Panel Data Models with Interactive Fixed Effects and Relatively Small T

Ayden Higgins[†]

University of Exeter

September 19, 2025

Abstract

This paper studies the estimation of a linear panel data model with interactive fixed effects. A transformation is introduced which, after having been applied, renders the least squares (LS) estimator of Bai (2009) consistent and asymptotically unbiased when n is large and T is fixed. This is termed the *transformed* least squares (TLS) estimator. Going further, these properties are shown to also carry over to the large n, large T setting, provided $T/n \to 0$. This contrasts sharply with the usual case, where the LS estimator is, in general, inconsistent when n is large and T is fixed, and is asymptotically biased when both n and T are large.

Keywords: interactive fixed effects, panel data, factor models.

JEL classification: C13, C33, C38.

[*]This work was supported by the European Research Council through the grants ERC-2016-STG-715787-MiMo and ERC-2018-CoG-819086-PANEDA. For the purpose of open access, the author has applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising from this submission.

[†]Address: University of Exeter Business School, Rennes Drive, Exeter, Devon, EX4 4PU, United Kingdom. Email: a.higgins@exeter.ac.uk.

1 Introduction

1.1 Model

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 $i = 1, \ldots, n$ and $t = 1, \ldots, T$, outcomes are generated according to

$$\boldsymbol{y}_t = \boldsymbol{X}_t \boldsymbol{\beta} + \boldsymbol{\Lambda} \boldsymbol{f}_t + \boldsymbol{\varepsilon}_t, \tag{1.1}$$

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 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 $K \times 1$ 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 $\boldsymbol{\beta}$ 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 be inconsistent. 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. 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 $\boldsymbol{\beta}$. 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 adversely impacts the estimation of common parameters in a model. In long panels this problem can, to some extent, be overcome, and, in particular, it has been shown that the least squares interactive fixed effects (LS) estimator that treats both the factors and the loadings as fixed effects is consistent when both n and T are large, though it typically suffers from asymptotic bias (Bai, 2009; Moon and Weidner, 2017). It is, however, in short panels that the incidental parameter problem is felt most acutely, and when n is large and T fixed the LS estimator is inconsistent, in general.

This paper proposes a simple remedy to this problem, by introducing a transformation of the model, which, after having been applied, renders the LS estimator consistent when n is large and T is fixed. This is termed the *transformed* least squares interactive fixed effects (TLS) estimator. 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 TLS estimator retains many of the most attractive features of the LS estimator, including certain robustness properties and, crucially, the ability to profile out the factors and the loadings from the objective function, and to reduce estimation to a univariate optimisation. And yet, unlike the LS estimator, the TLS estimator is shown to not only be consistent, but also asymptotically unbiased when n is large and T is fixed.

A deeper understanding of the relationship between the LS and TLS estimators comes from also studying the TLS estimator when both n and T are large. This analysis reveals that the TLS estimator remains consistent and, indeed, remains asymptotically unbiased, provided $T/n \to 0$. However, when $T/n \to \gamma \in (0, \infty)$ it exhibits asymptotic bias analogous both in its origin and its functional form to that of the LS estimator. Inspection of the resultant expressions establishes that the bias of the TLS estimator is of a lower order than that of the LS estimator. And yet, at the same time, comparison of the asymptotic variance of the two estimators indicates that this is achieved at the expense of efficiency, with the former being less efficient than the latter, at least in some circumstances. Overall the results of this paper point to a bias-variance trade-off between the two estimators, which is most apparent, and most important, when n is large and T is fixed, and more generally when T is small relative to n.

1.2 Related Literature

In light of the fixed T-inconsistency of the LS estimator, alternative estimation approaches have been considered which are applicable when n is large and T is fixed. Yet unlike in 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 fixed T approaches rely on the possibility of correlation existing between these, and indeed lean into this as a means to derive alternative estimators. 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 (CCE) approaches, which originate from the seminal work of Pesaran (2006). At the core of this approach is an assumption that at least some model covariates also admit a factor decomposition, such that the factors can be estimated by taking cross-sectional averages of these covariates. This ultimately gives rise to estimators that are often consistent when n is large and T is fixed, as well as when both n and T are large. The properties of this approach have been extensively studied, e.g., in a likelihood setting (Bai and Li, 2014), with dynamic regressors (Everaert and Groote, 2016), with an unknown number of factors (Westerlund and Urbain, 2015). Other contributions in this line of research include Westerlund (2020), De 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 a 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).[1] Though in some sense less restrictive, these estimators are often much more difficult to implement than are CCE-type estimators, with it being necessary to directly estimate multiple nuisance parameters alongside the parameter of interest, and in some cases, to do so from a set of highly non-linear moment conditions. As a consequence, these estimators appear less frequently in applied work.[2]

