moments.

Assumption MD(i) assumes that the dynamic process is stationary, which allows $\boldsymbol{y}_t$ to be expanded as an infinite series by recursive substitution. Assumption MD(ii) imposes standard conditions on the moments of the covariates, the factors, and the loadings.

Assumption ER (Error)**.**

(i) $\mathbb{E}[\varepsilon_{it}|\mathcal{C}] = 0$ for $i = 1, ..., n$, $t = 1, ..., T$.

(ii) Let $\sigma^2_{ij,t\tau} = \mathbb{E}[\varepsilon_{it}\varepsilon_{j\tau}|\mathcal{C}]$. Then $|\sigma^2_{ij,t\tau}| < C$ uniformly for all i, j, t, τ, and the error term is weakly conditionally cross-sectionally and serially dependent, that is, $\sum_{i \neq j} |\sigma^2_{ij,t\tau}| \leq C$ uniformly for all j, t, τ, and $\sum_{t \neq \tau} |\sigma^2_{ij,t\tau}| \leq C$ uniformly for all i, j, τ.

Assumption ER(i) imposes strict exogeneity of the regressors as in Bai (2009) and Moon and Weidner (2015). Assumption ER(ii) limits the degree of dependence between the errors in the cross-section and across time, while allowing for heteroskedasticity in both dimensions of the panel. Different notions of dependence appear throughout the panel literature, and this can be modelled in several ways. Assumption ER(ii) is quite general in this regard.

It is important to point out that the least squares objective function given in (2.15) implicitly uses the reduced form of the dynamic process to generate an internal instrument for the autoregressive parameter. To see this, notice that $\boldsymbol{S}^{-1}(\alpha) = \boldsymbol{I}_T + \alpha \boldsymbol{G}(\alpha)$. Substituting this into the reduced form then yields

$$\begin{aligned}\tilde{\boldsymbol{Y}} &= \left(\sum_{k=1}^{K} \beta_k \tilde{\boldsymbol{X}}_k + \tilde{\boldsymbol{\Lambda}} \boldsymbol{F}^\top + \tilde{\boldsymbol{\varepsilon}} \right) \boldsymbol{S}^{-1}(\alpha) \\ &= \alpha \left(\sum_{k=1}^{K} \beta_k \tilde{\boldsymbol{X}}_k \boldsymbol{G}(\alpha) \right) + \sum_{k=1}^{K} \beta_k \tilde{\boldsymbol{X}}_k + (\tilde{\boldsymbol{\Lambda}} \boldsymbol{F}^\top + \tilde{\boldsymbol{\varepsilon}}) \boldsymbol{S}^{-1}(\alpha). \end{aligned} \tag{3.1}$$

In this way the role that $\sum_{k=1}^{K} \beta_k \tilde{\boldsymbol{X}}_k \boldsymbol{G}(\alpha)$ plays as an instrument for α is clear. Going forward it is useful to collect this instrument and the other exogenous covariates into a single matrix of regressors. Therefore let $\tilde{\boldsymbol{Z}}_1 \coloneqq \sum_{k=1}^{K} \beta_k \tilde{\boldsymbol{X}}_k \boldsymbol{G}(\alpha)$, $\tilde{\boldsymbol{Z}}_{k+1} \coloneqq \tilde{\boldsymbol{X}}_k$ for $k = 1, .., K$, $\boldsymbol{\delta} \cdot \tilde{\boldsymbol{Z}} \coloneqq \sum_{\kappa=1}^{K+1} \delta_k \tilde{\boldsymbol{Z}}_\kappa$ and define $\tilde{\boldsymbol{\mathcal{Z}}} \coloneqq (\mathrm{vec}(\tilde{\boldsymbol{Z}}_1), ..., \mathrm{vec}(\tilde{\boldsymbol{Z}}_{K+1})) \in \mathbb{R}^{KT^2 \times (K+1)}$.

Assumption CS (Consistency)**.**

(i) $R \geq R^0 \coloneqq \mathrm{rank}(\tilde{\boldsymbol{\Lambda}}^0 \boldsymbol{F}^{0\top})$.

(ii) $\min_{\boldsymbol{\delta} \in \mathbb{R}^{K+1} : ||\boldsymbol{\delta}||_2 = 1} \sum_{r=R+R^0+1}^{T} \mu_r \left(\frac{1}{nT} (\boldsymbol{\delta} \cdot \tilde{\boldsymbol{Z}})^\top (\boldsymbol{\delta} \cdot \tilde{\boldsymbol{Z}}) \right) \geq b > 0$.

Assumption CS(i) allows for the true number of factors R^0 to be unknown as long as the number of factors used in estimation R is no less than R^0. Notice also that this condition concerns the rank of $\tilde{\boldsymbol{\Lambda}}^0 \boldsymbol{F}^{0\top}$ and not of $\boldsymbol{\Lambda}^0 \boldsymbol{F}^{0\top}$, that is, R^0 is the number of factors correlated with the covariates. Assumption CS(ii) is a multicollinearity condition and can intuitively be understood by realising that it implies $\inf_{\tilde{\boldsymbol{\Lambda}} \in \mathbb{R}^{TK \times R^0}, \boldsymbol{F} \in \mathbb{R}^{T \times R}} \mu_{K+1}(\tilde{\boldsymbol{\mathcal{Z}}}^\top (\boldsymbol{M}_{\boldsymbol{F}} \otimes \boldsymbol{M}_{\tilde{\boldsymbol{\Lambda}}}) \tilde{\boldsymbol{\mathcal{Z}}})$ is bounded away from zero. This, therefore, asserts that the data matrix $\tilde{\boldsymbol{\mathcal{Z}}}^\top \tilde{\boldsymbol{\mathcal{Z}}}$ retains a sufficient level of variation, after having been projected orthogonal to arbitrary $R \times T$ factors and $R^0 \times TK$ loadings.

Proposition 1 (Consistency – General)**.** *Under Assumptions MD, ER and CS,*

$$||\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0||_2 = \mathcal{O}_p\left(\sqrt{\frac{T}{n}} \right). \tag{3.2}$$

Proposition 1 demonstrates that as $T/n \to 0$ the estimator is consistent. Moreover, where T is fixed, it is in fact $\sqrt{n}$ consistent. This result is obtained under quite general dependence in the error, and as long as the number of factors used in estimation is no less than the true number. Notice also that no assumptions have been made regarding the factors and the loadings other than bounded fourth moments; for instance, these may be strong, weak, or non-existent.

Proposition 1 can be compared directly to Theorem 4.1 in Moon and Weidner (2015), which, under similar terms, provides a consistency result for the PC estimator applied to the original model. Their result establishes that

$$||\boldsymbol{\theta}^0 - \hat{\boldsymbol{\theta}}||_2 = \mathcal{O}_p\left(\frac{1}{\sqrt{\min\{n,T\}}}\right), \tag{3.3}$$

with this rate being determined largely by the condition $||\boldsymbol{\varepsilon}||_2 = \mathcal{O}_p(\sqrt{\max\{n,T\}})$ (Assumption SN(ii)), under which[5]

$$\frac{||\boldsymbol{\varepsilon}||_2}{\sqrt{nT}} = \mathcal{O}_p\left(\frac{1}{\sqrt{\min\{n,T\}}}\right). \tag{3.4}$$

In similar fashion, the rate obtained in Proposition 1 can be attributed to the quantity $||\tilde{\boldsymbol{\varepsilon}}||_F$ which plays an analogous role in this paper. Under Assumption ER this can be shown to satisfy

$$\frac{||\tilde{\boldsymbol{\varepsilon}}||_F}{\sqrt{nT}} = \mathcal{O}_p\left(\sqrt{\frac{T}{n}}\right). \tag{3.5}$$

Recalling the discussion in Section 2.2, it is worth stressing again that while the difference between the quantities $\boldsymbol{\varepsilon}$ and $\tilde{\boldsymbol{\varepsilon}}$ may appear superficial, it is in fact of the utmost significance. For example, consider the textbook assumption of identically and independently distributed conditionally homoskedastic errors; i.e. $\mathbb{E}[\varepsilon_{it}\varepsilon_{j\tau}|\mathcal{C}] = \sigma^2$ for $i = j$, $t = \tau$ and is zero otherwise. In this case,

$$\begin{aligned}
\mathbb{E}[||\tilde{\boldsymbol{\varepsilon}}||_F^2] = \mathbb{E}[||\boldsymbol{Q}_{\mathcal{X}}^\top \boldsymbol{\varepsilon}||_F^2] &= \mathbb{E}\left[\sum_{t=1}^{TK}\sum_{\tau=1}^{T}\sum_{j=1}^{n}\sum_{i=1}^{n} \mathbb{E}[(Q_{\mathcal{X}})_{it}(Q_{\mathcal{X}})_{jt}\varepsilon_{i\tau}\varepsilon_{j\tau}|\mathcal{C}]\right] \\
&= \sigma^2 \sum_{t=1}^{TK}\sum_{\tau=1}^{T}\sum_{i=1}^{n} \mathbb{E}[(Q_{\mathcal{X}})_{it}^2] \\
&= \sigma^2 T \mathbb{E}\left[||\boldsymbol{Q}_{\mathcal{X}}||_F^2\right] \\
&= \sigma^2 T^2 K = \mathcal{O}(T^2),
\end{aligned} \tag{3.6}$$

[5] Moreover, (3.4) also proves to be important for the asymptotic expansion of the objective function; see Section 4.

from which it then follows by Markov's inequality that $||\tilde{\varepsilon}||_F = \mathcal{O}_p(T)$, and so

$$\frac{||\tilde{\varepsilon}||_F}{\sqrt{nT}} = \mathcal{O}_p\left(\sqrt{\frac{T}{n}}\right) = o_p(1), \tag{3.7}$$

as $T/n \to 0$. By comparison,

$$\frac{||\varepsilon||_2}{\sqrt{nT}} \geq \frac{1}{\sqrt{nT}} \frac{1}{\sqrt{\min\{n,T\}}} ||\varepsilon||_F \xrightarrow{p} \frac{\sigma}{\sqrt{\min\{n,T\}}}, \tag{3.8}$$

using $\frac{1}{\sqrt{\mathrm{rank}(\boldsymbol{A})}}||\boldsymbol{A}||_F \leq ||\boldsymbol{A}||_2$ and noting that, in this particular example,

$$\frac{||\varepsilon||_F}{\sqrt{nT}} = \frac{1}{\sqrt{nT}}\sqrt{\sum_{t=1}^{T}\sum_{i=1}^{n}\varepsilon_{it}^2} \xrightarrow{p} \sigma. \tag{3.9}$$

Therefore, even in this simple case, $\frac{||\varepsilon||_2}{\sqrt{nT}}$ cannot be $o_p(1)$ with T fixed, as long as σ is bounded from below by a constant.

