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

Ayden Higgins[†]
University of Cambridge

January 31, 2022

Abstract

This paper studies the estimation of a linear panel data model with interactive fixed effects, where one dimension of the panel, typically time, is fixed. To this end, a novel 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, in the central result of this paper, it is demonstrated that simply transforming the model and then applying the least squares interactive fixed effects (LS-IFE) estimator of Bai (2009) will deliver consistent estimates of regression slope coefficients with T fixed. Moreover, these estimates are shown to be asymptotically unbiased even where the error exhibits cross-sectional or temporal dependence and/or heteroskedasticity, and with the inclusion of dynamic regressors. This contrasts sharply with the usual case where the LS-IFE estimator is, in general, both inconsistent and biased with T fixed.

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: University of Cambridge, Faculty of Economics, Austin Robinson Building, Sidgwick Avenue, Cambridge CB3 9DD, United Kingdom. E-mail: amh239@cam.ac.uk.

1 Introduction

This paper contributes to the extensive literature on linear panel data models with interactive effects. These models have proven to be very popular since, in many situations, the existence of such structures is well motivated; for example, arising due to unobserved heterogeneity across individuals, or exposure to common shocks. The model studied in this paper assumes that, in a panel with entries indexed $i = 1, \dots, n$ 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 covariates, $\boldsymbol{\Lambda}^*$ is an $n \times R^*$ matrix of time-invariant factor loadings, and $\boldsymbol{f}_t^*$ is an $R^* \times 1$ vector of time-varying factors. It is assumed that both the outcomes and the covariates are observed by the econometrician, while the factors, the loadings, and the error terms are not. The parameter of interest in this model is the $(K+1) \times 1$ vector $\boldsymbol{\theta} := (\alpha, \boldsymbol{\beta}^\top)^\top$ comprised of the scalar autoregressive parameter α and the $K \times 1$ vector $\boldsymbol{\beta}$.

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

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

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 biased 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, 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 significant 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 has serious repercussions for 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 LS-IFE estimator which treats the factors and loadings as interactive fixed effects is consistent as both n and $T \to \infty$, though it typically suffers form an 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 LS-IFE estimator is, in general, both inconsistent and biased outside of exceptional circumstances.

Though perhaps unsurprising, the fixed T inconsistency of the LS-IFE estimator is unfortunate since this estimator is both intuitive and incredibly easy to implement. There are, however, alternative methods that can be applied to short panels including Pesaran (2006), Ahn et al. (2013), Hayakawa (2012) and Bai and Li (2014). Broadly, these methods can be grouped into two types: common correlated effects (CCE) approaches (Pesaran, 2006; Bai and Li, 2014), and quasi-difference (QD) approaches (Ahn et al., 2013; Hayakawa, 2012). The CCE approach, introduced in Pesaran (2006), is predicated on the assumption that the latent factors in the error term also impact some model covariates, such that the factors can be instrumented by cross-sectional averages. In such cases, these instruments can be levered to purge the factor term and, as a result, give rise to estimators that are often consistent with T fixed, as well as where both n and T diverge. Though often readily implementable, the drawback to the CCE approach, however, is that it relies on a particular functional form for the relationship between the latent factors and model covariates which, especially in microeconometric settings, may not be easy to justify. The QD approach, on the other hand, takes advantage of the inherent indeterminacy associated with factor models, to normalise the factors and loadings in a certain way, such that the model can be multiplied by a difference matrix to purge the factor term. This essentially exchanges the problem of estimating the factors and the loadings for estimating the difference matrix, which can be specified in such a way that its dimension stays fixed as $n \to \infty$, making this approach suitable for fixed T estimation. This is most often done using GMM which has the added benefit of providing a straightforward way to handle predetermined regressors by specifying appropriate moment conditions. However, the moment conditions this approach yields are highly nonlinear, and generate a difficult optimisation problem which may quickly become infeasible when T extends beyond a few observations. Moreover, Hayakawa (2016) has pointed out that several QD approaches may fail to satisfy the identification conditions (which are a necessary precursor to consistency) for GMM estimation, which arises as a consequence of the particular normalisation typically imposed on the factor term.

Notwithstanding these contributions, it is clear that there remains scope for a general and simple to implement method for the estimation of linear panel data models with interactive fixed effects where the time dimension is fixed. This paper seeks to resolve this and introduces a novel transformation of the model which, after having been applied, enables the LS-IFE estimator to produce consistent estimates with T fixed. In contrast to other 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. Perhaps 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 simply transforming the model and then applying the LS-IFE estimator will produce consistent and 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. This contrasts sharply with the usual case where the LS-IFE estimator is both inconsistent and biased, in general, with T fixed. In order to fully appreciate why this estimation approach works, it is useful to also study its large n, large T properties. The main message in this regard is that consistency and asymptotic unbiasedness will also carry over to the large n, large T setting, as long as the ratio $T/n \to 0$. Moreover, outside of estimation, this paper shows that with minor modification, the eigenvalue ratio test of Ahn and Horenstein (2013) can be used to detect the number of factors in the transformed model with T fixed.