The TLS estimator follows in this second line of research, in the sense that it does not impose a particular functional form for the relationship between the factors, the loadings, and the covariates. It does not, however, share the complexity of those approaches, as it consists of a univariate optimisation which depends only on the parameter of interest, and not (directly) on nuisance parameters arising through modelling the factor structure in the error. Nonetheless, subsequent sections show that the TLS estimator is closely related to two of these estimators in particular: the quasi-difference estimator of Ahn et al. (2013) (ALS), and the FIVU estimator of Robertson and Sarafidis (2015) (RS). A detailed comparison of these approaches reveals both similarities and subtle differences between the LS and TLS estimators, and comparable one-step ALS and RS estimators. This is an interesting finding, and goes some way to bridging the gap between the least squares and method of moments-based approaches to panel models with interactive effects.[3]

1.3 Outline

Section 2 sets out the estimation approach, introducing the transformation and providing some intuition behind the key differences that lie between the LS and TLS estimators. Section 3 establishes the asymptotic properties of the TLS estimator, including consistency and asymptotic normality, and draws comparisons with the corresponding results for the LS estimator. Section 4 examines the relationship between the TLS estimator and some alternative estimators, under a large n, fixed T asymptotic. Section 5 collects additional considerations, including inferential procedures, a method to detect the correct number of factors, and an extension to dynamic models. Section 6 contains simulations, and Section 7 concludes. Proofs of all results are to be found in the

[1] Freyberger (2018) also falls under this heading.

[2] Hsiao et al. (2022) describe an alternative approach also applicable to short panels.

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

Supplementary Material.

1.4 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, $\|\boldsymbol{a}\|_2 := \sqrt{\sum_{i=1}^{n} a_i^2}$. 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 spectral norm and Frobenius norm of $\boldsymbol{A}$ are denoted $\|\boldsymbol{A}\|_2 := \sqrt{\mu_{\max}(\boldsymbol{A}^\top \boldsymbol{A})}$ and $\|\boldsymbol{A}\|_F := \sqrt{\operatorname{tr}(\boldsymbol{A}^\top \boldsymbol{A})}$, respectively. The notation $\|\boldsymbol{A}\|_{\max}$ is used to denote $\max_{1 \leq i \leq n} \max_{1 \leq j \leq m} |A_{ij}|$. 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 $\boldsymbol{\iota}_n$. 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 $\operatorname{vec}(\cdot)$ applied to an $n \times m$ matrix $\boldsymbol{A}$ creates an $nm \times 1$ vector $\operatorname{vec}(\boldsymbol{A})$ by stacking the columns of $\boldsymbol{A}$.

2 TLS Estimator

Treating both the factors and the loadings as additional (nuisance) parameters, the LS estimator of (1.1) is obtained as the values $(\boldsymbol{\beta}, \boldsymbol{\Lambda}, \boldsymbol{F})$ 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 estimator of $\boldsymbol{\beta}$ is consistent when the number of factors is known and both n and T are large. 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 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 estimator to be inconsistent when T is fixed. Yet, as is shown subsequently, by first transforming the model, the LS estimator can be rendered consistent and asymptotically unbiased when n is large and T is fixed.

2.1 Transformation

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

$$\boldsymbol{Y} = \boldsymbol{X} \cdot \boldsymbol{\beta} + \boldsymbol{\Lambda} \boldsymbol{F}^\top + \boldsymbol{\varepsilon}. \tag{2.1}$$