Remark 5. Bai (2009) also obtains an initial consistency result under weaker conditions on the errors than $||\varepsilon||_2 = \mathcal{O}_p(\sqrt{\min\{n,T\}})$. However, this result is obtained assuming that $R = R^0$, and the factors and loadings are independent of the errors. Neither of these are assumed in Proposition 1.

4 Asymptotic Expansion

Typically the asymptotic distribution of an extremum estimator is obtained by expanding the objective function locally around the true parameter value. It is, however, difficult to obtain an expansion of the objective function (2.15) since this involves a summation over a certain number of eigenvalues of a matrix. Following Bai (2009), an alternative approach would be to proceed from the first order conditions of the optimisation problem, to avoid dealing with the fully concentrated objective function. Yet Moon and Weidner (2015) show that it is possible to analyse this objective function directly, by utilising perturbation theory for linear operators to derive an expansion for the perturbed eigenvalues of $\boldsymbol{F}^0\tilde{\boldsymbol{\Lambda}}^{0\top}\tilde{\boldsymbol{\Lambda}}^0\boldsymbol{F}^{0\top}/nT$. Key to this approach is demonstrating that the perturbation is asymptotically small, which in this case follows from Proposition 1, whereby $|\theta^0_\kappa - \hat{\theta}_\kappa|$ is small, and from assuming that the 'perturbation' stemming from the error term, $\frac{||\tilde{\varepsilon}||_2}{\sqrt{nT}}$, diminishes asymptotically. In light of the discussion in the previous section, the significance of transforming the errors is again highlighted as expansion of the objective function remains valid only so long as $\frac{||\tilde{\boldsymbol{\varepsilon}}||_2}{\sqrt{nT}}$ is asymptotically small. Since $||\tilde{\boldsymbol{\varepsilon}}||_2 \leq ||\boldsymbol{\varepsilon}||_2$, $\frac{||\tilde{\boldsymbol{\varepsilon}}||_2}{\sqrt{nT}}$ will be asymptotically small in situations where this will not be true of $\frac{||\boldsymbol{\varepsilon}||_2}{\sqrt{nT}}$.[6]

Assumption AE (Asymptotic Expansion)**.**

(i) $R = R^0$.

(ii) $\frac{1}{n}\tilde{\boldsymbol{\Lambda}}^{0\top}\tilde{\boldsymbol{\Lambda}}^0 = \frac{1}{n}\boldsymbol{\Lambda}^{0\top}\boldsymbol{P}_{\mathcal{X}}\boldsymbol{\Lambda}^0 \xrightarrow{p} \boldsymbol{\Sigma}_{\tilde{\boldsymbol{\Lambda}}^0}$ as $n \to \infty$, with $\mu_{R^0}(\boldsymbol{\Sigma}_{\tilde{\boldsymbol{\Lambda}}^0}) > 0$ and $\mu_1(\boldsymbol{\Sigma}_{\tilde{\boldsymbol{\Lambda}}^0}) < \infty$.

(iii) $\frac{1}{T}\boldsymbol{F}^{0\top}\boldsymbol{F}^0 \xrightarrow{p} \boldsymbol{\Sigma}_{\boldsymbol{F}^0} > 0$ as $T \to \infty$, with $\mu_{R^0}(\boldsymbol{\Sigma}_{\boldsymbol{F}^0}) > 0$ and $\mu_1(\boldsymbol{\Sigma}_{\boldsymbol{F}^0}) < \infty$.

In the absence of dynamics, Moon and Weidner (2015) show that, under certain conditions, the asymptotic distribution of the PC estimator is unaffected by overstatement of the number of factors. Though it might also be expected that a similar result could be obtained in the present case, the asymptotic distribution is derived under the assumption that the number of factors is correctly specified; that is $R = R^0$ as in Assumption AE(i). A method to detect the true number of factors is discussed in Section 6.3. Assumptions AE(ii) and AE(iii) assume the factors and the transformed factor loadings are strong and both have a nonnegligible impact on the variance of the term $\tilde{\boldsymbol{\Lambda}}^0\boldsymbol{F}^{0\top} + \tilde{\boldsymbol{\varepsilon}}$.

Proposition 2 (Asymptotic Expansion)**.** *Under Assumptions MD, ER and AE, if* $||\boldsymbol{\theta}^0 - \boldsymbol{\theta}||_2 = \mathcal{O}_p(1)$*, then, as* $T^2/n \to 0$*,*

$$\mathscr{Q}(\boldsymbol{\theta}) = \mathscr{Q}(\boldsymbol{\theta}^0) - \frac{2}{\sqrt{nT}}(\boldsymbol{\theta} - \boldsymbol{\theta}^0)^\top\boldsymbol{d} + (\boldsymbol{\theta} - \boldsymbol{\theta}^0)^\top\boldsymbol{D}(\boldsymbol{\theta} - \boldsymbol{\theta}^0) + \boldsymbol{r}(\boldsymbol{\theta}), \tag{4.1}$$

where $\boldsymbol{d} := \boldsymbol{c} + \boldsymbol{b}^{(1)} + \boldsymbol{b}^{(2)} + \boldsymbol{b}^{(3)} + \boldsymbol{b}^{(4)}$*, the elements of these vectors and matrices are given by*

$$D_{\kappa\kappa'} := \frac{1}{nT}\mathrm{tr}(\tilde{\boldsymbol{Z}}_\kappa\boldsymbol{M}_{\boldsymbol{F}^0}\tilde{\boldsymbol{Z}}_{\kappa'}^\top\boldsymbol{M}_{\tilde{\boldsymbol{\Lambda}}^0}), \tag{4.2}$$

$$c_\kappa := \frac{1}{nT}\mathrm{tr}(\tilde{\boldsymbol{Z}}_\kappa\boldsymbol{M}_{\boldsymbol{F}^0}\tilde{\boldsymbol{\varepsilon}}^\top\boldsymbol{M}_{\tilde{\boldsymbol{\Lambda}}^0}), \tag{4.3}$$

$$b_1^{(1)} := \frac{1}{\sqrt{nT}}\mathrm{tr}\left(\boldsymbol{M}_{\boldsymbol{F}^0}\boldsymbol{G}\boldsymbol{M}_{\boldsymbol{F}^0}\tilde{\boldsymbol{\varepsilon}}^\top\boldsymbol{M}_{\tilde{\boldsymbol{\Lambda}}^0}\tilde{\boldsymbol{\varepsilon}}\right), \tag{4.4}$$

$$b_\kappa^{(2)} := -\frac{1}{\sqrt{nT}}\mathrm{tr}\left(\boldsymbol{M}_{\boldsymbol{F}^0}\tilde{\boldsymbol{\varepsilon}}^\top\boldsymbol{M}_{\tilde{\boldsymbol{\Lambda}}^0}\tilde{\boldsymbol{Z}}_\kappa\boldsymbol{F}^0(\boldsymbol{F}^{0\top}\boldsymbol{F}^0)^{-1}(\tilde{\boldsymbol{\Lambda}}^{0\top}\tilde{\boldsymbol{\Lambda}}^0)^{-1}\tilde{\boldsymbol{\Lambda}}^{0\top}\tilde{\boldsymbol{\varepsilon}}\right), \tag{4.5}$$

$$b_\kappa^{(3)} := -\frac{1}{\sqrt{nT}}\mathrm{tr}\left(\boldsymbol{M}_{\boldsymbol{F}^0}\tilde{\boldsymbol{Z}}_\kappa^\top\boldsymbol{M}_{\tilde{\boldsymbol{\Lambda}}^0}\tilde{\boldsymbol{\varepsilon}}\boldsymbol{F}^0(\boldsymbol{F}^{0\top}\boldsymbol{F}^0)^{-1}(\tilde{\boldsymbol{\Lambda}}^{0\top}\tilde{\boldsymbol{\Lambda}}^0)^{-1}\tilde{\boldsymbol{\Lambda}}^{0\top}\tilde{\boldsymbol{\varepsilon}}\right), \tag{4.6}$$

$$b_\kappa^{(4)} := -\frac{1}{\sqrt{nT}}\mathrm{tr}\left(\boldsymbol{M}_{\boldsymbol{F}^0}\tilde{\boldsymbol{\varepsilon}}^\top\boldsymbol{M}_{\tilde{\boldsymbol{\Lambda}}^0}\tilde{\boldsymbol{\varepsilon}}\boldsymbol{F}^0(\boldsymbol{F}^{0\top}\boldsymbol{F}^0)^{-1}(\tilde{\boldsymbol{\Lambda}}^{0\top}\tilde{\boldsymbol{\Lambda}}^0)^{-1}\tilde{\boldsymbol{\Lambda}}^{0\top}\tilde{\boldsymbol{Z}}_\kappa\right), \tag{4.7}$$

[6] The inequality $||\tilde{\boldsymbol{\varepsilon}}||_2 \leq ||\boldsymbol{\varepsilon}||_2$ is obtained by the submultiplicity of the spectral norm and noting that $||\boldsymbol{Q}_{\mathcal{X}}||_2 = 1$.

and $b_\kappa^{(1)} := 0$ *for* $\kappa = 2, ...K+1$. *Moreover, the order of the term* $\boldsymbol{r}(\boldsymbol{\theta})$ *is* $\boldsymbol{\mathcal{O}}_p\left(\frac{(1+\sqrt{nT}||\boldsymbol{\theta}^0-\boldsymbol{\theta}||_2)^2}{nT}\right)$.

As will be seen in the subsequent section, the term $\boldsymbol{c}$ plays a central role in determining the asymptotic distribution of the estimator. Term $\boldsymbol{b}^{(1)}$ arises due to the presence of a lagged outcome. When applying the PC estimator to the original model, an equivalent term arises and is the source of a bias, as described in Moon and Weidner (2017). Terms $\boldsymbol{b}^{(2)}, \boldsymbol{b}^{(3)}$ and $\boldsymbol{b}^{(4)}$ appear due to cross-sectional and serial dependence in the error term, and, again, have corresponding terms described in both Bai (2009) and Moon and Weidner (2015, 2017) which give rise to additional asymptotic biases. Under Assumptions MD, ER and AE, it can be established that $\boldsymbol{b}^{(1)}, \boldsymbol{b}^{(2)}, \boldsymbol{b}^{(3)}$ and $\boldsymbol{b}^{(4)}$ are $\boldsymbol{\mathcal{O}}_p(T^{1/5}/\sqrt{n})$ which suggests that the estimator is asymptotically unbiased where $T^3/n \to 0$. This is of course trivially satisfied where T is fixed. Using this, and the expression given in Proposition 2, the following result can be obtained.

