Panel Data Models with Interactive Fixed Effects and Relatively Small T

Ayden Higgins[†]
University of Oxford

December 4, 2022

Abstract

This paper studies the estimation of a linear panel data model with interactive fixed effects where T is small relative to n. This encompasses the situation where T is fixed and $n \to \infty$, and also where both $n, T \to \infty$ but $T/n \to 0$. A transformation is introduced which reduces the model to a lower dimension, and, in doing so, relives the model of incidental parameters in the cross-section. The consequences of this transformation turn out to be remarkably far-reaching, and it is shown that simply transforming the model and then applying the least squares interactive fixed effects (LS-IFE) estimator of Bai (2009) will deliver estimates that are consistent, unbiased, and asymptotically normal when T is small relative to n, even where the error exhibits cross-sectional or temporal dependence and/or heteroskedasticity. This contrasts sharply with the usual case where the LS-IFE estimator is, in general, inconsistent with T fixed, and suffers from asymptotic bias which impedes inference when $n, T \to \infty$ but $T/n \to 0$.

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.
[†]Address: Department of Economics, University of Oxford, 10 Manor Road, Oxford, OX1 3UQ, United Kingdom. E-mail: `ayden.higgins@economics.ox.ac.uk` .

1 Introduction

1.1 Overview

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 $i = 1, \dots, n$ and $t = 1, \dots, T$, 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 (with respect to ε) 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.[1] 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 $\boldsymbol{\theta} \coloneqq (\alpha, \boldsymbol{\beta}^\top)^\top$ 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

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

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, and, as a result, naive estimation approaches will typically produce inconsistent estimates. One possible remedy 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 the main alternative,

[1] For ease of presentation a first order autoregressive process is assumed throughout this paper. It is straightforward to extend the analysis and the results to cover higher order autoregressive processes.

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, generates complications for the estimation of the parameter of interest $\boldsymbol{\theta}$. These complications arise as a consequence of the incidental parameter problem (Neyman and Scott, 1948) which describes the situation where the presence of high-dimensional nuisance parameters aversely impacts the estimation of other parameters in the model. In long panels this problem can, to some extent, be overcome, and in particular it has been shown that the least squares estimator that treats both the factors and the loadings as fixed effects is consistent as both n and $T \to \infty$, though it typically suffers form asymptotic bias (Bai, 2009; Moon and Weidner, 2017). It is, however, in short panels that the incidental parameter problem is felt most acutely, and with T fixed the least squares interactive fixed effects (LS-IFE) estimator is generally inconsistent outside of exceptional circumstances.

This paper proposes a simple remedy to the problem by introducing a transformation of the model, which, after having been applied, enables the LS-IFE estimator to produce consistent estimates with T fixed. In contrast to typical approaches, this transformation is not designed to purge the incidental parameters from the model entirely. Instead, the aim is to reduce the dimension of the model, and, in doing so, relieve it of incidental parameters in the cross-section. 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. And yet, despite this simplicity, the transformation is shown to have remarkably far-reaching consequences, and, in the main result of this paper, it is found that that simply transforming the model and then applying the LS-IFE estimator will produce not only consistent, but also asymptotically unbiased estimates with T fixed, irrespective of the possible inclusion of dynamic regressors, and the presence of cross-sectional and serial dependence and/or heteroskedasticity in the error. The precise mechanisms through which this is achieved are studied in close detail which reveals that, under certain conditions, these properties also carry over to the large n, large T setting when the ratio $T/n \to 0$. These results contrast sharply with the usual case where, as well as being generally inconsistent with T fixed, the LS-IFE estimator suffers from asymptotic bias which impedes inference when $n, T \to \infty$ but $T/n \to 0$.

1.2 Related Literature

Given that the LS-IFE estimator is generally inapplicable to short panels, and that such datasets have been historically, and still remain today, very common in economics, a great deal of research - both theoretical and applied - has focused on alternative estimation strategies. Yet unlike in the large panels, where little if anything need be assumed about the relationship between the factors, the loadings and the covariates, many, if not most of the approaches designed for short panels rely on the possibility

of correlation existing between these and indeed lean into this possibility as a means to derive alternative estimators.[2] In this line of research, approaches may broadly be placed into one of two groups: those that impose a specific functional form for the relationship between the factor term and covariates, and those that do not.

The first group consists of common correlated effects approaches, which originate from the seminal paper of Pesaran (2006). At the core of this approach is the assumption that at least some model covariates also admit a factor decomposition, such that the factors can be instrumented by taking cross-sectional averages of these covariates. In such cases, these instruments can be levered to purge the factor term, and, ultimately, give rise to estimators that are often consistent with T fixed, as well as where both n and T diverge. The properties of this approach have been extensively studied: a likelihood setting (Bai and Li, 2014), with dynamics (Everaert and Groote, 2016), with an unknown number of factors (Westerlund and Urbain, 2015). Other contributions in this line of research include Westerlund (2020), Vos and Everaert (2021) and Juodis and Sarafidis (2022b). Though these methods are often easy to implement, the imposition of a particular relationship between the factors and the covariates can be restrictive, and whether or not this is reasonable assumption is largely a matter of context. This leads naturally to the second group of methods that seek to exploit possible correlation between observed covariates and the factor structure, without imposing any particular functional form for this relationship.

This second group might simply be termed correlated effects approaches and includes quasi-difference approaches (Holtz-Eakin et al., 1988; Ahn et al., 2001, 2013), the instrumental variables estimators of Robertson and Sarafidis (2015), and the hybrid approach of Juodis and Sarafidis (2022a). Either directly or indirectly, these approaches also hinge on the existence of correlation between the factor term and observed covariates, which can be exploited to derive moment conditions that then can be used to produce estimators that are fixed T-consistent. The present paper is closely related to this line of research, though takes a different approach centred around the LS-IFE estimator. While subsequent sections show that when applied to the transformed model, the LS-IFE estimator is, in some instances, equivalent to the quasi-difference and instrumental variables approaches, implementing those estimators comes with varying degrees of complexity, with it being necessary to directly estimate multiple nuisance parameters alongside the parameter of interest $\boldsymbol{\theta}$, and in some cases to do so from a set of highly non-linear moment conditions. On the other hand, this paper shows that estimation can sometimes be reduced to the optimisation of a univariate objective function which depends only on the parameters of interest, including when the model is dynamic. Bridging the gap between the LS-IFE estimator and method of moments-based approaches is also useful since at present little is known about the properties of the latter approaches outside of the fixed T setting including whether or not, and in which instances, they may still

[2] An exception to this is Hsiao et al. (2021).

produce consistent and asymptotically unbiased estimates.[3]

Outline: Section 2 sets out the estimation approach, introducing the transformation and discusses the mechanics of the LS-IFE estimator. Section 3 establishes consistency, under quite general conditions, and draws a comparison with existing results. The asymptotic distribution of the estimator is derived in Section 4 culminating in Theorem 1, the main result of this paper. This is followed by a discussion of the large n, large T properties of the estimator in Section 5. Some additional considerations are collected in Section 6 including an alternative interpretation of the estimator, and a method to detect the correct number of factors. Monte Carlo simulations follow in Section 7. Section 8 concludes. Additional discussion and results, as well as proofs, 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, $||\boldsymbol{a}||_1 \coloneqq \sum_{i=1}^n |a_i|$, $||\boldsymbol{a}||_2 \coloneqq \sqrt{\sum_{i=1}^n a_i^2}$, $||\boldsymbol{a}||_\infty \coloneqq \max_{1\leq i\leq n} |a_i|$. 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 $\mu_{\min}(\boldsymbol{A}) \coloneqq \mu_n(\boldsymbol{A}) \leq \ldots \leq \mu_1(\boldsymbol{A}) =: \mu_{\max}(\boldsymbol{A})$. The following matrix norms are those induced by their vector counterparts: $||\boldsymbol{A}||_1 \coloneqq \max_{1\leq j\leq m} \sum_{i=1}^n |A_{ij}|$ which is the maximum absolute column sum of $\boldsymbol{A}$, $||\boldsymbol{A}||_2 \coloneqq \sqrt{\mu_{\max}(\boldsymbol{A}^\top \boldsymbol{A})}$, and $||\boldsymbol{A}||_\infty \coloneqq \max_{1\leq i\leq n} \sum_{j=1}^m |A_{ij}|$ which is the maximum absolute row sum of $\boldsymbol{A}$. The Frobenius norm of $\boldsymbol{A}$ is denoted $||\boldsymbol{A}||_F \coloneqq \sqrt{\text{tr}(\boldsymbol{A}^\top \boldsymbol{A})}$. The notation $||\boldsymbol{A}||_{\max}$ is used to denote $\max_{1\leq i,j\leq n} |A_{ij}|$. Let $\boldsymbol{P_A} \coloneqq \boldsymbol{A}(\boldsymbol{A}^\top \boldsymbol{A})^+ \boldsymbol{A}^\top$ and $\boldsymbol{M_A} \coloneqq \boldsymbol{I}_n - \boldsymbol{P_A}$, 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 $\boldsymbol{\iota}_n$, an $n \times 1$ vector of zeros is denoted $\boldsymbol{0}_n$, and an $n \times m$ matrix of zeros is denoted $\boldsymbol{0}_{n\times m}$. 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. The operation vec$(\cdot)$ applied to an $n \times m$ matrix $\boldsymbol{A}$ creates an $nm \times 1$ vector vec$(\boldsymbol{A})$ by stacking the columns of $\boldsymbol{A}$. The operation diag$(\boldsymbol{B})$ applied to an $n \times n$ matrix $\boldsymbol{B}$ creates an $n \times n$ diagonal matrix diag$(\boldsymbol{B})$ which contains the diagonal elements of $\boldsymbol{B}$ along its diagonal. The shorthand diagv$(\boldsymbol{B})$ is used to indicate diag$(\boldsymbol{B})\boldsymbol{\iota}_n$ and off$(\boldsymbol{B}) \coloneqq \boldsymbol{B} - \text{diag}(\boldsymbol{B})$.