Outline: Section 2 sets out the estimation approach, introducing the transformation and discussing 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 result in Section 5. Some additional considerations are collected in Section 6 including detection of the number of factors and alternative approaches to treating the initial condition. Monte Carlo simulations follow in Section 7. Section 8 concludes. Additional discussion and 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, $||\boldsymbol{a}||_1 := \sum_{i=1}^{n} |a_i|$, $||\boldsymbol{a}||_2 := \sqrt{\sum_{i=1}^{n} a_i^2}$, $||\boldsymbol{a}||_\infty := \max_{1\leq i\leq n} |a_i|$. Let $\boldsymbol{A}$ be an $n \times m$ matrix with elements A_{ij}. When $m = n$, and the eigenvalues of $\boldsymbol{A}$ are real, they are denoted $\mu_{\min}(\boldsymbol{A}) := \mu_n(\boldsymbol{A}) \leq \ldots \leq \mu_1(\boldsymbol{A}) =: \mu_{\max}(\boldsymbol{A})$. The following matrix norms are those induced by their vector counterparts: $||\boldsymbol{A}||_1 := \max_{1\leq j\leq m} \sum_{i=1}^{n} |A_{ij}|$ which is the maximum absolute column sum of $\boldsymbol{A}$, $||\boldsymbol{A}||_2 := \sqrt{\mu_{\max}(\boldsymbol{A}^\top \boldsymbol{A})}$, and $||\boldsymbol{A}||_\infty := \max_{1\leq i\leq n} \sum_{j=1}^{m} |A_{ij}|$ which is the maximum absolute row sum of $\boldsymbol{A}$. The Frobenius norm of $\boldsymbol{A}$ is denoted $||\boldsymbol{A}||_F := \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{m} A_{ij}^2} = \sqrt{\mathrm{tr}(\boldsymbol{A}^\top \boldsymbol{A})}$. Let $\boldsymbol{P}_{\boldsymbol{A}} := \boldsymbol{A}(\boldsymbol{A}^\top \boldsymbol{A})^+ \boldsymbol{A}^\top$ and $\boldsymbol{M}_{\boldsymbol{A}} := \boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{A}}$, where $\boldsymbol{I}_n$ is the $n \times n$ identity matrix and $+$ denotes the Moore-Penrose generalised inverse. An $n \times 1$ vector of ones is denoted $\boldsymbol{\iota}_n$, and an $n \times m$ matrix of zeros is denoted $\boldsymbol{0}_{n\times m}$. 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.

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 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 it is possible to resolve the fixed T inconsistency of the LS estimator, and to produce estimates that are consistent and asymptotically unbiased as $n \to \infty$.

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

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

where $\boldsymbol{s}(\alpha) := (\alpha, \boldsymbol{0}_{1\times(T-1)})^\top$. In any dynamic panel model where T is small, special care must be taken with the initial condition $\boldsymbol{y}_0 \boldsymbol{s}^\top(\alpha)$ since this may itself be endogenous. In what follows the initial condition is treated as an additional parameter in the model and is absorbed into the factor term.[1] As such, define $\boldsymbol{\Lambda} := (\boldsymbol{y}_0, \boldsymbol{\Lambda}^*)$, $\boldsymbol{F}(\alpha) := (\boldsymbol{s}(\alpha), \boldsymbol{F}^*)$, and $R := R^* + 1$, whereby (2.1) becomes

$$\boldsymbol{Y}\boldsymbol{S}(\alpha) = \sum_{k=1}^{K} \beta_k \boldsymbol{X}_k + \boldsymbol{\Lambda}\boldsymbol{F}^\top + \boldsymbol{\varepsilon}, \tag{2.2}$$

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

$$\mathcal{G} := \{\boldsymbol{Q} \in \mathcal{O}(n) : \boldsymbol{Q}\mathfrak{X} = \mathfrak{X}\}, \tag{2.3}$$

[1] An alternative approach which involves conditioning on the first observation is discussed in Section 6.2.

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

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

$$\begin{aligned}
\tilde{\boldsymbol{Y}} &:= \boldsymbol{Q}_{\mathfrak{X}}^{\top}\boldsymbol{Y}, \\
\tilde{\boldsymbol{X}}_k &:= \boldsymbol{Q}_{\mathfrak{X}}^{\top}\boldsymbol{X}_k, \\
\tilde{\boldsymbol{\Lambda}} &:= \boldsymbol{Q}_{\mathfrak{X}}^{\top}\boldsymbol{\Lambda}, \\
\tilde{\boldsymbol{\varepsilon}} &:= \boldsymbol{Q}_{\mathfrak{X}}^{\top}\boldsymbol{\varepsilon},
\end{aligned}$$

in which case premultiplying (2.2) by $\boldsymbol{Q}_{\mathfrak{X}}^{\top}$ yields the transformed model

$$\tilde{\boldsymbol{Y}}\boldsymbol{S}(\alpha) = \sum_{k=1}^{K} \beta_k \tilde{\boldsymbol{X}}_k + \tilde{\boldsymbol{\Lambda}}\boldsymbol{F}^{\top} + \tilde{\boldsymbol{\varepsilon}}. \tag{2.4}$$