Proposition 3 ($\sqrt{nT}$ Consistency)**.** *Under Assumptions MD, ER, CS and AE, and assuming that* $||\boldsymbol{c}||_2 = \mathcal{O}_p(1)$*, then, as* $T^3/n \to 0$*,*

$$\sqrt{nT}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0) = \boldsymbol{\mathcal{O}}_p(1). \tag{4.8}$$

Proposition 3 verifies the $\sqrt{n}$ consistency of the estimator in the event that T is fixed, and also that with the number of factors known, and n increasing sufficiently fast relative to T, $\sqrt{nT}$ consistency of the estimator can be obtained as both n and $T \to \infty$. The origin of the condition $T^3/n \to 0$ is explained more fully in the next section where it is shown to arise from the inclusion of a dynamic regressor and that, under stronger conditions, it is possible to reduce this to $T/n \to 0$.

5 Asymptotic Distribution

This section studies the asymptotic distribution of the estimator, culminating in the central result of this paper which establishes that, under more restrictive conditions, with T fixed, the estimator is asymptotically unbiased despite the presence of cross-sectional dependence, serial dependence, and with the inclusion of dynamic regressors. However, in order to really appreciate the impact that transforming the model through $\boldsymbol{Q}_{\mathfrak{X}}$ has, this section first presents a more general result which is derived under the assumption that $T/n \to c$ with $c \in [0, K^{-1}]$. This is useful to paint a more complete picture of the asymptotic properties of the estimator, and leads to a greater appreciation of the fundamental effects of transforming the model. To begin, some additional assumptions are introduced that are utilised in both cases.

Assumption ER* (Error)**.** The errors are generated as $\boldsymbol{\varepsilon} = \boldsymbol{\Sigma}_n^{\frac{1}{2}}\boldsymbol{U}\boldsymbol{\Sigma}_T^{\frac{1}{2}}$, where $\boldsymbol{\Sigma}_n^{\frac{1}{2}}$ and $\boldsymbol{\Sigma}_T^{\frac{1}{2}}$ are symmetric matrices of dimension $n \times n$ and $T \times T$, respectively, both of which are uniformly bounded in absolute row and column sums, and have eigenvalues uniformly bounded from below by a positive constant. $\boldsymbol{U}$ is an $n \times T$ matrix, with elements u_{it} which are independent of the exogenous covariates, the factors and the loadings, and identically and independently distributed across i and t, with $\mathbb{E}[u_{it}] = 0$, $\mathbb{E}[u_{it}^2] = 1$ and $\mathbb{E}[u_{it}^4] \leq C$.

Assumption ER* still allows for heteroskedasticity in both dimensions of the panel, but serves to place further limits on possible dependence in the errors across the cross-section and between time periods. Moreover, this assumption now imposes that the error terms are independent of the factors, the loadings and the covariates. One significant consequence of Assumption ER* is that $||\tilde{\boldsymbol{\varepsilon}}||_2 = \mathcal{O}_p(T^{\frac{3}{4}})$, which plays an important role in relaxing the requirement $T^2/n \to 0$ in Proposition 2.

The cross-sectional covariance matrix associated with the transformed error $\tilde{\boldsymbol{\varepsilon}}$ is $\boldsymbol{\Sigma}_n^{\frac{1}{2}}\boldsymbol{P}_{\boldsymbol{\mathcal{X}}}\boldsymbol{\Sigma}_n^{\frac{1}{2}}$. Therefore an additional assumption is required to manage the degree of dependence in this matrix.

Assumption AD (Asymptotic Distribution)**.** There exists an $n \times n$ nonstochastic matrix $\boldsymbol{\Pi}$ with elements π_{ij}, which is uniformly bounded in absolute row and column sums, and such that, for any $\boldsymbol{\mathcal{X}}$, $|(\Sigma_n^{\frac{1}{2}}P_{\mathcal{X}}\Sigma_n^{\frac{1}{2}})_{ij}| \leq |\pi_{ij}|$ for all i, j.

This assumption asserts that, irrespective of realisations of the covariates, the matrix $\boldsymbol{\Sigma}_n^{\frac{1}{2}}\boldsymbol{P}_{\boldsymbol{\mathcal{X}}}\boldsymbol{\Sigma}_n^{\frac{1}{2}}$ is dominated by a nonstochastic matrix which is uniformly bounded in absolute row and column sums.

5.1 Asymptotic Distribution: $T/n \to c \geq 0$

Where $T/n \to c > 0$, Assumption ER* alone will not be enough to apply Proposition 2 because in this case $||\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0||_2 = \mathcal{O}_p(1)$ does not follow from Proposition 1. In order to resolve this, the following additional assumption is imposed under which it is possible to obtain a faster rate of consistency.

Assumption BE (Bounded Elements)**.** The sum of the absolute value of the off-diagonal elements of $\boldsymbol{\Sigma}_T$ are bounded by a constant, that is, $\sum_{t=1}^{T}\sum_{\tau\neq t}^{T}|(\Sigma_T)_{t\tau}| < c_{\Sigma_T}$.

Assumption BE places further limits on the degree of inter-temporal dependence. The precise origin of this assumption will become clear shortly. With this assumption in place, the following faster rate of consistency can be obtained.

Proposition 4 (Consistency – Faster)**.** *Under Assumptions MD, CS, ER* and BE,*

$$||\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0||_2 = \mathcal{O}_p\left(\frac{T^{\frac{1}{4}}}{\sqrt{n}}\right). \tag{5.1}$$

Proposition 4 refines the rate of consistency given in Proposition 1, though does so under more stringent conditions. This result enables the derivation of Theorem 1 presented below.

Theorem 1 (Asymptotic Distribution)**.** *Under Assumptions MD, CS, AE, AD, ER*, and BE, as $T/n \to c$ with $c \in [0, K^{-1}]$,*

$$\sqrt{nT}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0) + \boldsymbol{\Delta}^{-1}(\boldsymbol{\psi}^{(0)} + \boldsymbol{\psi}^{(1)} + \boldsymbol{\psi}^{(2)} + \boldsymbol{\psi}^{(3)}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{\Delta}^{-1}(\boldsymbol{\Omega} + \boldsymbol{\Upsilon}^{(2)} + \boldsymbol{\Xi} + \bar{\boldsymbol{\Phi}})\boldsymbol{\Delta}^{-1}), \tag{5.2}$$

where,

$$\boldsymbol{\psi}^{(0)} := \frac{1}{\sqrt{nT}}\begin{pmatrix} \mathrm{tr}(\boldsymbol{M}_{\tilde{\boldsymbol{\Lambda}}^0}\tilde{\boldsymbol{\Sigma}}_n)\mathrm{tr}(\boldsymbol{G}\boldsymbol{\Sigma}_T) \\ \mathbf{0}_{K\times 1} \end{pmatrix}, \tag{5.3}$$

$$\boldsymbol{\psi}^{(1)} := \frac{1}{\sqrt{nT}}\begin{pmatrix} \mathrm{tr}(\tilde{\boldsymbol{\Sigma}}_n)(\mathrm{tr}(\boldsymbol{\Sigma}_T\boldsymbol{M}_{\boldsymbol{F}^0}\boldsymbol{G}\boldsymbol{P}_{\boldsymbol{F}^0}) + \mathrm{tr}(\boldsymbol{P}_{\boldsymbol{F}^0}\boldsymbol{\Sigma}_T\boldsymbol{G})) \\ \mathbf{0}_{K\times 1} \end{pmatrix}, \tag{5.4}$$

$$\psi^{(2)}_{\kappa} := \frac{1}{\sqrt{nT}}\mathrm{tr}(\boldsymbol{\Sigma}_T)\mathrm{tr}(\tilde{\boldsymbol{\Sigma}}_n\boldsymbol{M}_{\tilde{\boldsymbol{\Lambda}}^0}\tilde{\boldsymbol{Z}}_{\kappa}\boldsymbol{F}^0(\boldsymbol{F}^{0\top}\boldsymbol{F}^0)^{-1}(\tilde{\boldsymbol{\Lambda}}^{0\top}\tilde{\boldsymbol{\Lambda}}^{0})^{-1}\tilde{\boldsymbol{\Lambda}}^{0\top}), \tag{5.5}$$

$$\psi^{(3)}_{\kappa} := \frac{1}{\sqrt{nT}}\mathrm{tr}(\tilde{\boldsymbol{\Sigma}}_n)\mathrm{tr}(\boldsymbol{\Sigma}_T\boldsymbol{F}^0(\boldsymbol{F}^{0\top}\boldsymbol{F}^0)^{-1}(\tilde{\boldsymbol{\Lambda}}^{0\top}\tilde{\boldsymbol{\Lambda}}^{0})^{-1}\tilde{\boldsymbol{\Lambda}}^{0\top}\tilde{\boldsymbol{Z}}_{\kappa}\boldsymbol{M}_{\boldsymbol{F}^0}), \tag{5.6}$$

$$\boldsymbol{\Omega} := \frac{1}{nT}\tilde{\boldsymbol{\mathcal{Z}}}^{\top}(\boldsymbol{M}_{\boldsymbol{F}^0}\otimes\boldsymbol{M}_{\tilde{\boldsymbol{\Lambda}}^0})(\boldsymbol{\Sigma}_T\otimes\tilde{\boldsymbol{\Sigma}}_n)(\boldsymbol{M}_{\boldsymbol{F}^0}\otimes\boldsymbol{M}_{\tilde{\boldsymbol{\Lambda}}^0})\tilde{\boldsymbol{\mathcal{Z}}}, \tag{5.7}$$

$$\boldsymbol{\Upsilon}^{(1)} := \frac{1}{nT}\begin{pmatrix} \mathrm{tr}(\tilde{\boldsymbol{\Sigma}}_n)\mathrm{tr}(\boldsymbol{G}\boldsymbol{\Sigma}_T\boldsymbol{G}^{\top}) & \mathbf{0}_{1\times K} \\ \mathbf{0}_{K\times 1} & \mathbf{0}_{K\times K} \end{pmatrix}, \tag{5.8}$$

$$\boldsymbol{\Upsilon}^{(2)} := \frac{1}{nT}\frac{1}{2}\begin{pmatrix} \mathrm{tr}(\tilde{\boldsymbol{\Sigma}}_n\tilde{\boldsymbol{\Sigma}}_n)(\mathrm{tr}(\boldsymbol{G}\boldsymbol{\Sigma}_T\boldsymbol{G}\boldsymbol{\Sigma}_T) + 2\mathrm{tr}(\boldsymbol{G}\boldsymbol{\Sigma}_T\boldsymbol{G}^{\top}\boldsymbol{\Sigma}_T) + \mathrm{tr}(\boldsymbol{G}^{\top}\boldsymbol{\Sigma}_T\boldsymbol{G}^{\top}\boldsymbol{\Sigma}_T)) & \mathbf{0}_{1\times K} \\ \mathbf{0}_{K\times 1} & \mathbf{0}_{K\times K} \end{pmatrix}, \tag{5.9}$$

$$\boldsymbol{\Xi} := \frac{(\upsilon^{(4)} - 3)}{nT} \begin{pmatrix} \mathrm{tr}(\mathrm{diagv}(\boldsymbol{\Sigma}_T^{\frac{1}{2}}\boldsymbol{G}\boldsymbol{\Sigma}_T^{\frac{1}{2}})^\top \mathrm{diagv}(\boldsymbol{\Sigma}_T^{\frac{1}{2}}\boldsymbol{G}\boldsymbol{\Sigma}_T^{\frac{1}{2}}))\mathrm{tr}(\mathrm{diagv}(\boldsymbol{\Sigma}_n^{\frac{1}{2}}\boldsymbol{P}_{\mathcal{X}}\boldsymbol{\Sigma}_n^{\frac{1}{2}})^\top \mathrm{diagv}(\boldsymbol{\Sigma}_n^{\frac{1}{2}}\boldsymbol{P}_{\mathcal{X}}\boldsymbol{\Sigma}_n^{\frac{1}{2}})) & \mathbf{0}_{1\times K} \\ \mathbf{0}_{K\times 1} & \mathbf{0}_{K\times K} \end{pmatrix}, \tag{5.10}$$