2 Estimation Approach

Treating both the factors and the loadings as additional (nuisance) parameters in the model, the LS-IFE estimator of (1.1) is obtained as the set of parameter values $(\boldsymbol{\theta}, \boldsymbol{\Lambda}, \boldsymbol{F})$ which minimise the sum of squared residuals. In seminal work, Bai (2009) studies the properties of this estimator and shows that with, strictly exogenous covariates, the LS-IFE estimator delivers consistent estimates of regression slope coefficients, and of

[3] This paper also closely related to the influential work of Balestra and Nerlove (1966), Nickell (1981), and Chamberlain and Moreira (2009).

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, including lagged outcomes. These authors establish the asymptotic properties of the LS-IFE estimator and, in particular, document asymptotic biases that arise in the presence of cross-sectional and serial dependence and/or heteroskedasticity, and due to inclusion of predetermined regressors. These biases originate from the incidental parameter problem and ultimately cause the LS-IFE estimator to be inconsistent when T is fixed. Yet, as is shown subsequently, where the covariates are strictly exogenous, by first transforming the model, the LS-IFE estimator can be used to produce estimates that are consistent and asymptotically unbiased when T is small relative to n.

2.1 Transformation of the Model

It is useful to begin by re-writing the model in matrix form. Let the $n \times T$ matrix $\boldsymbol{Y} \coloneqq (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_T)$, $\boldsymbol{Y}_L \coloneqq (\boldsymbol{y}_0, \ldots, \boldsymbol{y}_{T-1})$, $\boldsymbol{X}_k$ be the $n \times T$ matrix containing observations of the k-th covariate, and the $T \times R$ matrix $\boldsymbol{F} \coloneqq (\boldsymbol{f}_1, \ldots, \boldsymbol{f}_T)^\top$. With this notation, the model can be written more succinctly as

$$\begin{aligned} \boldsymbol{Y} &= \alpha \boldsymbol{Y}_L + \sum_{k=1}^{K} \beta_k \boldsymbol{X}_k + \boldsymbol{\Lambda}\boldsymbol{F}^\top + \boldsymbol{\varepsilon} \\ &=: \sum_{\kappa=1}^{K+1} \theta_\kappa \boldsymbol{Z}_\kappa + \boldsymbol{\Lambda}\boldsymbol{F}^\top + \boldsymbol{\varepsilon}, \end{aligned} \tag{2.1}$$

where $\boldsymbol{Z}_1 \coloneqq \boldsymbol{Y}_L$ and $\boldsymbol{Z}_\kappa \coloneqq \boldsymbol{X}_{\kappa-1}$ for $\kappa = 2, \ldots, K+1$. Now, define the $n \times TK$ matrix $\boldsymbol{\mathcal{X}} \coloneqq (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_K)$ which is assumed to have full column rank.[4] Consider the following group of transformations $\mathcal{G}$, where each element in this group is a bijective mapping from the sample space to itself:

$$\mathcal{G} \coloneqq \{\boldsymbol{Q} \in \mathfrak{O}(n) : \boldsymbol{Q}\boldsymbol{\mathcal{X}} = \boldsymbol{\mathcal{X}}\}, \tag{2.2}$$

with $\mathfrak{O}(n)$ being the group of $n \times n$ orthogonal matrices. This group $\mathcal{G}$ contains orthogonal transformations that preserve $\boldsymbol{\mathcal{X}}$. Take some $\boldsymbol{Q} \in \mathcal{G}$. This can be partitioned as $\boldsymbol{Q} =: (\boldsymbol{Q}_{\mathcal{X}}, \boldsymbol{Q}_{\mathcal{X}^\perp})$, 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.

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

A simple way to construct $\boldsymbol{Q}_{\mathcal{X}}$ is as $\mathcal{X}(\mathcal{X}^\top\mathcal{X})^{-\frac{1}{2}}$, and, with this in hand, the following transformed variables can be defined:

$$\begin{aligned}\tilde{\boldsymbol{Y}} &\coloneqq \boldsymbol{Q}_{\mathcal{X}}^\top \boldsymbol{Y},\\ \tilde{\boldsymbol{Z}}_\kappa &\coloneqq \boldsymbol{Q}_{\mathcal{X}}^\top \boldsymbol{Z}_\kappa,\\ \tilde{\boldsymbol{\Lambda}} &\coloneqq \boldsymbol{Q}_{\mathcal{X}}^\top \boldsymbol{\Lambda},\\ \tilde{\boldsymbol{\varepsilon}} &\coloneqq \boldsymbol{Q}_{\mathcal{X}}^\top \boldsymbol{\varepsilon},\end{aligned}$$

in which case premultiplying (2.1) by $\boldsymbol{Q}_{\mathcal{X}}^\top$ yields the transformed model

$$\tilde{\boldsymbol{Y}} = \sum_{\kappa=1}^{K+1} \theta_\kappa \tilde{\boldsymbol{Z}}_\kappa + \tilde{\boldsymbol{\Lambda}}\boldsymbol{F}^\top + \tilde{\boldsymbol{\varepsilon}}. \tag{2.3}$$

Looking at (2.3) there are three significant consequences of transforming the model through $\boldsymbol{Q}_{\mathcal{X}}$ that need to be highlighted. First, the resultant matrices $\tilde{\boldsymbol{Y}}$, $\tilde{\boldsymbol{Z}}_k$ and $\tilde{\boldsymbol{\Lambda}}\boldsymbol{F}^\top$ are of dimension $TK \times T$, since the entirety of the model has been transformed by $\boldsymbol{Q}_{\mathcal{X}}$ and projected into the TK-dimensional subspace spanned by the columns of the covariates. Hence, the dimension of the transformed factor term $\tilde{\boldsymbol{\Lambda}}\boldsymbol{F}^\top$ will no longer depend on n, and the model is relived of incidental parameters as $n \to \infty$.[5] Second, the transformation leads to no loss of information in the exogenous covariates since, by construction, transforming the model through $\boldsymbol{Q}_{\mathcal{X}}$ preserves the column space of $\mathcal{X}$. Thirdly, since the covariates used in the construction of $\boldsymbol{Q}_{\mathcal{X}}$ are strictly exogenous, under quite general conditions, including broad cross-sectional and serial dependence, the transformation serves to reduce the order of the error term which, ultimately, is key to estimating (2.3) using the LS-IFE estimator.[6]

2.2 Principal Components

The underlying mechanics of the LS-IFE estimator are most easily understood with the intuition that, given the factors and the loadings, the coefficients can be estimated by a linear regression, and, similarly, given $\boldsymbol{\theta}$, estimating the factors and loadings is a standard principal components problem. Where T is small relative to n, it is the 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 estimation of the original model, and

[5] 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; see Section 5.

[6] 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 $\mathcal{X}$.

of its transformed counterpart.

Assume that $\boldsymbol{\theta}$ is observed and define $\dot{\boldsymbol{Y}} \coloneqq \boldsymbol{Y} - \sum_{\kappa=1}^{K+1} \theta_\kappa \boldsymbol{Z}_\kappa = \boldsymbol{\Lambda}\boldsymbol{F}^\top + \boldsymbol{\varepsilon}$ 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.[7] Consider the problem of trying to estimate $\check{\boldsymbol{\Lambda}}$ from the variance of $\dot{\boldsymbol{Y}}$. With suitable conditions on the errors, the factors, and the loadings, as $n \to \infty$,

$$\frac{1}{nT}\dot{\boldsymbol{Y}}\dot{\boldsymbol{Y}}^\top\check{\boldsymbol{\Lambda}} = \frac{1}{nT}\check{\boldsymbol{\Lambda}}\check{\boldsymbol{F}}^\top\check{\boldsymbol{F}} + \frac{1}{nT}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top\check{\boldsymbol{\Lambda}} + \boldsymbol{\mathcal{O}}_p(1). \tag{2.4}$$

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. Where both n and T are large, several authors have shown that, in spite of this distortionary term, estimating $\check{\boldsymbol{\Lambda}}$ 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 $\frac{1}{nT}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top\check{\boldsymbol{\Lambda}} = \boldsymbol{\mathcal{O}}_p(1)$ 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 estimate $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}$, and then $\check{\boldsymbol{\Lambda}}$ 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 in the most exceptional 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 when T is small relative to n.

Consider, on the other hand, the transformed model. Let $\dddot{\boldsymbol{\Lambda}}$ denote an analogue of $\check{\boldsymbol{\Lambda}}$. With $\frac{1}{nT}||\tilde{\boldsymbol{\varepsilon}}\boldsymbol{F}\tilde{\boldsymbol{\Lambda}}^\top\dddot{\boldsymbol{\Lambda}}||_2 = \mathcal{O}_p(1)$, one arrives at a similar expression to (2.4),

$$\frac{1}{nT}\dddot{\boldsymbol{Y}}\dddot{\boldsymbol{Y}}^\top\dddot{\boldsymbol{\Lambda}} = \frac{1}{nT}\dddot{\boldsymbol{\Lambda}}\check{\boldsymbol{F}}^\top\check{\boldsymbol{F}} + \frac{1}{nT}\tilde{\boldsymbol{\varepsilon}}\tilde{\boldsymbol{\varepsilon}}^\top\dddot{\boldsymbol{\Lambda}} + \boldsymbol{\mathcal{O}}_p(1). \tag{2.5}$$

Yet now, since the covariates used to construct $\boldsymbol{Q}_{\boldsymbol{X}}$ are strictly exogenous, under quite general conditions $\frac{1}{nT}||\tilde{\boldsymbol{\varepsilon}}\tilde{\boldsymbol{\varepsilon}}^\top||_2 = \frac{1}{nT}||\boldsymbol{\varepsilon}^\top\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{\varepsilon}||_2 = \mathcal{O}_p(1)$, even with T fixed. As a consequence, asymptotically, $\dddot{\boldsymbol{\Lambda}}$ will be an eigenvector of $\frac{1}{nT}\dddot{\boldsymbol{Y}}\dddot{\boldsymbol{Y}}^\top$ 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 LS-IFE estimator to the transformed model is able to control for $\tilde{\boldsymbol{\Lambda}}$ and to deliver consistent estimates of $\boldsymbol{\theta}$ when T is small relative to n.