Looking at (2.4) there are three significant consequences of transforming the model through $\boldsymbol{Q}_{\mathfrak{X}}$ that need to be highlighted. First, the resultant matrices $\tilde{\boldsymbol{Y}}$, $\tilde{\boldsymbol{X}}_k$ and $\tilde{\boldsymbol{\Lambda}}\boldsymbol{F}^{\top}$ are of dimension $TK \times T$, since the entirety of the model has been transformed by $\boldsymbol{Q}_{\mathfrak{X}}$ and projected into the TK-dimensional subspace spanned by the columns of the covariates. Hence, the dimension of the factor term $\tilde{\boldsymbol{\Lambda}}\boldsymbol{F}^{\top}$ will no longer depend on n, thereby reliving the model of incidental parameters as $n \to \infty$.[3] Second, the transformation leads to no loss of information in the covariates since, by construction, transforming the model through $\boldsymbol{Q}_{\mathfrak{X}}$ preserves the column space of $\mathfrak{X}$. Thirdly, since the covariates used in the construction of $\boldsymbol{Q}_{\mathfrak{X}}$ are strictly exogenous, under quite general conditions, including broad cross-sectional and serial dependence, the transformation serves to reduce the order of the error term which, ultimately, is key to estimating (2.4) using the LS-IFE

[3] 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.

estimator.[4]

2.2 Principal Components

The underlying mechanics of the LS-IFE estimator are perhaps 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 of its transformed counterpart.

Assume that $\boldsymbol{\theta}$ is observed and define $\dot{\boldsymbol{Y}} := \boldsymbol{Y}\boldsymbol{S}(\alpha) - \sum_{k=1}^{K} \beta_k \boldsymbol{X}_k = \boldsymbol{\Lambda}\boldsymbol{F}^\top + \boldsymbol{\varepsilon}$ which has a pure factor structure. Let $\check{\boldsymbol{\Lambda}}$ and $\check{\boldsymbol{F}}$ be $n \times R$ and $T \times R$ matrices, respectively, which satisfy $\check{\boldsymbol{\Lambda}}\check{\boldsymbol{F}}^\top = \boldsymbol{\Lambda}\boldsymbol{F}^\top$, $\frac{1}{n}\check{\boldsymbol{\Lambda}}^\top\check{\boldsymbol{\Lambda}} = \boldsymbol{I}_R$ and $\check{\boldsymbol{F}}^\top\check{\boldsymbol{F}}$ being diagonal.[5] 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 $n \to \infty$,

$$\frac{1}{nT}\dot{\boldsymbol{Y}}\dot{\boldsymbol{Y}}^\top\check{\boldsymbol{\Lambda}} = \frac{1}{nT}\check{\boldsymbol{\Lambda}}\check{\boldsymbol{F}}^\top\check{\boldsymbol{F}} + \frac{1}{nT}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top\check{\boldsymbol{\Lambda}} + o_p(1)\,. \tag{2.5}$$

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 $||\boldsymbol{\varepsilon}||_2 = O_p(\sqrt{\max\{n,T\}})$ employed in Moon and Weidner (2015), dependence in the error term is sufficiently limited that $\frac{1}{nT}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top\check{\boldsymbol{\Lambda}} = o_p(1)$ 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

[4] 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 $\mathfrak{X}$.

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

case generally applies to the situation where T is fixed.

Now consider, on the other hand, the transformed model. Let $\check{\tilde{\boldsymbol{\Lambda}}}$ denote an analogue of $\check{\boldsymbol{\Lambda}}$. With $\frac{1}{nT}||\tilde{\boldsymbol{\varepsilon}}\boldsymbol{F}\tilde{\boldsymbol{\Lambda}}^\top\check{\tilde{\boldsymbol{\Lambda}}}||_2 = \mathcal{O}_p\left(1\right)$, one arrives at a similar expression to (2.5),

$$\frac{1}{nT}\dot{\tilde{\boldsymbol{Y}}}\dot{\tilde{\boldsymbol{Y}}}^\top\check{\tilde{\boldsymbol{\Lambda}}} = \frac{1}{nT}\check{\tilde{\boldsymbol{\Lambda}}}\check{\boldsymbol{F}}^\top\check{\boldsymbol{F}} + \frac{1}{nT}\tilde{\boldsymbol{\varepsilon}}\tilde{\boldsymbol{\varepsilon}}^\top\check{\tilde{\boldsymbol{\Lambda}}} + \boldsymbol{\mathcal{O}}_p\left(1\right). \tag{2.6}$$