$\tilde{\boldsymbol{\Sigma}}_n := \boldsymbol{Q}_{\mathcal{X}}^\top \boldsymbol{\Sigma}_n \boldsymbol{Q}_{\mathcal{X}}$, $\boldsymbol{\Delta} := \boldsymbol{D} + \boldsymbol{\Upsilon}^{(1)}$, $\bar{\boldsymbol{\Phi}} := \upsilon^{(3)}(\boldsymbol{\Phi} + \boldsymbol{\Phi}^\top)/nT$, $\boldsymbol{\Phi} := (\boldsymbol{\phi}, \mathbf{0}_{(K+1)\times K})$, with $\boldsymbol{\phi}$ being a $(K+1)\times 1$ vector with κ-th element

$$\phi_\kappa := \mathrm{vec}((\mathrm{diagv}(\boldsymbol{\Sigma}_T^{\frac{1}{2}}\boldsymbol{G}\boldsymbol{\Sigma}_T^{\frac{1}{2}}) \otimes \mathrm{diagv}(\boldsymbol{\Sigma}_n^{\frac{1}{2}}\boldsymbol{P}_{\mathcal{X}}\boldsymbol{\Sigma}_n^{\frac{1}{2}})))^\top \mathrm{vec}(\boldsymbol{\Sigma}_n^{\frac{1}{2}}\boldsymbol{Q}\boldsymbol{M}_{\tilde{\boldsymbol{\Lambda}}^0}\tilde{\boldsymbol{Z}}_\kappa \boldsymbol{M}_{\boldsymbol{F}^0}\boldsymbol{\Sigma}_T^{\frac{1}{2}}), \tag{5.11}$$

$\upsilon^{(3)}$ and $\upsilon^{(4)}$ being the third and fourth moments of u_{it}, respectively, and where $\boldsymbol{D}$ is defined in Proposition 2.

What is perhaps most significant in Theorem 1 are the four bias terms $\boldsymbol{\psi}^{(0)}, \boldsymbol{\psi}^{(1)}, \boldsymbol{\psi}^{(2)}$ and $\boldsymbol{\psi}^{(3)}$. The first two of these, $\boldsymbol{\psi}^{(0)}$ and $\boldsymbol{\psi}^{(1)}$, arise due to the presence of the dynamic regressor and reflect the complex temporal interaction, represented by $\boldsymbol{G}$, that occurs with the covariance matrices $\boldsymbol{\Sigma}_T$ and $\tilde{\boldsymbol{\Sigma}}_n$, the factors and the loadings. Terms $\boldsymbol{\psi}^{(2)}$ and $\boldsymbol{\psi}^{(3)}$ arise due to cross-sectional and time series dependence and, as is discussed shortly, are notably absent in the case of identically and independently distributed errors. Studying all these terms more closely shows the order of these biases to be:[7]

$$\boldsymbol{\psi}^{(0)} = \boldsymbol{\mathcal{O}}_p\left(\sqrt{\frac{T}{n}}\right), \tag{5.12}$$

$$\boldsymbol{\psi}^{(1)} = \boldsymbol{\mathcal{O}}_p\left(\sqrt{\frac{T}{n}}\right), \tag{5.13}$$

$$\boldsymbol{\psi}^{(2)} = \boldsymbol{\mathcal{O}}_p\left(\sqrt{\frac{T}{n}}\right), \tag{5.14}$$

$$\boldsymbol{\psi}^{(3)} = \boldsymbol{\mathcal{O}}_p\left(\sqrt{\frac{T}{n}}\right), \tag{5.15}$$

which reveals something fundamental: projection of the entire model into the time dimension of the panel does not make the incidental problem in the cross-section disappear entirely, it instead shifts it into the time dimension, where it may interact with the extant problem in that dimension in complicated ways.

Theorem 1 also reveals the origin of the requirement $T^3/n \to 0$ in Proposition 3, as well as the rationale behind Assumption BE, both of which are a direct consequence of the

[7] Although random, strictly speaking the elements of the vectors $\boldsymbol{\psi}^{(0)}$ and $\boldsymbol{\psi}^{(1)}$ are bounded by constants and not just in probability.

order of the bias $\boldsymbol{\psi}^{(0)}$. For the first component of $\boldsymbol{\psi}^{(0)}$, it can be established that, under Assumptions AE and ER*,

$$aT - b \leq \mathrm{tr}(\boldsymbol{M}_{\tilde{\boldsymbol{\Lambda}}^0}\tilde{\boldsymbol{\Sigma}}_n), \tag{5.16}$$

where a and b are positive constants.[8] The second component, $\mathrm{tr}(\boldsymbol{G}\boldsymbol{\Sigma}_T)$, is a weighted summation of the lower triangular elements of $\boldsymbol{\Sigma}_T$, and, under Assumptions MD ER*, it can be established that

$$|\mathrm{tr}(\boldsymbol{G}\boldsymbol{\Sigma}_T)| \leq T||\boldsymbol{G}||_2||\boldsymbol{\Sigma}_T||_2 = \mathcal{O}_p(T). \tag{5.17}$$

Indeed, without further restrictions, it is possible for this term to be exactly of that order, in which case

$$\boldsymbol{\psi}^{(0)} = \mathcal{O}_p\left(\frac{T^{1.5}}{\sqrt{n}}\right). \tag{5.18}$$

This lays bare the origin of the requirement $T^3/n \to 0$ in Proposition 3. Moreover, (5.18) also suggests that, without further restrictions, rates of consistency faster than $\sqrt{T/n}$ should not be expected, even where the number of factors is known. The addition of Assumption BE ensures $\mathrm{tr}(\boldsymbol{G}\boldsymbol{\Sigma}_T) = \mathcal{O}_p(1)$, which reduces the order of $\boldsymbol{\psi}^{(0)}$ to that given in (5.12). The following subsections examine more closely 4 special cases of Theorem 1 that are of particular interest.

5.1.1 Diagonal $\boldsymbol{\Sigma}_T$

A special case of Assumption BE is where the off-diagonal elements of $\boldsymbol{\Sigma}_T$ are exactly zero, in which case the following corollary is obtained.

Corollary 1 (Diagonal $\boldsymbol{\Sigma}_T$)**.** *Assume $\boldsymbol{\Sigma}_T$ is a positive diagonal matrix with bounded entries. Then, under the assumptions of Theorem 1, as $T/n \to c$ with $c \in [0, K^{-1}]$,*

$$\sqrt{nT}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0) + \boldsymbol{\Delta}^{-1}(\boldsymbol{\psi}^{(1)} + \boldsymbol{\psi}^{(2)} + \boldsymbol{\psi}^{(3)}) \xrightarrow{d} \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Delta}^{-1}(\boldsymbol{\Omega} + \boldsymbol{\Upsilon}^{(2)})\boldsymbol{\Delta}^{-1}), \tag{5.19}$$

where the nonzero element in $\boldsymbol{\Upsilon}^{(2)}$ simplifies to $2\mathrm{tr}(\tilde{\boldsymbol{\Sigma}}_n\tilde{\boldsymbol{\Sigma}}_n)\mathrm{tr}(\boldsymbol{G}\boldsymbol{\Sigma}_T\boldsymbol{\Sigma}_T\boldsymbol{G}^\top)$.

When $\boldsymbol{\Sigma}_T$ is diagonal, the variance-covariance matrix is greatly simplified and the bias term $\boldsymbol{\psi}^{(0)}$ is identically zero.[9] As a consequence, the estimator is $\sqrt{nT}$ consistent as $T/n \to c$

[8] See Appendix A.2.

[9] This can be seen by noticing that $\mathrm{tr}(\boldsymbol{G}\boldsymbol{\Sigma}_T) = \sum_{t=1}^{T-1}\sum_{\tau=1}^{t}\alpha^{\tau-1}(\boldsymbol{\Sigma}_T)_{(t+1,t+1-\tau)}$, which is a summation over the lower triangular elements of $\boldsymbol{\Sigma}_T$. Therefore, as long as $\boldsymbol{\Sigma}_T$ is upper triangular, or indeed diagonal, this will be exactly zero.

with $c \in [0, K^{-1}]$, and is in fact always at least $\sqrt{n}$ consistent irrespective of T. Moreover, where $T/n \to 0$, the estimator is asymptotically unbiased, though, in the event that $T/n \to c > 0$, a bias of order $\sqrt{T/n}$ is present.

5.1.2 IID Errors

Another corollary of Theorem 1 is obtained when the errors are identically and independently distributed. In this case $\boldsymbol{\psi}^{(2)} = \boldsymbol{\psi}^{(3)} = \mathbf{0}_{(K+1)\times 1}$ because where $\boldsymbol{\Sigma}_T \propto \boldsymbol{I}_T$, the orthogonal projector $\boldsymbol{M}_{\boldsymbol{F}^0}$ directly multiplies with $\boldsymbol{F}^0$ and, similarly, when $\boldsymbol{\Sigma}_n \propto \boldsymbol{I}_n$, $\boldsymbol{M}_{\tilde{\boldsymbol{\Lambda}}^{\mathbf{0}}}$ directly multiples with $\tilde{\boldsymbol{\Lambda}}^0$.

Corollary 2 (IID)**.** *Assume $\boldsymbol{\Sigma}_n = \sigma_0^2 \boldsymbol{I}_n$ and $\boldsymbol{\Sigma}_T = \sigma_0^2 \boldsymbol{I}_T$. Then, under the assumptions of Theorem 1, as $T/n \to c$ with $c \in [0, K^{-1}]$,*

$$\sqrt{nT}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0) + \boldsymbol{\Delta}^{-1}\boldsymbol{\psi}^{(1)} \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{\Delta}^{-1}(\boldsymbol{\Omega} + \boldsymbol{\Upsilon}^{(2)})\boldsymbol{\Delta}^{-1}), \tag{5.20}$$

where the nonzero element in $\boldsymbol{\Upsilon}^{(2)}$ simplifies to $2TK\sigma_0^4 \mathrm{tr}(\boldsymbol{G}\boldsymbol{G}^\top)$.