[7] 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 $\check{\boldsymbol{\Lambda}}\check{\boldsymbol{F}}^\top = \boldsymbol{\Lambda}\boldsymbol{F}^\top$.

2.3 Objective Function

The transformed model (2.3) can be estimated by minimising the following least squares objective function

$$\mathcal{Q}(\boldsymbol{\theta}, \tilde{\boldsymbol{\Lambda}}, \boldsymbol{F}) \coloneqq \frac{1}{nT}\text{tr}\left(\left(\tilde{\boldsymbol{Y}} - \sum_{\kappa=1}^{K+1}\theta_\kappa \tilde{\boldsymbol{Z}}_\kappa - \tilde{\boldsymbol{\Lambda}}\boldsymbol{F}^\top\right)^\top \left(\tilde{\boldsymbol{Y}} - \sum_{\kappa=1}^{K+1}\theta_\kappa \tilde{\boldsymbol{Z}}_\kappa - \tilde{\boldsymbol{\Lambda}}\boldsymbol{F}^\top\right)\right).^{8} \tag{2.6}$$

Both the factors and the transformed loadings can be concentrated out of (2.6), in which case one arrives at an objective function involving $\boldsymbol{\theta}$ alone,

$$\mathcal{Q}(\boldsymbol{\theta}) \coloneqq \frac{1}{nT}\sum_{r=R+1}^{T}\mu_r\left(\left(\tilde{\boldsymbol{Y}} - \sum_{\kappa=1}^{K+1}\theta_\kappa \tilde{\boldsymbol{Z}}_\kappa\right)^\top \left(\tilde{\boldsymbol{Y}} - \sum_{\kappa=1}^{K+1}\theta_\kappa \tilde{\boldsymbol{Z}}_\kappa\right)\right), \tag{2.7}$$

that is, the profile objective function now involves the sum of the $(T-R)$ smallest eigenvalues of the right hand-side matrix.[9] Using this, the estimator $\hat{\boldsymbol{\theta}}$ can then be defined as

$$\hat{\boldsymbol{\theta}} \coloneqq \underset{\boldsymbol{\theta}\in\Theta}{\arg\min}\, \mathcal{Q}(\boldsymbol{\theta}). \tag{2.8}$$

3 Consistency

Throughout the following, both $\boldsymbol{\Lambda}$ and $\boldsymbol{F}$ are treated as fixed parameters in estimation and the superscript 0 is now introduced to distinguish true parameter values. Moreover, let $\mathcal{C}$ denote $\sigma(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_K)$, that is, the sigma 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 $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_K$, $\boldsymbol{\Lambda}^0$ and $\boldsymbol{F}^0$ have uniformly bounded fourth moments.

(iii) Let $\boldsymbol{\Sigma}_{nT} \coloneqq \mathbb{E}[\text{vec}(\boldsymbol{\varepsilon})\text{vec}(\boldsymbol{\varepsilon})^\top|\mathcal{C}]$. Then $||\boldsymbol{\Sigma}_{nT}||_1 \leq c$ uniformly over n and T.

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.[10] Assumption MD(ii) imposes standard conditions on the moments of the covariates, the factors, and the

[8] When estimating the original model, the least squares objective function can be interpreted as the negative of a quasi-likelihood function that uses the standard normal distribution. The objective function (2.6) can then be interpreted as a marginal quasi-likelihood which uses only a part of this.

[9] See equation (3.3) in Moon and Weidner (2015) for details.

[10] Stationarity is imposed to also allow for the possibility that T grows with n in subsequent analysis. In a purely fixed T setting this can be dispensed with.

loadings. Assumption MD(iii) serves to limit the degree of dependence between the errors and the covariates. It still allows for heteroskedasticity in both dimensions of the panel, as well as quite generous serial and cross-sectional dependence in the errors. It is also weaker than the assumption that some or all of the covariates, the loadings and the errors are independently (and sometimes also identically) distributed over i, as is frequently assumed in the literature on short panels with interactive effects (see, for example, Pesaran (2006), Ahn et al. (2013) and Robertson and Sarafidis (2015)). In aid of the following assumption let $\boldsymbol{\delta} \cdot \tilde{\boldsymbol{Z}} \coloneqq \sum_{\kappa=1}^{K+1} \delta_k \tilde{\boldsymbol{Z}}_\kappa$.

Assumption CS (Consistency)**.**

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

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

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 requires there to remain sufficient amount of variation in the regressors after having been projected orthogonal to arbitrary $R \times T$ factors and $R^0 \times TK$ loadings.

Proposition 1 (Consistency – General)**.** *Under Assumptions MD and CS, as* $n \to \infty$,

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

Proposition 1 demonstrates that as $T/n \to 0$ the estimator is consistent. Moreover, where T is fixed it is $\sqrt{n}$-consistent. This result is obtained allowing for generous 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. Indeed, Proposition 1 neither requires that the factors or loadings be correlated with the covariates, nor for that matter, uncorrelated with the error term.[11]

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

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

[11] Bai (2009) also obtains an initial consistency result under weaker conditions on the errors than $||\boldsymbol{\varepsilon}||_2 = \mathcal{O}_p(\sqrt{\min\{n,T\}})$. 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.

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

$$\frac{||\boldsymbol{\varepsilon}||_2}{\sqrt{nT}} = \mathcal{O}_p\left(\frac{1}{\sqrt{\min\{n,T\}}}\right).^{12} \tag{3.1}$$

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 MD(iii) this can be shown to satisfy

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

Recalling the discussion in Section 2.2, it is worth stressing again the importance of the difference between $\boldsymbol{\varepsilon}$ and $\tilde{\boldsymbol{\varepsilon}}$. To highlight this, consider the rudimentary example 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,

$$\mathbb{E}[||\tilde{\boldsymbol{\varepsilon}}||_F^2] = \mathbb{E}[||\boldsymbol{Q}_{\boldsymbol{X}}^\top \boldsymbol{\varepsilon}||_F^2] = \mathbb{E}\left[\mathbb{E}[\mathrm{tr}(\boldsymbol{\varepsilon}^\top \boldsymbol{P}_{\boldsymbol{X}} \boldsymbol{\varepsilon})|\mathcal{C}]\right] = \sigma^2 T^2 K = \mathcal{O}(T^2),$$

from which it then follows by Markov's inequality that $||\tilde{\boldsymbol{\varepsilon}}||_F = \mathcal{O}_p(T)$, and so $\frac{||\tilde{\boldsymbol{\varepsilon}}||_F}{\sqrt{nT}} = \mathcal{O}_p(1)$ as $T/n \to 0$. By comparison,

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

using $\frac{1}{\sqrt{\mathrm{rank}(\boldsymbol{A})}}||\boldsymbol{A}||_F \leq ||\boldsymbol{A}||_2$. Therefore, even in this simple case, $\frac{||\boldsymbol{\varepsilon}||_2}{\sqrt{nT}}$ cannot be $\mathcal{O}_p(1)$ with T fixed, as long as σ is bounded from below by a constant.

4 Asymptotic Distribution

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 (2.7) 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 and 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 $(\tilde{\boldsymbol{\Lambda}}^0 \boldsymbol{F}^{0\top})^\top \tilde{\boldsymbol{\Lambda}}^0 \boldsymbol{F}^{0\top}/nT$. Key to this approach is demon-

[12] Moreover, (3.1) also proves to be important for the asymptotic expansion of the objective function; see Section 4.

strating that the perturbation is asymptotically small, which in this case follows from Proposition 1, whereby $|\theta_\kappa^0 - \hat{\theta}_\kappa|$ is small, and from assuming that the 'perturbation' stemming from the error term, $\frac{||\tilde{\boldsymbol{\varepsilon}}||_2}{\sqrt{nT}}$, 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 $\frac{||\boldsymbol{\varepsilon}||_2}{\sqrt{nT}}$.[13]

Assumption AE (Asymptotic Expansion)**.**

(i) $R = R^0$.