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

2.3 Objective Function

The transformed model (2.4) can be estimated by minimising the following least squares objective function:

$$\begin{aligned}&\mathcal{Q}(\boldsymbol{\theta}, \tilde{\boldsymbol{\Lambda}}, \boldsymbol{F})\\ &:= \frac{1}{nT}\mathrm{tr}\left(\left(\tilde{\boldsymbol{Y}}\boldsymbol{S}(\alpha) - \sum_{k=1}^{K}\beta_k\tilde{\boldsymbol{X}}_k - \tilde{\boldsymbol{\Lambda}}\boldsymbol{F}^\top\right)^\top\left(\tilde{\boldsymbol{Y}}\boldsymbol{S}(\alpha) - \sum_{k=1}^{K}\beta_k\tilde{\boldsymbol{X}}_k - \tilde{\boldsymbol{\Lambda}}\boldsymbol{F}^\top\right)\right).\end{aligned}\tag{2.7}$$
[6]

Both the factors and the transformed loadings can be concentrated out of (2.7), in which case one arrives at an objective function involving $\boldsymbol{\theta}$ alone,

$$\mathcal{Q}(\boldsymbol{\theta}) := \frac{1}{nT}\sum_{r=R+1}^{T}\mu_r\left(\left(\tilde{\boldsymbol{Y}}\boldsymbol{S}(\alpha) - \sum_{k=1}^{K}\beta_k\tilde{\boldsymbol{X}}_k\right)^\top\left(\tilde{\boldsymbol{Y}}\boldsymbol{S}(\alpha) - \sum_{k=1}^{K}\beta_k\tilde{\boldsymbol{X}}_k\right)\right), \tag{2.8}$$

that is, the profile objective function now involves the sum of the $(T - R)$ smallest eigenvalues of the right hand-side matrix.[7] Using this, the estimator $\hat{\boldsymbol{\theta}}$ can then

[6] When estimating the original model, the least squares objective function can be interpreted as the negative of a quasi-maximum likelihood function that uses the standard normal distribution. The objective function (2.7) can then be interpreted as a marginal quasi-likelihood which uses only a part of this.

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

be defined as

$$\hat{\boldsymbol{\theta}} := \arg\min_{\boldsymbol{\theta}\in\Theta} \mathcal{Q}(\boldsymbol{\theta}). \tag{2.9}$$

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 $\boldsymbol{S} := \boldsymbol{S}(\alpha^0)$, $\boldsymbol{G} := \boldsymbol{S}^{-1}\boldsymbol{W}$, and $\mathcal{C}$ 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.

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

Assumption ER (Error)**.**

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

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

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

It is important to point out that the least squares objective function (2.8) implicitly uses the reduced form of the dynamic process to generate an internal instrument for the autoregressive parameter. To see this, notice that $\boldsymbol{S}^{-1}(\alpha) = \boldsymbol{I}_T + \alpha \boldsymbol{G}(\alpha)$.[8] Substituting this into the reduced form then yields

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

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

Assumption CS (Consistency)**.**

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

(ii) $\min_{\boldsymbol{\delta} \in \mathbb{R}^{K+1} : ||\boldsymbol{\delta}||_2 = 1} \sum_{r=R+R^0+1}^{T} \mu_r \left(\frac{1}{nT} (\boldsymbol{\delta} \cdot \tilde{\boldsymbol{Z}})^\top (\boldsymbol{\delta} \cdot \tilde{\boldsymbol{Z}}) \right) \geq b > 0$, 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 can intuitively be understood by realising that it implies $\inf_{\tilde{\boldsymbol{\Lambda}} \in \mathbb{R}^{TK \times R^0}, \boldsymbol{F} \in \mathbb{R}^{T \times R}} \mu_{K+1}(\tilde{\boldsymbol{\mathcal{Z}}}^\top (\boldsymbol{M}_{\boldsymbol{F}} \otimes \boldsymbol{M}_{\tilde{\boldsymbol{\Lambda}}}) \tilde{\boldsymbol{\mathcal{Z}}})$ is bounded away from zero. This, therefore, asserts that the data matrix $\tilde{\boldsymbol{\mathcal{Z}}}^\top \tilde{\boldsymbol{\mathcal{Z}}}$ retains a sufficient level of variation, after having been projected orthogonal to arbitrary $R \times T$ factors and $R^0 \times TK$ loadings.

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

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

Proposition 1 demonstrates that as $T/n \to 0$ the estimator is consistent. Moreover, where T is fixed, it is essentially $\sqrt{n}$ consistent. This result is obtained under quite general dependence in the error, and as long as the number of factors used in estimation is no less than the true number. Notice also that no assumptions

[8] See Lemma D.2(i) in Appendix D.

have been made regarding the factors and the loadings other than bounded fourth moments; for instance, these may be strong, weak, or non-existent.[9]

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

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

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

$$\frac{||\boldsymbol{\varepsilon}||_2}{\sqrt{nT}} = \mathcal{O}_p\left(\frac{1}{\sqrt{\min\{n,T\}}}\right).^{10} \tag{3.1}$$

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 ER 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,