In this exceptional case the only remaining bias $\boldsymbol{\psi}^{(1)}$ reduces to

$$\psi_\sigma^{(1)} := \frac{\sigma_0^2}{\sqrt{nT}} \mathrm{tr}(\boldsymbol{P}_{\mathfrak{X}}) \mathrm{tr}(\boldsymbol{G}\boldsymbol{P}_{\boldsymbol{F}^0}). \tag{5.21}$$

This, in fact, is a generalisation of the bias described in (Nickell, 1981). To see this, recall that the model of individual effects nests as a special case of interactive fixed effects in which a single heterogeneous loading vector is interacted with a constant factor:

$$\boldsymbol{\lambda}_{IFE}^0 := \begin{pmatrix} \lambda_1^0 \\ \vdots \\ \lambda_n^0 \end{pmatrix}, \ \boldsymbol{F}_{IFE}^0 := \boldsymbol{\iota}_T, \tag{5.22}$$

where $\boldsymbol{\iota}_T$ is a $T \times 1$ vector of ones. In this case $\boldsymbol{P}_{\boldsymbol{F}_{IFE}^0} = \frac{1}{T}\boldsymbol{\iota}_T\boldsymbol{\iota}_T^\top$, and therefore,

$$\psi_\sigma^{(2)} = \frac{\sigma_0^2}{\sqrt{nT}} \frac{1}{T} \mathrm{tr}(\boldsymbol{P}_{\mathfrak{X}}) \mathrm{tr}(\boldsymbol{G}\boldsymbol{\iota}_T\boldsymbol{\iota}_T^\top). \tag{5.23}$$

Notice the structure of $\boldsymbol{G}$:

$$\boldsymbol{G}\boldsymbol{\iota}_T\boldsymbol{\iota}_T^\top = \begin{pmatrix} \vdots & \vdots & \vdots \\ 1 + \alpha^0 + (\alpha^0)^2 & 1 + \alpha^0 + (\alpha^0)^2 & 1 + \alpha^0 + (\alpha^0)^2 \\ 1 + \alpha^0 & 1 + \alpha^0 & 1 + \alpha^0 \\ 1 & 1 & 1 \\ 0 & 0 & 0 \end{pmatrix}, \tag{5.24}$$

and so $\operatorname{tr}(\boldsymbol{G}\boldsymbol{\iota}_T\boldsymbol{\iota}_T^\top) = \sum_{t=1}^{T-1}\sum_{\tau=1}^{t}(\alpha^0)^{\tau-1}$. A bit of algebra reveals that[10]

$$\sum_{t=1}^{T-1}\sum_{\tau=1}^{t}(\alpha^0)^{\tau-1} = \frac{T}{(1-\alpha^0)}\left(1 - \frac{1}{T}\frac{(1-(\alpha^0)^T)}{1-\alpha^0}\right). \tag{5.25}$$

Now, since the trace of a projector is equal to its rank, $\operatorname{tr}(\boldsymbol{P}_{\mathcal{X}}) = TK$, and the following expression is obtained:

$$\psi_\sigma^{(1)} = \sqrt{\frac{T}{n}}\frac{K}{(1-\alpha)}\left(1 - \frac{1}{T}\frac{(1-\alpha^T)}{1-\alpha}\right). \tag{5.26}$$

This reveals a simple expression for the bias. Notice the significance of the transformation $\boldsymbol{Q}_{\mathcal{X}}$: it reduces the rank of the cross-sectional covariance matrix to TK. Without the transformation $\operatorname{tr}(\tilde{\boldsymbol{\Sigma}}_n) = \operatorname{tr}(\boldsymbol{\Sigma}_n) = \operatorname{tr}(\boldsymbol{I}_n) = n$, and so

$$\psi_\sigma^{(1)} = \sqrt{\frac{n}{T}}\frac{1}{(1-\alpha)}\left(1 - \frac{1}{T}\frac{(1-\alpha^T)}{1-\alpha}\right), \tag{5.27}$$

which matches (up to scale by $\sqrt{nT}$) exactly expression (27) derived in Nickell (1981). This again highlights the fact that transforming the model by $\boldsymbol{Q}_{\mathcal{X}}$ does not eliminate all traces of the incidental parameter problem that would have existed in the cross-section. It simply transfers it to the time dimension where, as exemplified by comparing (5.26) and (5.27), it will likely manifest itself in similar ways.

5.1.3 Standard Normality

As would be expected, under standard normality of the errors particularly simple expressions are obtained.

Assumption SN (Standard Normality)**.** Conditional on the covariates, the factors and the loadings, the elements of the $n \times T$ matrix $\boldsymbol{\varepsilon}$ are drawn from independent standard normal distributions.

Corollary 3 (Standard Normality)**.** *Under Assumptions MD, CS, AE, AD, and SN, as $T/n \to c$ with $c \in [0, K^{-1}]$,*

$$\sqrt{nT}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0) - \boldsymbol{\Delta}^{-1}\boldsymbol{\psi}_{SN} \xrightarrow{d} \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Delta}^{-1}), \tag{5.28}$$

where

$$\boldsymbol{\psi}_{SN} \coloneqq K\sqrt{\frac{T}{n}}\begin{pmatrix}\operatorname{tr}(\boldsymbol{G}\boldsymbol{P}_{\boldsymbol{F}^0})\\ \boldsymbol{0}_{K\times 1}\end{pmatrix}, \tag{5.29}$$

and $\boldsymbol{\Delta}$ is defined in Theorem 1.

[10] See Appendix A.3.

Following the same steps used to obtain (5.26), a simple expression for $\boldsymbol{\psi}_{SN}$ can be derived under individual effects. In general, however, using the inequality $\text{tr}(\boldsymbol{A}) \leq \text{rank}(\boldsymbol{A})||\boldsymbol{A}||_1$, and noticing that $||\boldsymbol{G}||_1 = \sum_{t=1}^{T-1}(\alpha^0)^{t-1}$, one obtains the following bound:

$$\sqrt{\frac{T}{n}}K|\text{tr}(\boldsymbol{G}\boldsymbol{P}_{\boldsymbol{F}^0})| \leq \sqrt{\frac{T}{n}}KR^0\frac{(1-(\alpha^0)^T)}{1-\alpha^0}, \tag{5.30}$$

and hence the bias remains comparable, even where the structure of $\boldsymbol{P}_{\boldsymbol{F}^0}$ is left unrestricted.

5.1.4 Static Model

Corollary 4 (Static Model)**.** *In the absence of a dynamic regressor, then, under the assumptions of Theorem 1, as $T/n \to c$ with $c \in [0, K^{-1}]$,*

$$\sqrt{nT}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0) + \boldsymbol{D}^{-1}(\boldsymbol{\psi}^{(2)} + \boldsymbol{\psi}^{(3)}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{D}^{-1}\boldsymbol{\Omega}\boldsymbol{D}^{-1}). \tag{5.31}$$

This result is the analogue of Theorem 3 in Bai (2009) with the important difference being the order of the bias terms, which in this case are both of order $\sqrt{T/n}$. Moreover, similarly to Corollary 1, this result implies that, in the absence of dynamics, the estimator must be at least $\sqrt{n}$ consistent irrespective of T.

5.2 Asymptotic Distribution: Fixed T

This section presents the main result of this paper in the form of the following theorem.

Theorem 2 (Fixed T)**.** *Under Assumptions MD, CS, AE, AD, and ER*, with T fixed and $n \to \infty$,*

$$\sqrt{nT}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{D}^{-1}\boldsymbol{\Omega}\boldsymbol{D}^{-1}). \tag{5.32}$$

Theorem 2 demonstrates that with T fixed, the estimator is asymptotically unbiased in the presence of cross-sectional dependence, serial dependence, and with the inclusion of dynamic regressors. Notice also that Assumption BE is not required to obtain this result because the bias $\boldsymbol{\psi}^{(0)} = \mathcal{O}_p(1)$ when T is fixed.

Remark 6. Although the result for the static model is presented as a corollary of Theorem 1, where there is are no lagged outcome $\boldsymbol{\psi}^{(0)}$ does not appear, and so this result can be obtained without requiring Assumption BE.

Remark 7. While the biases $\boldsymbol{\psi}^{(1)}, \boldsymbol{\psi}^{(2)}$ and $\boldsymbol{\psi}^{(3)}$ have analogues described in Bai (2009), Moon and Weidner (2015), and Moon and Weidner (2017), $\boldsymbol{\psi}^{(0)}$ does not. That is because the former two of these papers study the static model, while the latter derives the asymptotic distribution under the assumption that the errors are independent across time, whereby $\boldsymbol{\Sigma}_T$ is diagonal, and $\boldsymbol{\psi}^{(0)}$ does not appear.

6 Further Matter

6.1 Faster Rates of Consistency with $R \geq R^0$.

Some of the situations discussed in Section 5 also give rise to faster rates of consistency with $R \geq R^0$. This section serves to highlight a few of these cases.

Proposition 5 (Consistency – Static Model)**.** *In the absence of a dynamic regressor, under Assumptions MD, CS, and ER*,*

$$||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0||_2 = \mathcal{O}_p\left(\frac{T^{\frac{1}{4}}}{\sqrt{n}}\right). \tag{6.1}$$

At the core of this result is the fact that under Assumption ER* $||\tilde{\boldsymbol{\varepsilon}}||_2 = \mathcal{O}_p(T^{\frac{3}{4}})$. Because, trivially, it is also the case that $||\tilde{\boldsymbol{\varepsilon}}||_2 \leq ||\boldsymbol{\varepsilon}||_2$, the estimator must also at least be consistent at the rate $\frac{1}{\sqrt{T}}$, which is straightforward to obtain following Theorem 4.1 in Moon and Weidner (2015).

Owing to the fact that the standard normal distribution is invariant to orthogonal transformations, especially favourable rates of consistency can be achieved in this case. Key to showing this is the following result.

Proposition 6. *Under Assumption SN,* $||\tilde{\boldsymbol{\varepsilon}}||_2 = \mathcal{O}_p(\sqrt{T})$.

Proof: Since the normal distribution is invariant to orthogonal transformation, it follows that $\boldsymbol{Q}_{\mathcal{X}}^{\top}\boldsymbol{\varepsilon}$ is a $TK \times T$ matrix with independent standard normal entries. Latala (2005) shows that such a matrix will be $\mathcal{O}_p(\max\{\sqrt{TK}, \sqrt{T}\}) = \mathcal{O}_p(\sqrt{T})$. □

Using Proposition 6, a faster rate of consistency can thus be obtained.

Proposition 7 (Consistency – Standard Normality)**.** *Under Assumptions MD, CS, and SN,*

$$||\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0||_2 = \mathcal{O}_p\left(\frac{1}{\sqrt{n}}\right). \tag{6.2}$$

This result demonstrates that, under standard normality, and with $R \geq R^0$, the rate of consistency is in fact independent of T.