(ii) $\frac{1}{n}\tilde{\boldsymbol{\Lambda}}^{0\top}\tilde{\boldsymbol{\Lambda}}^0 = \frac{1}{n}\boldsymbol{\Lambda}^{0\top}\boldsymbol{P}_{\boldsymbol{\mathcal{X}}}\boldsymbol{\Lambda}^0 \xrightarrow{p} \boldsymbol{\Sigma}_{\tilde{\boldsymbol{\Lambda}}^0}$ 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) For T fixed, the matrix $\frac{1}{T}\boldsymbol{F}^{0\top}\boldsymbol{F}^0 > 0$, otherwise $\frac{1}{T}\boldsymbol{F}^{0\top}\boldsymbol{F}^0 \xrightarrow{p} \boldsymbol{\Sigma}_{\boldsymbol{F}^0} > 0$ 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 LS-IFE estimator is unaffected by over-statement of the number of factors. Though it is conjectured that a similar result can be obtained in the present case, doing so is beyond the scope of this paper and 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 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 and AE, if* $||\boldsymbol{\theta} - \boldsymbol{\theta}^0||_2 = \mathcal{O}_p(1)$*, then, as* $n \to \infty$ *and* $T^2/n \to 0$*,*

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

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

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

by

$$
\begin{aligned}
D_{\kappa\kappa'} &:= \frac{1}{nT}\text{tr}(\tilde{\boldsymbol{Z}}_\kappa \boldsymbol{M}_{\boldsymbol{F}^0}\tilde{\boldsymbol{Z}}_{\kappa'}^\top \boldsymbol{M}_{\tilde{\boldsymbol{\Lambda}}^0}), \\
c_\kappa &:= \frac{1}{\sqrt{nT}}\text{tr}(\tilde{\boldsymbol{Z}}_\kappa \boldsymbol{M}_{\boldsymbol{F}^0}\tilde{\boldsymbol{\varepsilon}}^\top \boldsymbol{M}_{\tilde{\boldsymbol{\Lambda}}^0}), \\
b_\kappa^{(1)} &:= -\frac{1}{\sqrt{nT}}\text{tr}\left(\boldsymbol{M}_{\boldsymbol{F}^0}\tilde{\boldsymbol{\varepsilon}}^\top \boldsymbol{M}_{\tilde{\boldsymbol{\Lambda}}^0}\tilde{\boldsymbol{Z}}_\kappa \boldsymbol{F}^0(\boldsymbol{F}^{0\top}\boldsymbol{F}^0)^{-1}(\tilde{\boldsymbol{\Lambda}}^{0\top}\tilde{\boldsymbol{\Lambda}}^0)^{-1}\tilde{\boldsymbol{\Lambda}}^{0\top}\tilde{\boldsymbol{\varepsilon}}\right), \\
b_\kappa^{(2)} &:= -\frac{1}{\sqrt{nT}}\text{tr}\left(\boldsymbol{M}_{\boldsymbol{F}^0}\tilde{\boldsymbol{Z}}_\kappa^\top \boldsymbol{M}_{\tilde{\boldsymbol{\Lambda}}^0}\tilde{\boldsymbol{\varepsilon}}\boldsymbol{F}^0(\boldsymbol{F}^{0\top}\boldsymbol{F}^0)^{-1}(\tilde{\boldsymbol{\Lambda}}^{0\top}\tilde{\boldsymbol{\Lambda}}^0)^{-1}\tilde{\boldsymbol{\Lambda}}^{0\top}\tilde{\boldsymbol{\varepsilon}}\right), \\
b_\kappa^{(3)} &:= -\frac{1}{\sqrt{nT}}\text{tr}\left(\boldsymbol{M}_{\boldsymbol{F}^0}\tilde{\boldsymbol{\varepsilon}}^\top \boldsymbol{M}_{\tilde{\boldsymbol{\Lambda}}^0}\tilde{\boldsymbol{\varepsilon}}\boldsymbol{F}^0(\boldsymbol{F}^{0\top}\boldsymbol{F}^0)^{-1}(\tilde{\boldsymbol{\Lambda}}^{0\top}\tilde{\boldsymbol{\Lambda}}^0)^{-1}\tilde{\boldsymbol{\Lambda}}^{0\top}\tilde{\boldsymbol{Z}}_\kappa\right).
\end{aligned} \tag{4.1}
$$

Moreover, $r(\boldsymbol{\theta})$ *is* $\mathcal{O}_p(\frac{(1+\sqrt{nT}||\boldsymbol{\theta}^0-\boldsymbol{\theta}||_2)^2}{nT})$.

As will be seen shortly, the term $\boldsymbol{c}$ plays a central role in determining the asymptotic distribution of the estimator. Terms $\boldsymbol{b}^{(1)}, \boldsymbol{b}^{(2)}$ and $\boldsymbol{b}^{(3)}$ 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 asymptotic bias. Under Assumptions MD and AE, it can be established that $\boldsymbol{b}^{(1)}, \boldsymbol{b}^{(2)}$ and $\boldsymbol{b}^{(3)}$ are $\boldsymbol{\mathcal{O}}_p(T^{1/5}/\sqrt{n})$ which suggests that the estimator will be asymptotically unbiased where $T^3/n \to 0$. This is formalised in the following result and is of course trivially satisfied where T is fixed.

In order to establish the asymptotic distribution of the estimator, it is necessary to place some restrictions on the relationship between the factors, the loadings and the errors. In aid of this some additional notation is introduced. Let $\boldsymbol{\pi}_J$ be a $J \times 1$ vector of all zeros, except the first element which equals 1. Moreover, let $\boldsymbol{W}$ be a $T \times T$ shift matrix with zeros everywhere, except those elements directly above the main diagonal, which take a value of 1. Define $\boldsymbol{G}(\alpha) := (\boldsymbol{I} - \alpha\boldsymbol{W})^{-1}\boldsymbol{W}$, $\boldsymbol{G} := \boldsymbol{G}(\alpha^0)$, $\boldsymbol{g}(\alpha) := (\boldsymbol{I}_T + \alpha\boldsymbol{G})^\top\boldsymbol{\pi}$, $\boldsymbol{g} := \boldsymbol{g}(\alpha^0)$, $\boldsymbol{H}_0 := \sum_{h=0}^{\infty}(\alpha^0)^h\boldsymbol{X}_{-h}\boldsymbol{\beta}^0\boldsymbol{g}^\top$, $\boldsymbol{H}_1 := \sum_{k=1}^{K}\beta_k^0\boldsymbol{X}_k\boldsymbol{G}$ and $\boldsymbol{H}_\kappa := \boldsymbol{X}_{\kappa-1}$ for $\kappa = 2, \dots K+1$, $\boldsymbol{\mathcal{H}} := (\text{vec}(\boldsymbol{H}_1), \dots, \text{vec}(\boldsymbol{H}_{K+1}))$, and $\boldsymbol{\mathcal{H}}_+ := \boldsymbol{\mathcal{H}} + (\text{vec}(\boldsymbol{H}_0), \boldsymbol{0}_{nT}, \dots, \boldsymbol{0}_{nT})$.

Assumption AD. (Asymptotic Distribution I)

(i) For each (n,T) there exists a matrix $\boldsymbol{\Sigma}_{nT}$ such that $\text{vec}(\boldsymbol{\varepsilon}) = \boldsymbol{\Sigma}_{nT}^{\frac{1}{2}}\text{vec}(\boldsymbol{U})$ where $\boldsymbol{U}$ is an $n \times T$ matrix with elements which are independent of the exogenous covariates, the factors and the loadings, and independently distributed across i and t, with $\mathbb{E}[u_{it}] = 0$, $\mathbb{E}[u_{it}^2] = 1$ and $\mathbb{E}[u_{it}^4] \leq c$ uniformly over i and t, and $\boldsymbol{\Sigma}_{nT}$ is a non-stochastic, symmetric and positive definite matrix with $||\boldsymbol{\Sigma}_{nT}||_1 \leq c$ uniformly over n and T.

(ii) The elements of the matrix $\boldsymbol{M}_{\tilde{\boldsymbol{\Lambda}}^0}\boldsymbol{H}_\kappa\boldsymbol{M}_{\boldsymbol{F}^0}$ have uniformly bounded fourth moments.

Assumption AD(i) limits the degree of dependence in the error while still allowing for heteroskedasticty and a more limited degree of cross-sectional and temporal dependence. Alternative mixing-type assumptions could also be imposed. What is important, however, is the fact that the errors are now assumed to be independent of the factors and the loadings. Assumption AD(ii) assumes the that the transformed variables $\boldsymbol{M}_{\tilde{\boldsymbol{\Lambda}}^0}\boldsymbol{H}_\kappa\boldsymbol{M}_{\boldsymbol{F}^0}$ have a sufficient number of finite moments.

Theorem 1 (Asymptotic Distribution I)**.** *Assume* $||\boldsymbol{c}||_2 = \mathcal{O}_p(1)$*. Under Assumptions MD, CS, AE, and AD, as* $n \to \infty$ *and* $T^3/n \to 0$*,*

$$\sqrt{nT}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0) \xrightarrow{d} \mathcal{N}(\mathbf{0}_{K+1}, \boldsymbol{\Delta}_+^{-1}\boldsymbol{\Omega}_+\boldsymbol{\Delta}_+^{-1}), \tag{4.2}$$

where

$$\begin{aligned}\boldsymbol{\Delta}_+ &:= \frac{1}{nT}\tilde{\boldsymbol{\mathcal{H}}}_+^\top(\boldsymbol{M}_{\boldsymbol{F}^0} \otimes \boldsymbol{M}_{\tilde{\boldsymbol{\Lambda}}^0})\tilde{\boldsymbol{\mathcal{H}}}_+,\\ \boldsymbol{\Omega}_+ &:= \frac{1}{nT}\tilde{\boldsymbol{\mathcal{H}}}_+^\top(\boldsymbol{M}_{\boldsymbol{F}^0} \otimes \boldsymbol{M}_{\tilde{\boldsymbol{\Lambda}}^0})\tilde{\boldsymbol{\Sigma}}_{nT}(\boldsymbol{M}_{\boldsymbol{F}^0} \otimes \boldsymbol{M}_{\tilde{\boldsymbol{\Lambda}}^0})\tilde{\boldsymbol{\mathcal{H}}}_+,\end{aligned}$$

and $\tilde{\boldsymbol{\Sigma}}_{nT} =: (\boldsymbol{I}_T \otimes \boldsymbol{Q}_{\boldsymbol{\mathcal{X}}}^\top)\boldsymbol{\Sigma}_{nT}(\boldsymbol{I}_T \otimes \boldsymbol{Q}_{\boldsymbol{\mathcal{X}}})$*.*

Theorem 1 constitutes the most general result in this paper and shows that the estimator is unbiased as $n \to \infty$, whether T is fixed or $T \to \infty$, as long as $T^3/n \to 0$. This is in spite of the lagged outcome, as well as the possibility of heteroskedasticity and/or serial and cross-sectional dependence afforded under Assumption AD. Theorem 1 also contrasts sharply with results for the LS-IFE applied to the usual model in which case the estimator is known to be inconsistent with T fixed, and even in the large n, large T setting, would suffer from a bias of diverging order when T is small relative to n, which would impede inference. The following section investigates in greater detail the properties of the estimator when both n and T are large and provides additional assumptions under which it is possible to establish consistency and unbiasedness under the weaker condition $T/n \to 0$.