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

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

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

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

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

using $\frac{1}{\sqrt{\text{rank}(\boldsymbol{A})}}||\boldsymbol{A}||_F \leq ||\boldsymbol{A}||_2$. Therefore, even in this simple case, $\frac{||\boldsymbol{\varepsilon}||_2}{\sqrt{nT}}$ cannot be $o_p(1)$ 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 (2.8) 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 $\boldsymbol{F}^0\tilde{\boldsymbol{\Lambda}}^{0^\top}\tilde{\boldsymbol{\Lambda}}^0\boldsymbol{F}^{0^\top}/nT$. Key to this approach is demonstrating that the perturbation is asymptotically small, which in this case follows from Proposition 1, whereby $|\theta^0_\kappa - \hat{\theta}_\kappa|$ is small, and from assuming that the 'perturbation' stemming from the error term, $\frac{||\tilde{\boldsymbol{\varepsilon}}||_2}{\sqrt{nT}}$, 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 the $\frac{||\tilde{\boldsymbol{\varepsilon}}||_2}{\sqrt{nT}}$ 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 $\frac{||\boldsymbol{\varepsilon}||_2}{\sqrt{nT}}$.[11]

Assumption AE (Asymptotic Expansion)**.**

(i) $R = R^0$.

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

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

In the absence of dynamics, Moon and Weidner (2015) show that, under certain conditions, the asymptotic distribution of the LS-IFE estimator is unaffected by overstatement of the number of factors. Though it might also be expected that a similar result could be obtained in the present case, doing so is beyond the scope of this paper and the asymptotic distribution is derived under the assumption

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

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 6.2. 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, ER and AE, if* $||\boldsymbol{\theta} - \boldsymbol{\theta}^0||_2 = o_p(1)$*, then, as* $T^2/n \to 0$*,*

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

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

$$\begin{aligned}
D_{\kappa\kappa'} &:= \frac{1}{nT}\mathrm{tr}(\tilde{\boldsymbol{Z}}_\kappa \boldsymbol{M}_{\boldsymbol{F}^0} \tilde{\boldsymbol{Z}}_{\kappa'}^\top \boldsymbol{M}_{\tilde{\boldsymbol{\Lambda}}^0}), \qquad (4.1)\\
c_\kappa &:= \frac{1}{\sqrt{nT}}\mathrm{tr}(\tilde{\boldsymbol{Z}}_\kappa \boldsymbol{M}_{\boldsymbol{F}^0} \tilde{\boldsymbol{\varepsilon}}^\top \boldsymbol{M}_{\tilde{\boldsymbol{\Lambda}}^0}),\\
b_1^{(1)} &:= \frac{1}{\sqrt{nT}}\mathrm{tr}\left(\boldsymbol{M}_{\boldsymbol{F}^0} \boldsymbol{G} \boldsymbol{M}_{\boldsymbol{F}^0} \tilde{\boldsymbol{\varepsilon}}^\top \boldsymbol{M}_{\tilde{\boldsymbol{\Lambda}}^0} \tilde{\boldsymbol{\varepsilon}}\right),\\
b_\kappa^{(2)} &:= -\frac{1}{\sqrt{nT}}\mathrm{tr}\left(\boldsymbol{M}_{\boldsymbol{F}^0} \tilde{\boldsymbol{\varepsilon}}^\top \boldsymbol{M}_{\tilde{\boldsymbol{\Lambda}}^0} \tilde{\boldsymbol{Z}}_\kappa \boldsymbol{F}^0 (\boldsymbol{F}^{0\top}\boldsymbol{F}^0)^{-1} (\tilde{\boldsymbol{\Lambda}}^{0\top} \tilde{\boldsymbol{\Lambda}}^0)^{-1} \tilde{\boldsymbol{\Lambda}}^{0\top} \tilde{\boldsymbol{\varepsilon}}\right),\\
b_\kappa^{(3)} &:= -\frac{1}{\sqrt{nT}}\mathrm{tr}\left(\boldsymbol{M}_{\boldsymbol{F}^0} \tilde{\boldsymbol{Z}}_\kappa^\top \boldsymbol{M}_{\tilde{\boldsymbol{\Lambda}}^0} \tilde{\boldsymbol{\varepsilon}} \boldsymbol{F}^0 (\boldsymbol{F}^{0\top}\boldsymbol{F}^0)^{-1} (\tilde{\boldsymbol{\Lambda}}^{0\top} \tilde{\boldsymbol{\Lambda}}^0)^{-1} \tilde{\boldsymbol{\Lambda}}^{0\top} \tilde{\boldsymbol{\varepsilon}}\right),\\
b_\kappa^{(4)} &:= -\frac{1}{\sqrt{nT}}\mathrm{tr}\left(\boldsymbol{M}_{\boldsymbol{F}^0} \tilde{\boldsymbol{\varepsilon}}^\top \boldsymbol{M}_{\tilde{\boldsymbol{\Lambda}}^0} \tilde{\boldsymbol{\varepsilon}} \boldsymbol{F}^0 (\boldsymbol{F}^{0\top}\boldsymbol{F}^0)^{-1} (\tilde{\boldsymbol{\Lambda}}^{0\top} \tilde{\boldsymbol{\Lambda}}^0)^{-1} \tilde{\boldsymbol{\Lambda}}^{0\top} \tilde{\boldsymbol{Z}}_\kappa\right),
\end{aligned}$$

and $b_\kappa^{(1)} := 0$ *for* $\kappa = 2, \dots K+1$*. Moreover,* $\boldsymbol{r}(\boldsymbol{\theta})$ *is* $o_p\left(\frac{(1+\sqrt{nT}||\boldsymbol{\theta}^0 - \boldsymbol{\theta}||_2)^2}{nT}\right)$.