6.2 Low Rank Covariates

Low rank covariates often appear in applied work, with obvious examples being those that are either time or cross-sectionally invariant. In models with interactive effects, identifying the coefficients associated with these covariates can be challenging since they present another low rank structure in the model, in addition to the factor term. Mirroring the result obtained in Moon and Weidner (2017), where such covariates are present it is, however, still possible to obtain consistent estimates under appropriate conditions. Let $\boldsymbol{\vartheta}$ denote a reordering of the parameter vector $\boldsymbol{\theta}$ such that the first K_{L} coefficients, indexed $l = 1, ..., K_{\mathrm{L}}$, are those associated with low rank regressors, and the remaining K_{H} coefficients, indexed $h = 1, ..., K_{\mathrm{H}}$, denote those associated with the regressors which have full rank. For simplicity it is assumed that the low rank regressors have rank 1, though the following results extend naturally to the more general case. The l-th low rank covariate can be decomposed as $\boldsymbol{X}_l = \boldsymbol{v}_l \boldsymbol{w}_l^\top$, with $\boldsymbol{v}_l$ and $\boldsymbol{w}_l$ being $n \times 1$ and $T \times 1$ vectors, respectively. These vectors can then be gathered into the matrices $\boldsymbol{\mathcal{V}} := (\boldsymbol{v}_1, ..., \boldsymbol{v}_{K_{\mathrm{L}}})$ and $\boldsymbol{\mathcal{W}} := (\boldsymbol{w}_1, ..., \boldsymbol{w}_{K_{\mathrm{L}}})$. When some of the covariates are low rank, special care must be taken in the construction of $\boldsymbol{\mathcal{X}}$. In this case $\boldsymbol{\mathcal{X}}$ can be constructed as $(\boldsymbol{\mathcal{V}}, \boldsymbol{X}_1, ..., \boldsymbol{X}_{K_{\mathrm{H}}})$ to ensure that $\boldsymbol{\mathcal{X}}^\top \boldsymbol{\mathcal{X}}$ is invertible. Let $\tilde{\boldsymbol{\mathcal{V}}} = \boldsymbol{Q}_{\boldsymbol{\mathcal{X}}}^\top \boldsymbol{\mathcal{V}}$ and $\boldsymbol{\delta}_{\mathrm{H}} \cdot \tilde{\boldsymbol{Z}}_{\mathrm{H}} := \sum_{\kappa=1}^{K_{\mathrm{H}}} \delta_\kappa \tilde{\boldsymbol{Z}}_\kappa$.

Assumption LR (Low Rank)**.**

(i) $\min_{\boldsymbol{\delta}_{\mathrm{H}} \in \mathbb{R}^{K_{\mathrm{H}}} : ||\boldsymbol{\delta}_{\mathrm{H}}||_2 = 1} \sum_{r=R+R^0+K_{\mathrm{L}}+1}^{T} \mu_r \left(\frac{1}{nT} (\boldsymbol{\delta}_{\mathrm{H}} \cdot \tilde{\boldsymbol{Z}}_{\mathrm{H}})^\top (\boldsymbol{\delta}_{\mathrm{H}} \cdot \tilde{\boldsymbol{Z}}_{\mathrm{H}}) \right) \geq b > 0.$

(ii) There exists a constant $c > 0$ such that $\frac{1}{n} \tilde{\boldsymbol{\Lambda}}^{0\top} \boldsymbol{M}_{\tilde{\boldsymbol{\mathcal{V}}}} \tilde{\boldsymbol{\Lambda}}^0 > c\boldsymbol{I}_{R^0}$ and $\frac{1}{T} \boldsymbol{F}^{0\top} \boldsymbol{M}_{\boldsymbol{\mathcal{W}}} \boldsymbol{F}^0 > c\boldsymbol{I}_{R^0}$, w.p.a.1.

Assumption LR is analogous to Assumption 4(ii) in Moon and Weidner (2017) and requires what amounts to a strengthening of Assumption CS(ii), and an additional condition to ensure that the low rank regressors are sufficiently distinct from the factors and the transformed loadings so as to be able to distinguish one from the other. Here however, special care must be taken with Assumption LR(ii) because

$$\frac{1}{n} \tilde{\boldsymbol{\Lambda}}^{0\top} \boldsymbol{M}_{\tilde{\boldsymbol{\mathcal{V}}}} \tilde{\boldsymbol{\Lambda}}^0 = \frac{1}{n} \boldsymbol{\Lambda}^{0\top} (\boldsymbol{P}_{\boldsymbol{\mathcal{X}}} - \boldsymbol{P}_{\boldsymbol{\mathcal{V}}}) \boldsymbol{\Lambda}^0. \tag{6.3}$$

Since the transforming the model by $\boldsymbol{Q}_{\boldsymbol{\mathcal{X}}}^\top$ has the effect of projecting the model into the column space of the covariates, it is not enough that the loadings be distinct from each $\boldsymbol{v}_l$, as in Moon and Weidner (2017). In this context what is required is that the projection of the loadings onto the column space of the all the covariates is different from the projection onto the column space of just the low rank covariates, which, clearly, will require there to be some high rank model covariates.

Proposition 8 (Consistency – Low Rank)**.** *Under Assumptions MD, AE, ER, and LR,* [11]

$$||\hat{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta}^0||_2 = \mathcal{O}_p\left(\sqrt{\frac{T}{n}}\right). \tag{6.4}$$

6.3 Estimating the Number of Factors

Results established in Sections 3 and 6.1 demonstrate that in many instances the estimator will remain consistent with the number of factors overestimated. However, since overestimation of the number of factors will typically lead to a loss of efficiency in finite samples, it is desirable to input the correct number of factors. One approach to detecting this number involves first estimating the coefficients with the number of factors overestimated, and using these estimates to construct a pure factor model. Then, methods devised to detect the number of factors in a pure factor model can be applied. Examples of these detection methods include Bai (2003), Onatski (2009) and Ahn and Horenstein (2013). This section focuses on one of these, the eigenvalue ratio test of Ahn and Horenstein (2013), and considers how, after transforming the model, this method can be applied to detect the number of factors with T fixed.

More generally, however, this section seeks to make two points. First, after having transformed the model, other results which exist in the literature for the large n, large T setting may be ported to that with T fixed, potentially with the additional benefit of relaxing assumptions regarding dependence in the errors. Second, in situations where factors exist in the error term which are uncorrelated with the covariates, alongside those which are correlated, transforming the model and detecting the number of factors may lead to efficiency gains, since only the number of factors which are correlated with the error term need be inputted into the estimation procedure. Let

$$\mu_r^* := \mu_r\left(\frac{1}{nT}\left(\tilde{\boldsymbol{Y}}\boldsymbol{S}(\hat{\alpha}) - \sum_{\kappa=1}^{K}\hat{\beta}_\kappa\tilde{\boldsymbol{X}}_\kappa\right)^\top\left(\tilde{\boldsymbol{Y}}\boldsymbol{S}(\hat{\alpha}) - \sum_{\kappa=1}^{K}\hat{\beta}_\kappa\tilde{\boldsymbol{X}}_\kappa\right) + \frac{1}{n}\boldsymbol{I}_T\right), \tag{6.5}$$

that is, μ_r^* is the r-th largest eigenvalue of the right-hand side matrix. Then define

$$\text{EigR}(r) := \frac{\mu_r^*}{\mu_{r+1}^*} \text{ for } r = 1, ..., T-1. \tag{6.6}$$

The main modification here from Ahn and Horenstein (2013)'s original specification is the addition of the matrix $\frac{1}{n}\boldsymbol{I}_T$. This is added because, unlike in the original setting, where

[11] Faster rates will also be possible in the circumstances discussed in Section 6.1.

covariates are present the eigenvalues in (6.5) need not be strictly positive, and there is a non-zero probability that some of them are exactly zero. The addition of the identity matrix ensures that this cannot happen and, because the eigenvalues are demonstrated to converge at a rate of $1/n$ or slower, this does not impact the properties of the test.

Proposition 9. *Under Assumptions MD, CS and ER*, as $T/n \to 0$,*

$$\Pr\left(\max_{1\leq r\leq T} \mu_r^* = R^0\right) \to 1. \tag{6.7}$$

6.4 Balestra and Nerlove's Approach

As mentioned previously, though often still yielding consistent estimates, using the PC estimator with the number of factors inputted R exceeding the true number of factors R^0 will result in a loss of efficiency in finite samples. The estimator $\hat{\boldsymbol{\theta}}$ studied thus far in this paper treats the initial condition $\boldsymbol{y}_0\boldsymbol{s}^\top(\alpha)$ as an additional parameter and, as a consequence, results in another factor appearing in the error term. An alternative approach which does not generate this additional factor is to follow Balestra and Nerlove (1966) and include the projection of the lagged outcome onto the column space of the exogenous variables as an additional explanatory variable on the right hand-side of the outcome equation. This approach is naturally wedded to this paper's, since projecting lagged outcomes embeds them in the TK-dimensional space spanned by the columns of $\boldsymbol{\mathfrak{X}}$. Consider the following outcome equation:

$$\boldsymbol{Y}^c = \alpha \boldsymbol{Y}_L^c + \sum_{k=1}^K \beta_k \boldsymbol{X}_k^c + \boldsymbol{\Lambda}^* \boldsymbol{F}^{*c\top} + \boldsymbol{\varepsilon}^c, \tag{6.8}$$

where $\boldsymbol{Y}_L^c := (\boldsymbol{y}_1, ..., \boldsymbol{y}_{T-1})$, and the matrices $\boldsymbol{Y}^c, \boldsymbol{X}_k^c, \boldsymbol{\Lambda}^*\boldsymbol{F}^{*c\top}$ and $\boldsymbol{\varepsilon}^c$ are $n \times T^c$, with $T^c := T - 1$. Clearly the trade off in adopting this approach is that, while no longer generating an additional factor, this does lead to the loss of a time period of data. Define $\boldsymbol{\mathfrak{X}}^c := (\boldsymbol{X}_1^c, ..., \boldsymbol{X}_K^c)$ and $\boldsymbol{Q}_{\boldsymbol{\mathfrak{X}}^c} := \boldsymbol{\mathfrak{X}}^c(\boldsymbol{\mathfrak{X}}^{c\top}\boldsymbol{\mathfrak{X}}^c)^{-1}$. Then, using $\sim$ to indicate transformed variables, as previously, consider the alternate objective function

$$Q^c(\boldsymbol{\theta}) := \frac{1}{nT^c}\operatorname{tr}\left(\left(\tilde{\boldsymbol{Y}}^c - \sum_{\kappa=1}^{K+1} \theta_\kappa \tilde{\boldsymbol{Z}}_\kappa^c - \tilde{\boldsymbol{\Lambda}}^* \boldsymbol{F}^{*c\top}\right)^\top \left(\tilde{\boldsymbol{Y}}^c - \sum_{\kappa=1}^{K+1} \theta_\kappa \tilde{\boldsymbol{Z}}_\kappa^c - \tilde{\boldsymbol{\Lambda}}^* \boldsymbol{F}^{*c\top}\right)\right), \tag{6.9}$$

where $\tilde{\boldsymbol{Z}}_1^c := \boldsymbol{Q}_{\boldsymbol{\mathfrak{X}}^c}^\top \boldsymbol{Y}_L^c$ and $\tilde{\boldsymbol{Z}}_\kappa^c := \tilde{\boldsymbol{X}}_\kappa^c$ for $\kappa = 2, ..., K+1$. An alternative estimator $\hat{\boldsymbol{\theta}}_{BN}$ may then be defined as

$$\hat{\boldsymbol{\theta}}_{BN} := \arg\min_{\boldsymbol{\theta}\in\Theta} Q^c(\boldsymbol{\theta}). \tag{6.10}$$

This estimator retains all of the essential properties of the $\hat{\boldsymbol{\theta}}$, including fixed T consistency and an analogous asymptotic distribution. Moreover, this approach is especially appealing since it involves simply transforming the data by $\boldsymbol{Q}_{\mathfrak{X}}$ and then applying the usual PC estimator with no other modifications.