5 Large n, Large T Properties

In order to paint a more complete picture of the asymptotic properties of the estimator, it is necessary to study it under an asymptotic regime where both $n, T \to \infty$ and possibly $T/n \to c > 0$. To this end the following additional assumption is introduced. Let $\tilde{\boldsymbol{\Sigma}}_{nT}^{\frac{1}{2}} := (\boldsymbol{I}_T \otimes \boldsymbol{Q}_{\boldsymbol{\mathcal{X}}}^\top)\boldsymbol{\Sigma}_{nT}^{\frac{1}{2}}$.

Assumption AD* (Asymptotic Distribution II)**.** In addition to requirements of Assumption AD, $||\text{vec}(\text{off}(\boldsymbol{\Sigma}_{nT}))||_1 \leq c$ and $||\tilde{\boldsymbol{\Sigma}}_{nT}^{\frac{1}{2}}||_1 \leq c$, uniformly over n and T.

This assumption requires that the transformed covariance matrix $\tilde{\boldsymbol{\Sigma}}_{nT}$ is also bounded in absolute row and column sums, and limits further the degree of the dependence in

the errors. The condition $||\text{vec}(\text{off}(\boldsymbol{\Sigma}_{nT}))||_1 \leq c$ in particular is intimately related to the condition $T^3/n \to 0$ imposed in Theorem 1. When the model contains both a lagged outcome and neglected dependence in the error these two sources of temporal dependence may become conflated. Yet Theorem 1 demonstrates that as long as n grows sufficiently fast relative to T, neglected serial dependence in the error does not result in inconsistency or biasedness of the estimator. However, when $T/n \to c \geq 0$ this is no longer the case and it becomes necessary to further limit the degree of serial dependence in the error. Assumption AD* is also slightly stronger than would be necessary since it limits cross-sectional dependence as well as temporal dependence.[14] Naturally, Assumption AD* can be dispensed with in the absence of dynamics.

Before establishing the asymptotic distribution, it must, in the first instance, be established that the estimator remains consistent as $T/n \to c$ with potentially $c > 0$.

Proposition 3 (Consistency II)**.** *Under Assumptions MD, CS and AD*, as* $n, T \to \infty$

$$||\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0||_2 = \mathcal{O}_p\left(n^{-\frac{1}{3}}\right).$$

Proposition 3 establishes a rate of consistency when both n and T diverge. If T increases only very slowly with n, then it is known from Proposition 1 that the estimator is close to $\sqrt{n}$-consistent. On the other hand, in the large n, large T setting, when $T \propto n$, similar to Theorem 4.1 in Moon and Weidner (2015) it can be established that the estimator is also $\sqrt{n}$-consistent. Between these two scenarios the worst case order $n^{-\frac{1}{3}}$ in Proposition 3 is obtained. Notice also that this result does not require Assumption AE and therefore does not require the factors to be strong, nor indeed that the correct number of factors is known. Moreover, though this result is obtained under Assumption AD*, the assumption that the errors u_{it} are independent of the factors and the loadings would not be required.

Theorem 2 (Asymptotic Distribution II)**.** *Assume* $||\boldsymbol{c}||_2 = \mathcal{O}_p(1)$*. Under Assumptions MD, CS, AE, and AD*, as* $n, T \to \infty$ *with* $T/n \to c$ *and* $c \geq 0$*,*

$$\sqrt{nT}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0) + \boldsymbol{\Delta}^{-1}(\boldsymbol{\psi}^{(0)} + \boldsymbol{\psi}^{(1)} + \boldsymbol{\psi}^{(2)} + \boldsymbol{\psi}^{(3)}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{\Delta}^{-1}\boldsymbol{\Omega}\boldsymbol{\Delta}^{-1}), \tag{5.1}$$

where,

$$\begin{aligned}
\boldsymbol{\psi}^{(0)} &:= -\frac{1}{\sqrt{nT}}\text{tr}(\tilde{\boldsymbol{\Sigma}}_{nT}(\boldsymbol{G} \otimes \boldsymbol{M}_{\tilde{\boldsymbol{\Lambda}}^0}))\boldsymbol{\pi}_{K+1}, \\
\boldsymbol{\psi}^{(1)} &:= \frac{1}{\sqrt{nT}}\text{tr}(\tilde{\boldsymbol{\Sigma}}_{nT}((\boldsymbol{P}_{\boldsymbol{F}^0}\boldsymbol{G}\boldsymbol{M}_{\boldsymbol{F}^0} + \boldsymbol{G}\boldsymbol{P}_{\boldsymbol{F}^0}) \otimes \boldsymbol{I}_{TK}))\boldsymbol{\pi}_{K+1}, \\
\psi_\kappa^{(2)} &:= \frac{1}{\sqrt{nT}}\text{tr}(\tilde{\boldsymbol{\Sigma}}_{nT}(\boldsymbol{I}_T \otimes \boldsymbol{M}_{\tilde{\boldsymbol{\Lambda}}^0}\tilde{\boldsymbol{H}}_\kappa\boldsymbol{F}^0(\boldsymbol{F}^{0\top}\boldsymbol{F}^0)^{-1}(\tilde{\boldsymbol{\Lambda}}^{0\top}\tilde{\boldsymbol{\Lambda}}^0)^{-1}\tilde{\boldsymbol{\Lambda}}^{0\top})), \\
\psi_\kappa^{(3)} &:= \frac{1}{\sqrt{nT}}\text{tr}(\tilde{\boldsymbol{\Sigma}}_{nT}(\boldsymbol{F}^0(\boldsymbol{F}^{0\top}\boldsymbol{F}^0)^{-1}(\tilde{\boldsymbol{\Lambda}}^{0\top}\tilde{\boldsymbol{\Lambda}}^0)^{-1}\tilde{\boldsymbol{\Lambda}}^{0\top}\tilde{\boldsymbol{H}}_\kappa\boldsymbol{M}_{\boldsymbol{F}^0} \otimes \boldsymbol{I}_{TK})),
\end{aligned}$$

[14] Appendix B provides more discussion of this assumption.

and $\boldsymbol{\Delta} \coloneqq \boldsymbol{\Upsilon}^{(1)} + \boldsymbol{\Upsilon}^{(2)}$ *and* $\boldsymbol{\Omega} \coloneqq \boldsymbol{\Upsilon}^{(3)} + \boldsymbol{\Upsilon}^{(4)} + \boldsymbol{\Upsilon}^{(5)} + \boldsymbol{\Upsilon}^{(6)}$*, with*

$$\boldsymbol{\Upsilon}^{(1)} \coloneqq \frac{1}{nT}\tilde{\boldsymbol{\mathcal{H}}}^{\top}(\boldsymbol{M}_{\boldsymbol{F}^0} \otimes \boldsymbol{M}_{\tilde{\boldsymbol{\Lambda}}^0})\tilde{\boldsymbol{\mathcal{H}}},$$
$$\boldsymbol{\Upsilon}^{(2)} \coloneqq \frac{1}{nT}\boldsymbol{\pi}_{K+1}\mathrm{tr}(\tilde{\boldsymbol{\Sigma}}_{nT}(\boldsymbol{G}\boldsymbol{G}^{\top} \otimes \boldsymbol{I}_{TK}))\boldsymbol{\pi}_{K+1}^{\top},$$
$$\boldsymbol{\Upsilon}^{(3)} \coloneqq \frac{1}{nT}\tilde{\boldsymbol{\mathcal{H}}}^{\top}(\boldsymbol{M}_{\boldsymbol{F}^0} \otimes \boldsymbol{M}_{\tilde{\boldsymbol{\Lambda}}^0})\tilde{\boldsymbol{\Sigma}}_{nT}(\boldsymbol{M}_{\boldsymbol{F}^0} \otimes \boldsymbol{M}_{\tilde{\boldsymbol{\Lambda}}^0})\tilde{\boldsymbol{\mathcal{H}}},$$
$$\boldsymbol{\Upsilon}^{(4)} \coloneqq \frac{1}{nT}\boldsymbol{\pi}_{K+1}\mathrm{tr}(\tilde{\boldsymbol{\Sigma}}_{nT}(\boldsymbol{G} \otimes \boldsymbol{I}_{TK})\tilde{\boldsymbol{\Sigma}}_{nT}(\boldsymbol{G} \otimes \boldsymbol{I}_{TK})) + \mathrm{tr}(\tilde{\boldsymbol{\Sigma}}_{nT}(\boldsymbol{G} \otimes \boldsymbol{I}_{TK})\tilde{\boldsymbol{\Sigma}}_{nT}(\boldsymbol{G}^{\top} \otimes \boldsymbol{I}_{TK}))\boldsymbol{\pi}_{K+1}^{\top},$$
$$\boldsymbol{\Upsilon}^{(5)} \coloneqq (\boldsymbol{\Phi} + \boldsymbol{\Phi}^{\top}),$$
$$\boldsymbol{\Upsilon}^{(6)} \coloneqq \frac{1}{nT}\boldsymbol{\pi}_{K+1}\mathrm{tr}(\boldsymbol{\Psi}(\boldsymbol{\mathcal{M}}^{(4)} - 3\boldsymbol{I}_{nT})\boldsymbol{\Psi}^{\top})\boldsymbol{\pi}_{K+1}^{\top},$$