As will be seen shortly, the term $\boldsymbol{c}$ plays a central role in determining the asymptotic distribution of the estimator. Term $\boldsymbol{b}^{(1)}$ arises due to the presence of a lagged outcome. When using the LF-IFE estimator to estimate the original model, an equivalent term arises and is the source of a bias, as described in Moon and Weidner (2017). Terms $\boldsymbol{b}^{(2)}, \boldsymbol{b}^{(3)}$ and $\boldsymbol{b}^{(4)}$ appear due to cross-sectional and serial dependence in the error term, and, again, have corresponding terms described in both Bai (2009) and Moon and Weidner (2015, 2017) which give rise to additional asymptotic biases. Under Assumptions MD, ER and AE, it can be established that $\boldsymbol{b}^{(1)}, \boldsymbol{b}^{(2)}, \boldsymbol{b}^{(3)}$ and $\boldsymbol{b}^{(4)}$ are $\mathcal{O}_p(T^{1/5}/\sqrt{n})$ which suggests that the estimator will be asymptotically unbiased where $T^3/n \to 0$.[12] This is of course trivially satisfied where T is fixed. In order to establish that this is indeed the case, the following

[12] The origin of the requirement $T^3/n \to 0$ is discussed in Appendix B.

central limit theorem is assumed to hold.

Assumption CLT (Central Limit Theorem)**.**

$$\frac{1}{\sqrt{nT}}\tilde{\boldsymbol{\mathcal{Z}}}^{\top}(\boldsymbol{M}_{\boldsymbol{F}^0}\otimes\boldsymbol{M}_{\tilde{\boldsymbol{\Lambda}}^0})\text{vec}(\tilde{\boldsymbol{\varepsilon}})\xrightarrow{d}\mathcal{N}(\boldsymbol{0},\boldsymbol{\Omega}),$$

with $0 < c_1 \leq \mu_{K+1}(\boldsymbol{\Omega}) \leq \mu_1(\boldsymbol{\Omega}) \leq c_2 < \infty$.

Assumption CLT is a high level assumption and can easily be established under primitive conditions on the covariates, the factors, the loadings and the errors.[13] With the addition of this assumption the main result in this paper is obtained.

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

$$\sqrt{nT}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^0)\xrightarrow{d}\mathcal{N}(\boldsymbol{0},\boldsymbol{D}^{-1}\boldsymbol{\Omega}\boldsymbol{D}^{-1}). \tag{4.2}$$

Theorem 1 demonstrates that the LS-IFE estimator applied to the transformed model is indeed asymptotically unbiased with T fixed, in spite of the generous dependence in the error term permitted under Assumption ER, and notwithstanding with the possible inclusion of a lagged outcome.

5 Understanding the Main Result

In order to understand why the estimator is asymptotically unbiased with T fixed it is necessary to study the asymptotic properties of the estimator when $\frac{T}{n} \to c \geq 0$. This section provides an informal discussion of these properties with formal results given in Appendix B.[14] To simplify exposition, a separable structure is adopted for the variance-covariance matrix of the error term, such that $\boldsymbol{\varepsilon} = \boldsymbol{\Sigma}_n^{\frac{1}{2}}\boldsymbol{U}\boldsymbol{\Sigma}_T^{\frac{1}{2}}$, where the elements of $\boldsymbol{U}$ are independent of the exogenous covariates, the factors and the loadings, identically and independently distributed across i and t, with $\mathbb{E}[u_{it}] = 0$, $\mathbb{E}[u_{it}^2] = 1$ and $\mathbb{E}[u_{it}^4] \leq c$.

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

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

[13] Lemma C.7 in Appendix B provides such a result.