Remark 8. The estimation approach proposed in this paper shares a close kinship with the procedure suggested by Chamberlain (1984) for short panels with individual effects. In the present context, this could be understood as decomposing $\boldsymbol{\Lambda} = \boldsymbol{P}_{\mathfrak{X}}\boldsymbol{\Lambda} + \boldsymbol{M}_{\mathfrak{X}}\boldsymbol{\Lambda} =: \mathfrak{X}\boldsymbol{\Gamma} + \boldsymbol{e}$, where $\boldsymbol{\Gamma}$ is a $TK \times T$ parameter to be estimated, and $\boldsymbol{e}$ is subsumed into the error term. Chamberlain (1984) suggests a minimum distance approach to jointly estimate $\boldsymbol{\theta}$ and $\boldsymbol{\Gamma}$, however, if one instead applies least squares, then concentrating out $\boldsymbol{\Gamma}$ and minimising with respect to the factors will yield an identical estimator.

7 Monte Carlo Simulations

This section provides simulation results which highlight the different properties of the PC estimator when applied to the original and transformed models. In the following design the factors and loadings are both generated independently from standard normal distributions and the true number of factors is set equal to 2; i.e. $R^0 = 2$. Two covariates are generated: $\boldsymbol{X}_1 = \boldsymbol{\Lambda}\boldsymbol{F}^\top + \boldsymbol{\eta}$, where $\boldsymbol{\eta}$ has elements drawn independently from a standard normal distribution, and $\boldsymbol{X}_2$, which is also drawn from a standard normal. The entries of the error $\boldsymbol{\varepsilon}$ are generated as $\boldsymbol{\Sigma}_n^{\frac{1}{2}}\boldsymbol{U}\boldsymbol{\Sigma}_T^{\frac{1}{2}}$, where the elements of $\boldsymbol{U}$ are independently drawn from a standard normal distribution, and $\boldsymbol{\Sigma}_n$ and $\boldsymbol{\Sigma}_T$ are diagonal matrices with elements drawn uniformly between 0.5 and 2.5. The number of Monte Carlo replications is 10000. Tables 1a – 1c display the bias and the standard error of the standard least squares estimator (LS), the principal component estimator applied to the original model (PC), the approach described in Section 6.4 (BN) and the PC estimator applied to the transformed model (QPC).

Table 1a: Bias (SE) α

	LS			PC			BN			QPC		
$n \setminus T$	6	9	12	6	9	12	6	9	12	6	9	12
30	**-0.003**	**-0.002**	**-0.001**	**-0.106**	**-0.005**	**-0.002**	**-0.043**	**-0.004**	**-0.003**	**-0.105**	**-0.007**	**-0.002**
	(0.060)	(0.040)	(0.026)	(0.222)	(0.054)	(0.040)	(0.173)	(0.054)	(0.040)	(0.276)	(0.066)	(0.041)
60	**-0.002**	**-0.001**	**-0.001**	**-0.086**	**-0.007**	**-0.001**	**-0.013**	**-0.003**	**-0.001**	**-0.023**	**-0.005**	**-0.001**
	(0.038)	(0.031)	(0.026)	(0.184)	(0.049)	(0.030)	(0.093)	(0.043)	(0.030)	(0.120)	(0.050)	(0.029)
90	**-0.001**	**-0.001**	**0.000**	**-0.191**	**-0.004**	**-0.001**	**-0.010**	**-0.001**	**-0.001**	**-0.014**	**-0.002**	**-0.001**
	(0.031)	(0.025)	(0.021)	(0.252)	(0.041)	(0.026)	(0.077)	(0.035)	(0.026)	(0.086)	(0.039)	(0.025)
150	**-0.001**	**0.000**	**0.000**	**-0.237**	**-0.003**	**-0.001**	**-0.005**	**-0.001**	**0.000**	**-0.008**	**-0.001**	**0.000**
	(0.027)	(0.019)	(0.017)	(0.260)	(0.034)	(0.023)	(0.067)	(0.026)	(0.022)	(0.071)	(0.028)	(0.020)
300	**0.000**	**0.000**	**0.000**	**-0.085**	**-0.003**	**-0.001**	**-0.002**	**-0.001**	**0.000**	**-0.002**	**-0.001**	**0.000**
	(0.017)	(0.014)	(0.012)	(0.175)	(0.030)	(0.020)	(0.034)	(0.020)	(0.015)	(0.034)	(0.051)	(0.015)

Table 1b: Bias (SE) β_1

	LS			PC			BN			QPC		
$n \setminus T$	6	9	12	6	9	12	6	9	12	6	9	12
30	**0.473**	**0.480**	**0.486**	**0.218**	**0.064**	**0.042**	**0.154**	**0.041**	**0.031**	**0.126**	**0.057**	**0.047**
	(0.138)	(0.110)	(0.097)	(0.242)	(0.106)	(0.082)	(0.240)	(0.097)	(0.076)	(0.262)	(0.093)	(0.073)
60	**0.475**	**0.483**	**0.486**	**0.129**	**0.039**	**0.027**	**0.054**	**0.014**	**0.009**	**0.040**	**0.012**	**0.009**
	(0.123)	(0.101)	(0.087)	(0.189)	(0.077)	(0.055)	(0.141)	(0.063)	(0.046)	(0.135)	(0.063)	(0.047)
90	**0.476**	**0.484**	**0.487**	**0.265**	**0.031**	**0.022**	**0.041**	**0.007**	**0.004**	**0.027**	**0.006**	**0.005**
	(0.121)	(0.096)	(0.083)	(0.241)	(0.057)	(0.043)	(0.116)	(0.046)	(0.035)	(0.100)	(0.048)	(0.037)
150	**0.475**	**0.483**	**0.489**	**0.289**	**0.023**	**0.014**	**0.027**	**0.002**	**0.001**	**0.013**	**0.002**	**0.001**
	(0.118)	(0.093)	(0.080)	(0.252)	(0.042)	(0.031)	(0.095)	(0.033)	(0.028)	(0.078)	(0.036)	(0.030)
300	**0.475**	**0.485**	**0.488**	**0.121**	**0.027**	**0.010**	**0.007**	**0.001**	**0.000**	**0.003**	**0.001**	**0.000**
	(0.111)	(0.090)	(0.078)	(0.161)	(0.035)	(0.022)	(0.048)	(0.024)	(0.020)	(0.037)	(0.026)	(0.021)

Table 1c: Bias (SE) β_2

	LS			PC			BN			QPC		
$n \setminus T$	6	9	12	6	9	12	6	9	12	6	9	12
30	**0.001**	**0.000**	**0.001**	**-0.054**	**-0.003**	**-0.001**	**-0.020**	**-0.002**	**-0.001**	**-0.055**	**-0.003**	**-0.001**
	(0.142)	(0.100)	(0.088)	(0.197)	(0.095)	(0.081)	(0.199)	(0.097)	(0.082)	(0.236)	(0.104)	(0.085)
60	**-0.001**	**0.000**	**-0.001**	**-0.047**	**-0.004**	**-0.001**	**-0.009**	**-0.002**	**-0.001**	**-0.016**	**-0.004**	**-0.001**
	(0.096)	(0.077)	(0.065)	(0.134)	(0.075)	(0.059)	(0.118)	(0.076)	(0.059)	(0.135)	(0.081)	(0.062)
90	**-0.001**	**0.001**	**-0.001**	**-0.097**	**-0.002**	**0.000**	**-0.006**	**0.000**	**0.000**	**-0.008**	**-0.001**	**0.000**
	(0.081)	(0.062)	(0.053)	(0.141)	(0.059)	(0.048)	(0.095)	(0.060)	(0.048)	(0.104)	(0.065)	(0.051)
150	**0.000**	**0.000**	**0.000**	**-0.012**	**-0.002**	**-0.001**	**-0.003**	**0.000**	**0.000**	**-0.003**	**0.000**	**0.000**
	(0.067)	(0.048)	(0.043)	(0.137)	(0.045)	(0.039)	(0.081)	(0.045)	(0.040)	(0.089)	(0.050)	(0.042)
300	**0.000**	**0.000**	**0.000**	**-0.046**	**-0.002**	**-0.001**	**-0.002**	**0.000**	**0.000**	**-0.002**	**0.000**	**0.000**
	(0.042)	(0.035)	(0.031)	(0.087)	(0.033)	(0.028)	(0.044)	(0.033)	(0.028)	(0.049)	(0.036)	(0.031)

Inspecting Table 1a, the LS estimates of α appear to perform relatively well, which is expected since the model is not transformed in any way and the errors and factors are both drawn independently in each time period. The PC estimates of α on the other hand, suffer from a bias with fixed T originating from the implicit transformation of the model to remove the factor term, which generates Nickell bias in the autoregressive coefficient. As expected, both the BN and the QPC estimates of α are unbiased as n increases. For the coefficient β_1, the LS estimates are severely biased, with this bias being persistent irrespective of n and T. For small T, the PC estimates are also biased, which stems from the heteroskedasticity of the errors in both the cross-section and across time. Only where both n and T are large does this bias diminish. Owing to the significant heteroskedasticity in the design, when both n and T are small, BN and QPC have sizeable biases - though smaller in magnitude than LS or PC. This bias diminishes rapidly as n increases. Since $\boldsymbol{X}_2$ is neither dynamic, nor correlated with the factor term, estimates of β_2 generally perform well across all n and T. Tables 2a – 2b below present coverage probabilities of the estimators based on the asymptotic variance-covariance matrix, and with a nominal value of 95%.