where $\boldsymbol{\Phi} \coloneqq (\tilde{\boldsymbol{\mathcal{H}}}^{\top}(\boldsymbol{M}_{\boldsymbol{F}^0} \otimes \boldsymbol{M}_{\tilde{\boldsymbol{\Lambda}}^0})(\tilde{\boldsymbol{\Sigma}}_{nT}^{\frac{1}{2}})^{\top}\boldsymbol{\mathcal{M}}^{(3)}\mathrm{diagv}((\tilde{\boldsymbol{\Sigma}}_{nT}^{\frac{1}{2}})^{\top}(\boldsymbol{G} \otimes \boldsymbol{I}_{TK})\tilde{\boldsymbol{\Sigma}}_{nT}^{\frac{1}{2}})\boldsymbol{\pi}_{K+1}^{\top}$*, the matrix* $\boldsymbol{\Psi} \coloneqq \mathrm{diag}((\tilde{\boldsymbol{\Sigma}}_{nT}^{\frac{1}{2}})^{\top}(\boldsymbol{G} \otimes \boldsymbol{I}_{TK})\tilde{\boldsymbol{\Sigma}}_{nT}^{\frac{1}{2}})$*,* $\upsilon_i^{(p)} \coloneqq \mathbb{E}[u_i^p]$*, and* $\boldsymbol{\mathcal{M}}^{(p)}$ *is an* $nT \times nT$ *diagonal matrix with diagonal elements* $\upsilon_1^{(p)}, \ldots, \upsilon_{nT}^{(p)}$*.*

The most significant difference when comparing (5.1) to (4.2) 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 indeed $\boldsymbol{\psi}^{(1)}$ is the analogue of the bias characterised in Moon and Weidner (2017). The bias $\boldsymbol{\psi}^{(0)}$, on the other hand, appears not to have been described previously in the literature since it only arises in situations when there is some dependence in the errors across time in addition to dynamics.[15] Indeed it is $\boldsymbol{\psi}^{(0)}$ that is ultimately responsible for the requirement $T^3/n \to 0$ in Theorem 1, with Assumption AD* serving to reduce the order of this term in order to obtain Theorem 2. Terms $\boldsymbol{\psi}^{(2)}$ and $\boldsymbol{\psi}^{(3)}$ arise due to cross-sectional and time series dependence and are almost exact duplicates of the expressions found in Bai (2009). However, when comparing the expressions for $\boldsymbol{\psi}^{(0)}, \boldsymbol{\psi}^{(1)}, \boldsymbol{\psi}^{(2)}$ and $\boldsymbol{\psi}^{(3)}$ to those that would be obtained for the untransformed model, there is one significant difference: the transformed cross-sectional covariance matrix $\tilde{\boldsymbol{\Sigma}}_{nT}$ has rank T^2K rather than nT. As a consequence, the order of these bias terms are:

	$\boldsymbol{\psi}^{(0)}$	$\boldsymbol{\psi}^{(1)}$	$\boldsymbol{\psi}^{(2)}$	$\boldsymbol{\psi}^{(3)}$
Original Model	$\boldsymbol{\mathcal{O}}_p\left(\sqrt{\frac{n}{T}}\right)$	$\boldsymbol{\mathcal{O}}_p\left(\sqrt{\frac{n}{T}}\right)$	$\boldsymbol{\mathcal{O}}_p\left(\sqrt{\frac{T}{n}}\right)$	$\boldsymbol{\mathcal{O}}_p\left(\sqrt{\frac{n}{T}}\right)$
Transformed Model	$\boldsymbol{\mathcal{O}}_p\left(\sqrt{\frac{T}{n}}\right)$	$\boldsymbol{\mathcal{O}}_p\left(\sqrt{\frac{T}{n}}\right)$	$\boldsymbol{\mathcal{O}}_p\left(\sqrt{\frac{T}{n}}\right)$	$\boldsymbol{\mathcal{O}}_p\left(\sqrt{\frac{T}{n}}\right)$

This reveals something fundamental: projection of the entire model into the time dimension of the panel does not make the incidental parameters in the cross-section disappear entirely, it instead shifts them into the time dimension, where they may

[15] Moon and Weidner (2017) assume the errors are independent over time.

interact with the extant problem in that dimension in complicated ways. Moreover, it also suggests that as long as the ratio $T/n \to 0$ the LS-IFE estimator applied to the transformed model will remain consistent and unbiased. This is in stark contrast to the same estimator applied to the original model where, although the LS-IFE estimator is consistent as $n, T \to \infty$, the analogues of $\boldsymbol{\psi}^{(0)}, \boldsymbol{\psi}^{(1)}$, and $\boldsymbol{\psi}^{(3)}$ would be explosive in probability and therefore would impede inference. This is formalised in the following result.

Corollary 1 (Small T)**.** *Under Assumptions MD, CS, AE, and AD*, as $n, T \to \infty$ with $T/n \to 0$*

$$\sqrt{nT}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{\Delta}_*^{-1}\boldsymbol{\Omega}_*\boldsymbol{\Delta}_*^{-1}),$$

where $\boldsymbol{\Delta}_* \coloneqq \boldsymbol{\Upsilon}^{(1)}$ and $\boldsymbol{\Omega}_* \coloneqq \boldsymbol{\Upsilon}^{(3)}$.

It is useful to conclude this discussion with a simple example to demonstrate visibly the core message of the preceding paragraph. Consider the case in which the errors are identically and independently distributed, and the true factors and loadings take the form of individual effects, that is,

$$\boldsymbol{\lambda}^0 \coloneqq \begin{pmatrix} \lambda_1^0 & \dots & \lambda_n^0 \end{pmatrix}^\top, \ \boldsymbol{F}^0 \coloneqq \boldsymbol{\iota}_T,$$

where $\boldsymbol{\iota}_T$ is a $T \times 1$ vector of ones. In this case $\boldsymbol{\psi}^{(2)} = \boldsymbol{\psi}^{(3)} = \mathbf{0}_{K+1}$ because $\boldsymbol{\Sigma}_{nT} \propto \boldsymbol{I}_{nT}$. Moreover, because $\boldsymbol{\Sigma}_{nT}$ is diagonal $\boldsymbol{\psi}^{(0)} = \mathbf{0}_{K+1}$, leaving the only remaining bias as $\boldsymbol{\psi}^{(1)}$. Since $\boldsymbol{P}_{\boldsymbol{F}^0} = \frac{1}{T}\boldsymbol{\iota}_T\boldsymbol{\iota}_T^\top$, this reduces to

$$\psi_1^{(1)} \coloneqq \frac{\sigma_0^2}{\sqrt{nT}}\frac{1}{T}\mathrm{tr}(\boldsymbol{P}_{\boldsymbol{\mathfrak{X}}})\mathrm{tr}(\boldsymbol{G}\boldsymbol{\iota}_T\boldsymbol{\iota}_T^\top).$$

A bit of algebra reveals that

$$\mathrm{tr}(\boldsymbol{G}\boldsymbol{\iota}_T\boldsymbol{\iota}_T^\top) = \sum_{t=1}^{T-1}\sum_{\tau=1}^{t}(\alpha^0)^{\tau-1} = \frac{T}{(1-\alpha^0)}\left(1 - \frac{1}{T}\frac{(1-(\alpha^0)^T)}{1-\alpha^0}\right). \tag{5.2}$$

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

$$\psi_1^{(1)} = \sqrt{\frac{T}{n}}\frac{K}{(1-\alpha)}\left(1 - \frac{1}{T}\frac{(1-\alpha^T)}{1-\alpha}\right). \tag{5.3}$$

Notice again the significance of the transformation $\boldsymbol{Q}_{\boldsymbol{\mathfrak{X}}}$ in reducing the rank of the covariance matrix to T^2K. Without the transformation $\mathrm{tr}(\boldsymbol{\Sigma}_{nT}) = \mathrm{tr}(\boldsymbol{I}_{nT}) = nT$, and

so

$$\psi_1^{(1)} = \sqrt{\frac{n}{T}}\frac{1}{(1-\alpha)}\left(1 - \frac{1}{T}\frac{(1-\alpha^T)}{1-\alpha}\right), \tag{5.4}$$

which matches (up to scale by $\sqrt{nT}$) the familiar expression derived in Nickell (1981). This again highlights the fact that transforming the model by $\boldsymbol{Q}_{\boldsymbol{\mathcal{X}}}$ 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 (5.3) and (5.4), it will likely manifest itself in similar ways.