[14] See also previous versions of this paper.

where,

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

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

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

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

and $\boldsymbol{\Delta}$ and $\tilde{\boldsymbol{\Omega}}$ are defined in Appendix B. The most significant change when comparing (4.2) to (5.1) are the four bias terms $\boldsymbol{\psi}^{(0)}, \boldsymbol{\psi}^{(1)}, \boldsymbol{\psi}^{(2)}$ and $\boldsymbol{\psi}^{(3)}$. The first two of these, $\boldsymbol{\psi}^{(0)}$ and $\boldsymbol{\psi}^{(1)}$, arise due to the presence of the dynamic regressor, and indeed $\boldsymbol{\psi}^{(1)}$ is the exact 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 the lower triangular elements of $\boldsymbol{\Sigma}_T$ are non-zero. 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 $\tilde{\boldsymbol{\Sigma}}_n$ has rank TK rather than n. As a consequence, the order of these bias terms are:[15]

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

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-

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

section disappear entirely, it instead shifts them into the time dimension, where they may 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, in contrast to the original model where the analogues of $\boldsymbol{\psi}^{(0)}$, $\boldsymbol{\psi}^{(1)}$, and $\boldsymbol{\psi}^{(3)}$ would be explosive in probability.

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,

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

where $\boldsymbol{\iota}_T$ is a $T \times 1$ vector of ones. In this case $\boldsymbol{\psi}^{(2)} = \boldsymbol{\psi}^{(3)} = \mathbf{0}_{(K+1)\times 1}$ because $\boldsymbol{\Sigma}_T \propto \boldsymbol{I}_T$ and $\boldsymbol{\Sigma}_n \propto \boldsymbol{I}_n$. Moreover, because $\boldsymbol{\Sigma}_T$ is diagonal, $\boldsymbol{\psi}^{(0)} = \mathbf{0}_{(K+1)\times 1}$ 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

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

A bit of algebra reveals that

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

Now, since the trace of a projector is equal to its rank, $\operatorname{tr}(\boldsymbol{P}_{\mathfrak{X}}) = TK$, 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 $\boldsymbol{Q}_{\mathfrak{X}}$ in reducing the rank of the cross-sectional covariance matrix to TK. Without the transformation $\operatorname{tr}(\tilde{\boldsymbol{\Sigma}}_n) = \operatorname{tr}(\boldsymbol{\Sigma}_n) = \operatorname{tr}(\boldsymbol{I}_n) = n$, and so

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

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 $\boldsymbol{Q}_{\mathfrak{X}}$ 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 Low Rank Covariates

Low rank covariates often appear in applied work, with obvious examples being those that are either time or cross-sectionally invariant. In models with interactive effects, identifying the coefficients associated with these covariates can be challenging since they present another low rank structure in the model, in addition to the factor term. Mirroring the result obtained in Moon and Weidner (2017), where such covariates are present it is, however, still possible to obtain consistent estimates under appropriate conditions. Let $\boldsymbol{\vartheta}$ denote a reordering of the parameter vector $\boldsymbol{\theta}$ such that the first K_{L} coefficients, indexed $l = 1, \ldots, K_{\mathrm{L}}$, 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 $\boldsymbol{\mathcal{V}} := (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{K_{\mathrm{L}}})$ and $\boldsymbol{\mathcal{W}} := (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_{K_{\mathrm{L}}})$. 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 $\tilde{\boldsymbol{\mathcal{V}}} := \boldsymbol{Q}_{\boldsymbol{\mathcal{X}}}^\top \boldsymbol{\mathcal{V}}$ and $\boldsymbol{\delta}_{\mathrm{H}} \cdot \tilde{\boldsymbol{Z}}_{\mathrm{H}} := \sum_{\kappa=1}^{K_{\mathrm{H}}} \delta_\kappa \tilde{\boldsymbol{Z}}_\kappa$.

Assumption LR (Low Rank)**.**

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

(ii) There exists $c > 0$ 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

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

Since transforming the model by $\boldsymbol{Q}_{\mathcal{X}}^{\top}$ 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 3 (Consistency – Low Rank). *Under Assumptions MD, AE, ER, and LR, as* $T/n \to 0$

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

6.2 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 with T fixed.

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

$$\mu_r^* \coloneqq \mu_r\left(\frac{1}{nT}\left(\tilde{\boldsymbol{Y}}\boldsymbol{S}(\hat{\alpha})-\sum_{\kappa=1}^{K}\hat{\beta}_\kappa\tilde{\boldsymbol{X}}_\kappa\right)^\top\left(\tilde{\boldsymbol{Y}}\boldsymbol{S}(\hat{\alpha})-\sum_{\kappa=1}^{K}\hat{\beta}_\kappa\tilde{\boldsymbol{X}}_\kappa\right)+\frac{T}{n}\boldsymbol{I}_T\right), \quad (6.1)$$

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

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

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

Proposition 4. *Under Assumptions MD, CS and ER, as $T/n \to 0$,*

$$\Pr\left(\max_{1\le r\le T}\mu_r^* = R^0\right) \to 1. \quad (6.2)$$

6.3 Balestra and Nerlove's Approach

The estimator $\hat{\boldsymbol{\theta}}$ studied thus far in this paper treats the initial condition $\boldsymbol{y}_0\boldsymbol{s}^\top(\alpha)$ as an additional parameter and, as a consequence, results in another factor appearing in the error term. An alternative approach which does not generate this additional factor is to follow Balestra and Nerlove (1966) and include the projection of the lagged outcome onto the column space of the exogenous variables as an additional explanatory variable on the right hand-side of the outcome equation. This approach is naturally wedded to this paper's, since projecting lagged outcomes embeds them in the TK-dimensional space spanned by the columns of $\boldsymbol{\mathfrak{X}}$. Consider the following outcome equation:

$$\boldsymbol{Y}^c = \alpha\boldsymbol{Y}_L^c + \sum_{k=1}^{K}\beta_k\boldsymbol{X}_k^c + \boldsymbol{\Lambda}^*\boldsymbol{F}^{*c\top} + \boldsymbol{\varepsilon}^c, \quad (6.3)$$