Table 2a: Coverage α %

	LS			PC			BN			QPC		
$n \setminus T$	6	9	12	6	9	12	6	9	12	6	9	12
30	85.35	85.30	86.65	60.32	83.73	88.69	77.53	88.66	89.84	45.21	72.39	70.86
60	85.47	86.74	87.31	53.57	79.59	86.39	85.32	91.16	92.68	67.83	74.33	78.90
90	86.47	87.14	87.61	30.72	76.51	83.49	87.40	91.98	93.08	72.46	80.35	81.91
150	86.99	86.26	88.01	22.63	71.12	79.61	88.83	93.29	93.34	71.22	86.40	85.66
300	84.46	87.16	87.70	27.85	62.05	72.27	91.22	93.63	93.91	82.59	88.49	90.03

Table 2b: Coverage β_1 %

	LS			PC			BN			QPC		
$n \setminus T$	6	9	12	6	9	12	6	9	12	6	9	12
30	00.84	00.02	00.00	53.68	74.19	80.50	64.39	81.51	84.43	61.75	82.86	85.62
60	00.13	00.00	00.00	62.17	80.23	82.98	77.18	88.67	91.07	74.38	86.42	90.40
90	00.07	00.00	00.00	35.07	81.29	84.92	80.65	90.98	93.16	80.15	87.51	92.01
150	00.03	00.00	00.00	33.31	81.52	87.90	82.75	92.56	93.79	79.76	89.44	92.03
300	00.00	00.00	00.00	44.95	73.44	88.59	89.63	93.00	94.01	82.85	90.88	92.79

Table 2b: Coverage β_2 %

	LS			PC			BN			QPC		
$n \setminus T$	6	9	12	6	9	12	6	9	12	6	9	12
30	86.54	84.87	85.91	84.78	93.09	93.49	86.74	92.95	93.38	75.79	90.91	92,29
60	84.79	86.30	86.42	81.87	92.86	93.62	88.34	92.94	93.48	80.25	89.33	92.23
90	85.72	86.17	87.23	66.25	93.29	94.16	90.55	93.22	94.16	85.79	88.76	92.70
150	88.13	86.56	87.71	53.81	93.17	93.92	91.03	93.17	93.86	83.23	90.15	92.98
300	83.71	87.26	87.41	74.34	93.56	93.55	91.89	93.50	93.92	84.16	91.66	93.07

For α the coverage of LS remains consistently below its nominal value, while for PC it decreases with fixed T. In the case of the latter, this decrease in coverage is expected due to the fixed T bias, with coverage only improving when both n and T increase. In contrast, the coverage of BN and QPC readily improve as n increases, with T fixed or T increasing slowly. The story is similar for β_1 in Table 2b. The coverage of LS is incredibly poor, presenting near 0 across all n, T values. The coverage of PC is also poor with either n or T small, and improves only as both of these increase. BN and QPC present poor coverage with both n and T small, yet these rapidly improve as n increases. When comparing the performance of BN and QPC, it is, in general, the case that BN outperforms QPC. This is a consequence of the fact that, while omitting a time period, BN uses only 2 factors in estimation, whereas QPC uses 3, with the extra factor being present to control for a possibly endogenous initial condition. Clearly, including an additional factor in estimation has a noticeable impact on the efficiency of the estimator in finite samples, therefore it is useful to apply the eigenvalue ratio test described in Section 6.3 to uncover the appropriate number of factors to use in estimation.

Table 3: Number of Factors Chosen QPC %

	EigR = 2			EigR = 3		
$n \setminus T$	6	9	12	6	9	12
30	27.26	41.36	47.15	30.58	08.66	05.28
60	40.74	60.39	70.11	11.86	01.01	00.35
90	45.00	70.14	77.21	07.67	00.29	00.11
150	55.09	79.98	89.01	03.52	00.06	00.00
300	73.36	85.15	93.62	00.02	00.00	00.00

Table 3 presents the percentage of times that the number of factors is chosen to be either 2 or 3 when applying the modified eigenvalue ratio test (EigR) described in Section 6.3 to the QPC residuals. Only in the smallest sample size, $n = 30$, $T = 6$, is the number of factors chosen to be 3. In all other cases the percentage of times 3 factors is chosen is very small. Indeed, this number declines rapidly as either n or T grows, while the number of times 2 is chosen increases. This suggests that the impact of the initial condition becomes negligible as either dimension of the panel grows. In light of this, Tables 4a and 4b below present bias and coverages for QPC with $R = 2$. Comparing these results to those presented previously, these estimates are generally better than both BN and QPC with an additional factor. This is unsurprising since it uses both the lower number of factors, and retains an additional time period when compared to BN. However, BN still outperforms QPC when it comes to the autoregressive parameter α.

Table 4a: Bias (SE), QPC with $R = 2$

	α			β_1			β_2		
$n \setminus T$	6	9	12	6	9	12	6	9	12
30	**-0.034**	**-0.004**	**-0.002**	**0.126**	**0.057**	**0.047**	**-0.013**	**-0.002**	**-0.001**
	(0.150)	(0.054)	(0.041)	(0.211)	(0.108)	(0.088)	(0.171)	(0.097)	(0.083)
60	**-0.008**	**-0.003**	**-0.001**	**0.040**	**0.012**	**0.009**	**-0.006**	**-0.001**	**-0.001**
	(0.072)	(0.039)	(0.029)	(0.111)	(0.058)	(0.045)	(0.103)	(0.071)	(0.058)
90	**-0.006**	**-0.001**	**-0.001**	**0.027**	**0.006**	**0.005**	**-0.004**	**0.000**	**0.000**
	(0.058)	(0.032)	(0.025)	(0.087)	(0.043)	(0.036)	(0.084)	(0.057)	(0.048)
150	**-0.002**	**-0.001**	**0.000**	**0.013**	**0.002**	**0.001**	**-0.001**	**0.000**	**0.000**
	(0.048)	(0.024)	(0.020)	(0.064)	(0.032)	(0.027)	(0.069)	(0.044)	(0.038)
300	**-0.001**	**-0.001**	**0.000**	**0.003**	**0.001**	**0.000**	**-0.001**	**0.000**	**0.000**
	(0.027)	(0.018)	(0.015)	(0.034)	(0.024)	(0.020)	(0.039)	(0.034)	(0.028)

Table 4b: Bias (SE), QPC with $R = 2$

	α			β_1			β_2		
$n \setminus T$	6	9	12	6	9	12	6	9	12
30	59.75	78.99	76.18	67.00	76.62	78.19	88.19	92.71	93,31
60	82.18	84.03	86.08	81.02	89.55	91.50	90.95	93.61	94.16
90	83.72	87.77	86.74	84.62	91.37	93.03	92.85	93.25	94.01
150	84.75	91.32	89.04	87.86	93.05	93.80	92.26	93.36	93.89
300	92.34	92.39	92.11	90.95	93.42	94.06	92.39	93.47	94.02

8 Conclusion

In conclusion, this paper introduces a new method to estimate linear panel data models with interactive fixed effects designed for situations where T is small relative to n, or indeed may be fixed. By transforming the model and then applying the PC estimator of Bai (2009), the approach proposed in this paper is shown to deliver $\sqrt{n}$ consistent estimates of regression slope coefficients with T fixed which are, moreover, asymptotically unbiased in the presence of cross-sectional dependence, serial dependence, and with the inclusion of dynamic regressors. This contrasts starkly with the usual case where the PC estimator generally delivers biased and inconsistent estimates with T fixed. Several other consequences of this approach are also discussed, particularly the ability to apply other inferential procedures designed for the large n, large T setting to the transformed model. This is illustrated by modifying to the eigenvalue ratio test of Ahn and Horenstein (2013) to render it applicable in the present setting.

There are two natural extensions to this paper, both of which are currently in progress. The first is to notice that the estimator proposed in this paper can be obtained as a marginal likelihood associated with a maximal invariant statistic under the group of transformations (2.4). Using the full likelihood of the maximal invariant may potentially lead to improved estimation of the autoregressive parameter, as has been shown in a similar context by Barbosa and Moreira (2020). The second extension is to incorporate more general predetermined regressors which are intuitively difficult to handle in this framework.

The results and perspective introduced in this paper also suggest several other avenues for future research. For example, the transformation introduced in this paper, or similar, might be applied to other inferential procedures designed for a large panels, potentially rendering them amenable to the fixed T setting. Finally, it is worth stressing that arguably the most powerful concept developed in this research is the idea that multi-dimensional nuisance parameters may be removed from one (or possibly several) dimensions, by reducing the model into a lower dimensional subspace. This, really, is what lies at the heart of this paper and may well prove to be fruitful in other applications.

References

Ahn, S. C., Horenstein, A. R., 2013. Eigenvalue ratio test for the number of factors. Econometrica 81 (3), 120 – 1227.

Ahn, S. C., Lee, Y. H., Schmidt, P., 2013. Panel data models with multiple time-varying individual effects. Journal of Econometrics 174 (1), 1 – 14.

Bai, J., 2003. Inferential theory for factor models of large dimensions. Econometrica 71 (1), 135 – 171.

Bai, J., 2009. Panel data models with interactive fixed effects. Econometrica 77 (4), 1229 – 1279.

Balestra, P., Nerlove, M., 1966. Pooling cross section and time series data in the estimation of a dynamic model: The demand for natural gas. Econometrica 34 (3), 585–612.

Barbosa, J. D., Moreira, M. J., 2020. Likelihood inference and the role of initial conditions for the dynamic panel data model. Journal of econometrics (in press).

Chamberlain, G., 1984. Chapter 22: Panel data. In: Griliches, Z., Intriligator, M. (Eds.), Handbook of Econometrics. Vol. 2. Elsevier, pp. 1247 – 1318.

Hayakawa, K., 2016. Identification problem of GMM estimators for short panel data models with interactive fixed effects. Economics Letters 139, 22 – 26.

Latala, R., 2005. Some estimates of norms of random matrices. Proceedings of the American Mathematical Society 133 (5), 1273–1282.

Moon, H. R., Weidner, M., 2015. Linear regression for panel with unknown number of factors as interactive fixed effects. Econometrica 83 (4), 1543 – 1579.

Moon, H. R., Weidner, M., 2017. Dynamic linear panel regression models with interactive fixed effects. Econometric Theory 33 (1), 158 – 195.

Neyman, J., Scott, E. L., 1948. Consistent estimates based on partially consistent observations. Econometrica 16 (1), 1–32.

Nickell, S., 1981. Biases in dynamic models with fixed effects. Econometrica 49 (6), 1417–1426.

Onatski, A., 2009. Testing hypotheses about the number of factors in large factor models. Econometrica 77 (5), 1447 – 1479.

Pesaran, H., 2006. Estimation and inference in large heterogeneous panels with a multifactor error structure. Econometrica 74 (4), 967 – 1012.