6 Further Matter

6.1 A Method of Moments Interpretation

The estimation approach described in this paper can be recast in terms of moment conditions in which case the close kinship it shares with existing methods, in particular the quasi-difference estimator of Ahn et al. (2013), and the FIVU estimator of Robertson and Sarafidis (2015), becomes apparent. To see this suppose $\alpha^0 = 0$. Define $\boldsymbol{L}_i \coloneqq (\boldsymbol{I}_T \otimes \boldsymbol{l}_i)^\top$ where $\boldsymbol{l}_i$ is the $TK \times 1$ vector containing all the observations of the exogenous covariates for each i. Strict exogeneity of the error term implies the moment condition

$$\mathbb{E}\left[\boldsymbol{L}_i^\top \boldsymbol{\varepsilon}_i\right] = \mathbb{E}\left[\boldsymbol{L}_i^\top \left(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta} - (\boldsymbol{I}_T \otimes \boldsymbol{\lambda}_i^\top)\mathrm{vec}(\boldsymbol{F}^\top)\right)\right] = \boldsymbol{0}_{T^2K},$$

where $\boldsymbol{\varepsilon}_i$ is the $T \times 1$ vector of error terms associated with each i. Now, let

$$\mathbb{E}\left[\boldsymbol{L}_i^\top(\boldsymbol{\lambda}_i^\top \otimes \boldsymbol{I}_T)\right] = \mathbb{E}\left[(\boldsymbol{I}_T \otimes \boldsymbol{l}_i\boldsymbol{\lambda}_i^\top)\right] =: (\boldsymbol{I}_T \otimes \boldsymbol{\Psi}),$$

where the $TK \times R$ matrix $\boldsymbol{\Psi}$ captures correlation between the covariates and the factor loadings. The two equations above give rise to the finite sample analogue

$$\boldsymbol{\varphi}(\boldsymbol{\beta}, \boldsymbol{\Psi}, \boldsymbol{F}) \coloneqq \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{L}_i^\top(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}) - \mathrm{vec}(\boldsymbol{\Psi}\boldsymbol{F}^\top).$$

In the framework described by Robertson and Sarafidis (2015), the set of moment conditions above coincides with those that would occur under strict exogeneity. Therefore minimising the following unweighted GMM objective function gives rise to their minimum distance FIVU estimator

$$\mathcal{Q}_{FIVU}(\boldsymbol{\beta}, \boldsymbol{\Psi}, \boldsymbol{F}) =: \boldsymbol{\varphi}^\top(\boldsymbol{\beta}, \boldsymbol{\Psi}, \boldsymbol{F})\boldsymbol{\varphi}(\boldsymbol{\beta}, \boldsymbol{\Psi}, \boldsymbol{F}). \tag{6.1}$$

The parameter $\boldsymbol{\Psi}$ can be profiled out of this objective function which, after some algebra, gives rise to an alternative GMM estimator that is equivalent to minimising

$$\mathcal{Q}_{QD}(\boldsymbol{\beta}, \boldsymbol{F}) \coloneqq \boldsymbol{\varphi}_*^\top(\boldsymbol{\beta}, \boldsymbol{F})\boldsymbol{\varphi}_*(\boldsymbol{\beta}, \boldsymbol{F}) \tag{6.2}$$

where the vector of moment conditions is

$$\boldsymbol{\varphi}_*(\boldsymbol{\beta}, \boldsymbol{F}) \coloneqq \left(\frac{1}{n}\sum_{i=1}^{n}\left(\boldsymbol{\mathcal{V}}^\top(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}) \otimes \boldsymbol{l}_i\right)\right),$$

with the $T \times (T-R)$ matrix $\boldsymbol{\mathcal{V}}$ forming an orthonormal basis for the null space associated with the columns of $\boldsymbol{F}$.[16] This second vector of moment conditions coincides with that which would arise in the framework of Ahn et al. (2013) under strict exogeneity and thus minimising $\mathcal{G}_{QD}(\boldsymbol{\beta}, \boldsymbol{F})$ over $(\boldsymbol{\beta}, \boldsymbol{F})$ produces an unweighted version of the quasi-difference estimator described in that paper. Going further it is possible to wholly concentrate the factors $\boldsymbol{F}$ out of the objective function in (6.2) at which point one arrives at a univariate objective function which (up to scale) coincides with $\mathcal{Q}(\boldsymbol{\beta})$ given in (2.7). This reveals that, at least in some cases, there is an intimate relationship between the moment based estimation approaches of Ahn et al. (2013) and Robertson and Sarafidis (2015) and the LS-IFE estimator, with it sometimes being possible to reduce the former two estimation problems down to the optimisation of a univariate objective function.[17] Moreover, there is also the suggestion that the fixed T consistency and unbiasedness of those moment biased approaches will, under certain conditions, also carry over to the large n, large T setting, as long as $T/n \to 0$.[18] Of course when the model is dynamic then these approaches will no longer coincide. Nonetheless, this paper has demonstrated that it is in fact still possible to derive a set of moment conditions for the dynamic model which can be reduced down to a univariate optimisation problem by way of principle components.

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.

[16] To be clear $\boldsymbol{\mathcal{V}}$ is a function of $\boldsymbol{F}$, though this is suppressed for ease of notation.

[17] These various methods can, in turn, be related to the procedure suggested by Chamberlain (1984) for short panels with individual effects. In the present context, this could be understood as decomposing $\boldsymbol{\Lambda} = \boldsymbol{P}_{\boldsymbol{\mathcal{X}}}\boldsymbol{\Lambda} + \boldsymbol{M}_{\boldsymbol{\mathcal{X}}}\boldsymbol{\Lambda} =: \boldsymbol{\mathcal{X}}\boldsymbol{\Psi} + \boldsymbol{e}$, where $\boldsymbol{\Psi}$ is a $TK \times R$ parameter to be estimated, and $\boldsymbol{e}$ is subsumed into the error term.

[18] The relationship between these approaches also give an interesting interpretation to the transformed LS-IFE estimator, showing that the method can also be viewed as the solution to the principle component problem that involves a matrix of moment conditions rather than the covariance matrix the error $\boldsymbol{\Lambda}\boldsymbol{F}^\top + \boldsymbol{\varepsilon}$ as would conventionally be the case. I am grateful to a reader for pointing this out.

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 $l = 1, \ldots, K_{\mathrm{L}}$, are those associated with low rank regressors, and the remaining K_{H} coefficients, indexed $h = 1, \ldots, K_{\mathrm{H}}$, 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 result extends 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 $\boldsymbol{\mathcal{V}} \coloneqq (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{K_{\mathrm{L}}})$ and $\boldsymbol{\mathcal{W}} \coloneqq (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_{K_{\mathrm{L}}})$. 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 $(\boldsymbol{\mathcal{V}}, \boldsymbol{X}_1, \ldots, \boldsymbol{X}_{K_{\mathrm{H}}})$ to ensure that $\boldsymbol{\mathcal{X}}^\top \boldsymbol{\mathcal{X}}$ is invertible. Let $\tilde{\boldsymbol{\mathcal{V}}} \coloneqq \boldsymbol{Q}_{\boldsymbol{\mathcal{X}}}^\top \boldsymbol{\mathcal{V}}$ and $\boldsymbol{\delta}_{\mathrm{H}} \cdot \tilde{\boldsymbol{Z}}_{\mathrm{H}} \coloneqq \sum_{\kappa=1}^{K_{\mathrm{H}}} \delta_\kappa \tilde{\boldsymbol{Z}}_\kappa$.

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 $c > 0$ 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

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

Since 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 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 4 (Consistency – Low Rank)**.** *Under Assumptions MD, AE, AD*, and LR, as* $n \to \infty$ *and* $T/n \to 0$

$$||\hat{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta}^0||_2 = \mathcal{O}_p(1).$$

6.3 Estimating the Number of Factors

The result established in Section 3 demonstrates 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 when T is small relative to n.

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 might also be ported to the small T setting, 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

$$\mu_r^* \coloneqq \mu_r \left(\frac{1}{nT} \left(\tilde{\boldsymbol{Y}} - \sum_{\kappa=1}^{K+1} \hat{\theta}_\kappa \tilde{\boldsymbol{Z}}_\kappa \right)^\top \left(\tilde{\boldsymbol{Y}} - \sum_{\kappa=1}^{K+1} \hat{\theta}_\kappa \tilde{\boldsymbol{Z}}_\kappa \right) + \frac{T}{n} \boldsymbol{I}_T \right), \tag{6.3}$$

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

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

The main modification here from Ahn and Horenstein (2013)'s original specification is the addition of the matrix $\frac{T}{n}\boldsymbol{I}_T$.

Proposition 5. *MD, CS and AE as $n \to \infty$ and $T/n \to 0$,*

$$\Pr\left(\max_{1 \leq r \leq T} \mu_r^* = R^0 \right) \to 1. \tag{6.4}$$

6.4 Alternative Approach

An alternative estimation approach to that studied in this paper treats the initial condition, which has a rank of 1, as an additional factor. Though this increases the number of factors, it also preserves another time period of data. It is useful to compare the

performance of the two approaches.[19] More specifically, with $T^c \coloneqq T+1$

$$\begin{aligned}\boldsymbol{Y}^c\boldsymbol{S}^c(\alpha) &= \sum_{k=1}^{K}\beta_k\boldsymbol{X}_k^c + \alpha\boldsymbol{y}_{-1}\boldsymbol{\pi}_{T+1}^\top + \boldsymbol{\Lambda}\boldsymbol{F}^{c\top} + \varepsilon \\ &= \sum_{k=1}^{K}\beta_k\boldsymbol{X}_k^c + \boldsymbol{\Lambda}^*\boldsymbol{F}^{*c\top},\end{aligned}$$

with $\boldsymbol{S}^c(\alpha) \coloneqq \boldsymbol{I}_{T+1} - \alpha\boldsymbol{W}^c$, $\boldsymbol{\Lambda}^* \coloneqq (\boldsymbol{y}_{-1}, \boldsymbol{\Lambda})$ and $\boldsymbol{F}^{c*} \coloneqq (\alpha\boldsymbol{\pi}_{T+1}, \boldsymbol{F}^c)$, and where the matrices with superscript c now include an additional column relative to previous sections. Then, using $\sim$ to indicate transformed variables as previously, consider the alternate objective function

$$Q^c(\boldsymbol{\theta}) \coloneqq \frac{1}{nT}\mathrm{tr}\left(\left(\boldsymbol{S}^c(\alpha)\tilde{\boldsymbol{Y}}^c - \sum_{\kappa=1}^{K+1}\theta_\kappa\tilde{\boldsymbol{Z}}_\kappa^c - \tilde{\boldsymbol{\Lambda}}^*\boldsymbol{F}^{*c\top}\right)^\top\left(\boldsymbol{S}^c(\alpha)\tilde{\boldsymbol{Y}}^c - \sum_{\kappa=1}^{K+1}\theta_\kappa\tilde{\boldsymbol{Z}}_\kappa^c - \tilde{\boldsymbol{\Lambda}}^*\boldsymbol{F}^{*c\top}\right)\right).$$

The alternative estimator $\hat{\boldsymbol{\theta}}^c$ may then be defined as

$$\hat{\boldsymbol{\theta}}^c \coloneqq \underset{\boldsymbol{\theta}\in\Theta}{\arg\min}\, Q^c(\boldsymbol{\theta}).$$

This estimator retains all of the essential properties of $\hat{\boldsymbol{\theta}}$, including fixed T consistency and a similar asymptotic distribution.