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

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

where $\tilde{\boldsymbol{Z}}^c_1 := \boldsymbol{Q}^\top_{\mathfrak{X}^c}\boldsymbol{Y}^c_L$ and $\tilde{\boldsymbol{Z}}^c_\kappa := \tilde{\boldsymbol{X}}^c_\kappa$ for $\kappa = 2, \ldots, K+1$. An alternative estimator $\hat{\boldsymbol{\theta}}_{BN}$ may then be defined as

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

This estimator retains all of the essential properties of $\hat{\boldsymbol{\theta}}$, including fixed T consistency and an analogous asymptotic distribution. Moreover, this approach is especially appealing since it involves simply transforming the data by $\boldsymbol{Q}_{\mathfrak{X}}$ and then applying the usual LF-IFE estimator with no other modifications.[16]

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 $\boldsymbol{\Sigma}_n^{\frac{1}{2}}\boldsymbol{U}\boldsymbol{\Sigma}_T^{\frac{1}{2}}$, where the elements of $\boldsymbol{U}$ are independently drawn from a standard normal distribution, and $\boldsymbol{\Sigma}_n$ and $\boldsymbol{\Sigma}_T$ are diagonal matrices with elements drawn uniformly between 0.5 and 2.5. The number of Monte Carlo replications is 10000. Tables 1a – 1c display the bias and the standard error of the naive LS estimator (Naive), the LS-IFE estimator applied to the original model (IFE), the approach described in Section 6.3 (BN-IFE) and the LS-IFE estimator applied to the transformed model (Q-IFE).

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

Table 1a: Bias (SE) α

	Naive			IFE			BN-IFE			Q-IFE		
$n \setminus T$	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			BN-IFE			Q-IFE		
$n \setminus T$	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			BN-IFE			Q-IFE		
$n \setminus T$	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 BN-IFE and the Q-IFE 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 BN-IFE and Q-IFE 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 – 2b below present coverage probabilities of the estimators based on the asymptotic variance-covariance matrix, and with a nominal value of 95%.

Table 2a: Coverage α %

	Naive			IFE			BN-IFE			Q-IFE		
$n \setminus T$	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			BN-IFE			Q-IFE		
$n \setminus T$	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 2b: Coverage β_2 %

	Naive			IFE			BN-IFE			QP-IFE		
$n \setminus T$	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 BN-IFE and Q-IFE 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 BN-IFE and Q-IFE estimates present poor coverage with both n and T small, yet these rapidly improve as n increases. When comparing the performance of BN-IFE and Q-IFE, it is, in general, the case that BN-IFE outperforms Q-IFE. This is a consequence of the fact that, while omitting a time period, BN-IFE uses only 2 factors in estimation, whereas Q-IFE uses 3, with the extra factor being present to control for a possibly endogenous initial condition. Clearly, including an additional factor in estimation has a noticeable impact on the efficiency of the estimator in finite samples, therefore it is useful to apply the eigenvalue ratio test described in Section 6.2 to uncover the appropriate number of factors to use in estimation.

Table 3: Number of Factors Chosen Q-IFE %

	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 6.2 to the Q-IFE 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 negligible as either dimension of the panel grows. In light of this, Tables 4a and 4b below present bias and coverages for Q-IFE with $R = 2$. Comparing these results to those presented previously, these estimates are generally better than both BN-IFE and Q-IFE with an additional factor. However, BN still outperforms Q-IFE when it comes to the autoregressive parameter α.

Table 4a: Bias (SE), Q-IFE 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 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 introduces a new method to estimate linear panel data models with interactive fixed effects designed for situations where T is fixed. 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. This contrasts sharply with the usual case where the LS-IFE estimator generally delivers biased and inconsistent estimates with T fixed. Several other consequences of this approach are also discussed, particularly the ability to apply other inferential procedures designed for the large n, large T setting to the transformed model. This is illustrated by modifying the eigenvalue ratio test of Ahn and Horenstein (2013) to render it applicable in the present setting.

There are two natural extensions to this paper, both of which are currently in progress. The first is to notice that the estimator proposed in this paper can be obtained as a marginal likelihood associated with a maximal invariant statistic under the group of transformations (2.3). 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. 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., 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, oCLC: ocn243960539.

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

Hayakawa, K., 2012. GMM estimation of short dynamic panel data models with interactive fixed effects. Journal of the Japan Statistical Society 42 (2), 109 – 123.

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

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

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.