6.5 Standard Normality

Owing to the fact that the standard normal distribution is invariant to orthogonal transformations, especially favourable rates of consistency with $R \geq R^0$ can be achieved. Key to showing this is the following result.

Lemma 1. *Assume that the elements of the $n \times T$ matrix $\boldsymbol{\varepsilon}$ are standard normal and independent across i and t. Then $||\tilde{\boldsymbol{\varepsilon}}||_2 = \mathcal{O}_p(\sqrt{T})$.*

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

Using this result the following is obtained.

Proposition 6. *Under the Assumptions of Lemma 1, Assumptions MD and CS,*

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

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

[19] This alternative approach has been studied more extensively in previous versions of this paper.

7 Monte Carlo Simulations

This section provides simulation results which highlight the different properties of the LS-IFE estimator when applied to the original model, and to the transformed model. 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 $(\boldsymbol{\Sigma}_n^{\frac{1}{2}} \otimes \boldsymbol{\Sigma}_T^{\frac{1}{2}})\text{vec}(\boldsymbol{U})$, 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 naive LS estimator (Naive), the LS-IFE estimator applied to the original model (IFE), the LS-IFE estimator applied to the transformed model (Q-IFE-1), and approach described in Section 6.4 (Q-IFE-2).

Table 1a: Bias (SE) α

	Naive			IFE			Q-IFE-1			Q-IFE-2		
$n \setminus T$	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)
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

	Naive			IFE			Q-IFE-1			Q-IFE-2		
$n \setminus T$	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)
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

	Naive			IFE			Q-IFE-1			Q-IFE-2		
$n \setminus T$	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)
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 Naive 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 LS-IFE 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 bias in the autoregressive coefficient. As expected, both the Q-IFE-1 and the Q-IFE-2 estimates of α are unbiased as n increases. For the coefficient β_1, the Naive estimates are severely biased, with this bias being persistent irrespective of n and T. For small T, the LS-IFE estimates are also biased, which stems from the heteroskedasticity of the errors in both the cross-section and across time. Owing to the significant heteroskedasticity in the design, when both n and T are small, the Q-IFE-1 and Q-IFE-2 estimates have sizeable biases - though smaller in magnitude than the Naive or IFE estimates. 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 – 2c 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 %

	Naive			IFE			Q-IFE-1			Q-IFE-2		
$n \setminus T$	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
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 %

	Naive			IFE			Q-IFE-1			Q-IFE-2		
$n \setminus T$	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
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 2c: Coverage β_2 %

	Naive			IFE			Q-IFE-1			Q-IFE-2		
$n \setminus T$	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
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 the Naive estimates remains consistently below its nominal value, while for the IFE estimates 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 the Q-IFE-1 and Q-IFE-2 estimates improve as n increases, with T fixed or T increasing slowly. The story is similar for β_1 in Table 2b. The coverage of Naive estimates is incredibly poor, presenting near 0 across all n, T values. The coverage of the IFE estimates is also poor with either n or T small, and improves only as both of these increase. The Q-IFE-1 and Q-IFE-2 estimates present poor coverage with both n and T small, yet these rapidly improve as n increases. When comparing the performance of Q-IFE-1 and Q-IFE-2, it is, in general, the case that Q-IFE-1 outperforms Q-IFE-2. This is for two reasons. The first is that, while omitting a time period, Q-IFE-1 uses only 2 factors in estimation, whereas Q-IFE-2 uses 3, with the extra factor being present to control for a possibly endogenous initial condition. The second reason is that an approach which includes the lagged outcome as an additional regressor on the right-hand side of the outcome equation, benefits from additional variation coming from pre-sample exogenous covariates transmitted through the initial condition.[20] This variation is lost when treating the initial condition as an additional factor. As $T \to \infty$ alongside n, the impact of the initial condition, whichever way it is treated, becomes negligible and it is unclear which estimation approach should perform best. Furthermore, is it not necessarily clear what the correct number of factors to input into the estimation routine should be for Q-IFE-2 when both n and T are large. It is therefore 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 Q-IFE-2 %

	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
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 Q-IFE-2 residuals. Only in the smallest sample size, $n = 30$, $T = 6$, is the number of factors chosen to be 3. This suggests that the impact of the initial condition becomes quickly negligible. In light of this, Tables 4a and 4b below present bias and coverages for Q-IFE-2 with $R = 2$. Comparing these results to those presented previously, these estimates are generally better than Q-IFE-1. However, Q-IFE-1 still outperforms Q-IFE-2 when it comes to the autoregressive parameter α.

[20] This pre-sample variation is the origin of $\boldsymbol{H}_0$; see Section 4.

Table 4a: Bias (SE), Q-IFE-2 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)
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: Coverage Q-IFE-2 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
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 has introduced a new, simple method to estimate linear panel data models with interactive fixed effects designed for situations where T is small relative to n. By transforming the model and then applying the LS-IFE 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 asymptotically unbiased in the presence of cross-sectional dependence, serial dependence, and with the inclusion of dynamic regressors. Under certain conditions, these results have also been shown to extend settings where T grows with n, so long as the ratio $T/n \to 0$. This stands in contrast to the usual case where the LS-IFE estimator generally delivers inconsistent estimates with T fixed, and suffers from biases of the order n/T and T/n as $n, T \to \infty$. Careful study of this estimation approach has also revealed interesting connections between the LS-IFE estimator and several method of moments-based approaches, bridging the gap between what are, at present, two quite separate literatures. In particular, adapting the former approach to short panels has provided new insights on the properties of the latter group of approaches when applied to large panels. Several other consequences of this approach have also been 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 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 ob-

tained as a marginal likelihood associated with a maximal invariant statistic under the group of transformations (2.2). 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 (2021). The second extension is to incorporate more general predetermined regressors which are intuitively difficult to handle in this framework. Exploiting the established connections between the LS-IFE estimator and method of moments-based approaches may prove especially useful in this regard. Finally, it is worth stressing that arguably the most powerful concept developed in this paper 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., 2001. GMM estimation of linear panel data models with time-varying individual effects. Journal of Econometrics 101 (2), 219 – 255.

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.

Bai, J., Li, K., 2014. Theory and methods of panel data models with interactive effects. Annals of Statistics 42 (1), 142 – 170.

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., 2021. Likelihood inference and the role of initial conditions for the dynamic panel data model. Journal of Econometrics 221 (1), 160–179.

Bernstein, D. S., 2009. Matrix mathematics: theory, facts, and formulas. Princeton University Press, Princeton, New Jersey, USA.

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

Chamberlain, G., Moreira, M. J., 2009. Decision theory applied to a linear panel data model. Econometrica 77 (1), 107–133.

Everaert, G., Groote, T., 2016. Common correlated effects estimation of dynamic panels with cross-sectional dependence. Econometric Reviews 35 (3), 428 – 463.

Grenander, U., Szegö, G., 1955. Toeplitz Forms and Their Applications.

Holtz-Eakin, D., Newey, W., Rosen, H. S., 1988. Estimating vector autoregressions with panel data. Econometrica 56 (6), 1371–1395.

Hsiao, C., Shi, Z., Zhou, Q., 2021. Transformed estimation for panel interactive effects models. Journal of Business & Economic Statistics 0 (0), 1–18.

Juodis, A., Sarafidis, V., 2022a. An incidental parameters free inference approach for panels with common shocks. Journal of Econometrics 229 (1), 19–54.

Juodis, A., Sarafidis, V., 2022b. A linear estimator for factor-augmented fixed-T panels with endogenous regressors. Journal of Business & Economic Statistics 40 (1), 1–15.

Kato, T., 1980. Perturbation Theory for Linear Operators. Springer-Verlag, Berlin, Germany.

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.

Robertson, D., Sarafidis, V., 2015. IV estimation of panels with factor residuals. Journal of Econometrics 185 (2), 526–541.

Vos, I. D., Everaert, G., 2021. Bias-corrected common correlated effects pooled estimation in dynamic panels. Journal of Business & Economic Statistics 39 (1), 294–306.

Westerlund, J., 2020. A cross-section average-based principal components approach for fixed-T panels. Journal of Applied Econometrics 35 (6), 776–785.

Westerlund, J., Urbain, J., 2015. Cross-sectional averages versus principal components. Journal of Econometrics 185 (2), 372 – 377.

Yu, J., de Jong, R., fei Lee, L., 2008. Quasi-maximum likelihood estimators for spatial dynamic panel data with fixed effects when both n and t are large. Journal of Econometrics 146 (1), 118–134.