Define the $n \times TK$ matrix $\boldsymbol{\mathfrak{X}} \coloneqq (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_K)$. It is assumed that $\boldsymbol{\mathfrak{X}}$ has full rank, i.e., $\operatorname{rank}(\boldsymbol{\mathfrak{X}}) = \min\{n, TK\}$.[4] Take a singular value decomposition $\boldsymbol{\mathfrak{X}} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^\top$, where the matrices of singular vectors $\boldsymbol{U}$ and $\boldsymbol{V}$ are $n \times \operatorname{rank}(\boldsymbol{\mathfrak{X}})$ and $TK \times \operatorname{rank}(\boldsymbol{\mathfrak{X}})$, respectively, and the diagonal matrix of singular values $\boldsymbol{S}$ is $\operatorname{rank}(\boldsymbol{\mathfrak{X}}) \times \operatorname{rank}(\boldsymbol{\mathfrak{X}})$. Define the transformation matrix $\boldsymbol{Q}_{\mathfrak{X}} \coloneqq \boldsymbol{U}\boldsymbol{V}^\top$, with which the following transformed variables can be defined:

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

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

$$\tilde{\boldsymbol{Y}} = \tilde{\boldsymbol{X}} \cdot \boldsymbol{\beta} + \tilde{\boldsymbol{\Lambda}} \boldsymbol{F}^\top + \tilde{\boldsymbol{\varepsilon}}. \tag{2.2}$$

The key properties of $\boldsymbol{Q}_{\mathfrak{X}}$ are presented in Appendix A.1. Intuitively, the action of $\boldsymbol{Q}_{\mathfrak{X}}$ is most important when $TK < n$, in which case $\boldsymbol{Q}_{\mathfrak{X}}$ acts to reduce the model into the TK-dimensional subspace spanned by the columns of $\boldsymbol{\mathfrak{X}}$. In particular, the dimension of the transformed factor term $\tilde{\boldsymbol{\Lambda}} \boldsymbol{F}^\top$ will no longer depend on n, whereupon the model is relieved of incidental parameters in the cross-section.[5] A second important feature of $\boldsymbol{Q}_{\mathfrak{X}}$ is that transforming the model leads to no loss of information in the covariates, which is manifest in the property $\boldsymbol{Q}_{\mathfrak{X}} \boldsymbol{Q}_{\mathfrak{X}}^\top \boldsymbol{\mathfrak{X}} = \boldsymbol{\mathfrak{X}}$.[6] Finally, if the covariates

[4] If $\boldsymbol{\mathfrak{X}}$ is not full rank, or indeed, if one chooses to specify $\boldsymbol{\mathfrak{X}}$ using some, but not all of the columns of the covariates, subject to the satisfaction of the requisite conditions, the TLS estimator will retain its essential properties.

[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 3.4.

[6] See Appendix A.1.

used in the construction of $\boldsymbol{Q}_{\mathfrak{X}}$ are strictly exogenous, under quite general conditions, the transformation serves to reduce the order of the error term which, ultimately, is key to estimating (2.2) when n is large and T is fixed.[7]

2.2 Principal Components

The underlying mechanics of the LS 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 $\boldsymbol{\beta}$, estimating the factors and loadings is a standard principal component problem. Where n is large and T is fixed 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 the true $\boldsymbol{\beta}$ is observed, and that both $\boldsymbol{\Lambda}$ and $\boldsymbol{F}$ have full column rank. Then $\boldsymbol{Y} - \boldsymbol{X}\cdot\boldsymbol{\beta} = \boldsymbol{\Lambda}\boldsymbol{F}^\top + \boldsymbol{\varepsilon}$ has a pure factor structure. Decompose the factor term as $\boldsymbol{\Lambda}\boldsymbol{F}^\top = \dot{\boldsymbol{\Lambda}}\dot{\boldsymbol{F}}^\top$, where $\dot{\boldsymbol{\Lambda}}$ and $\dot{\boldsymbol{F}}$ are $n \times R$ and $T \times R$ matrices, respectively, which satisfy $n^{-1}\dot{\boldsymbol{\Lambda}}^\top\dot{\boldsymbol{\Lambda}} = \boldsymbol{I}_R$ and $\dot{\boldsymbol{F}}^\top\dot{\boldsymbol{F}}$ is diagonal.[8] Consider the following:

$$
\begin{aligned}
&\frac{1}{nT}(\boldsymbol{Y} - \boldsymbol{X}\cdot\boldsymbol{\beta})(\boldsymbol{Y} - \boldsymbol{X}\cdot\boldsymbol{\beta})^\top\frac{\dot{\boldsymbol{\Lambda}}}{\sqrt{n}} \\
&\qquad\qquad = \frac{1}{nT}\boldsymbol{\Lambda}\boldsymbol{F}^\top\boldsymbol{F}\boldsymbol{\Lambda}^\top\frac{\dot{\boldsymbol{\Lambda}}}{\sqrt{n}} + \frac{1}{nT}\left(\boldsymbol{\varepsilon}\boldsymbol{F}\boldsymbol{\Lambda}^\top + \boldsymbol{\Lambda}\boldsymbol{F}^\top\boldsymbol{\varepsilon}^\top + \boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top\right)\frac{\dot{\boldsymbol{\Lambda}}}{\sqrt{n}} \\
&\qquad\qquad =: \frac{\dot{\boldsymbol{\Lambda}}}{\sqrt{n}}\left(\frac{1}{T}\dot{\boldsymbol{F}}^\top\dot{\boldsymbol{F}}\right) + \boldsymbol{e}(\dot{\boldsymbol{\Lambda}}, \boldsymbol{\Lambda}, \boldsymbol{F}, \boldsymbol{\varepsilon}).
\end{aligned}
$$

Given that $n^{-1}\dot{\boldsymbol{\Lambda}}^\top\dot{\boldsymbol{\Lambda}} = \boldsymbol{I}_R$ and $\dot{\boldsymbol{F}}^\top\dot{\boldsymbol{F}}$ is diagonal, then, absent of the second term on the right, the columns of $\dot{\boldsymbol{\Lambda}}$ would be eigenvectors of $(nT)^{-1}(\boldsymbol{Y} - \boldsymbol{X}\cdot\boldsymbol{\beta})(\boldsymbol{Y} - \boldsymbol{X}\cdot\boldsymbol{\beta})^\top$. Where both n and T are large, several authors have shown that, in spite of this second term, estimating $\dot{\boldsymbol{\Lambda}}$ in this manner may still be possible. For example, under the condition $\|\boldsymbol{\varepsilon}\|_2 = \mathcal{O}_p(\sqrt{\max\{n,T\}})$ employed in Moon and Weidner (2015), dependence in the

[7]This paper focuses on the case where the regressors are strictly exogenous, as in Bai (2009). Lagged outcomes can also be accommodated as is discussed in Section 5.3. If a covariate $\boldsymbol{X}_k$ is endogenous but valid instruments are available, then those instruments can, in principle, substitute for $\boldsymbol{X}_k$ in the construction of $\mathfrak{X}$. A more complete treatment of the IV setting will be the focus of future work.

[8]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 $\dot{\boldsymbol{\Lambda}}$ be the R columns of $\sqrt{n}\boldsymbol{U}$ associated with the nonzero singular values, and $\dot{\boldsymbol{F}}$ be the corresponding R columns of $n^{-\frac{1}{2}}\boldsymbol{V}\boldsymbol{S}^\top$. As the columns of $\boldsymbol{U}$ and $\boldsymbol{V}$ are orthonormal, it follows that $n^{-1}\dot{\boldsymbol{\Lambda}}^\top\dot{\boldsymbol{\Lambda}} = \boldsymbol{I}_R$, $\dot{\boldsymbol{F}}^\top\dot{\boldsymbol{F}}$ is diagonal, and $\dot{\boldsymbol{\Lambda}}\dot{\boldsymbol{F}}^\top = \boldsymbol{\Lambda}\boldsymbol{F}^\top$.

error term is sufficiently limited that, with suitable conditions on the factors and the loadings, the second term is (uniformly) $o_p(1)$ when both $n, T \to \infty$. However, such arguments typically fail when T is fixed.

Consider instead the transformed model. Let $\ddot{\boldsymbol{\Lambda}}$ and $\ddot{\boldsymbol{F}}$ be $TK \times R$ and $T \times R$ matrices, respectively, which satisfy $\ddot{\boldsymbol{\Lambda}}\ddot{\boldsymbol{F}}^\top = \tilde{\boldsymbol{\Lambda}}\boldsymbol{F}^\top$, $n^{-1}\ddot{\boldsymbol{\Lambda}}^\top\ddot{\boldsymbol{\Lambda}} = \boldsymbol{I}_R$, and $\ddot{\boldsymbol{F}}^\top\ddot{\boldsymbol{F}}$ is diagonal.[9] Note that $\ddot{\boldsymbol{F}}$ typically differs from $\dot{\boldsymbol{F}}$. One arrives at a similar expression to before,

$$\begin{aligned}
&\frac{1}{nT}(\tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{X}} \cdot \boldsymbol{\beta})(\tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{X}} \cdot \boldsymbol{\beta})^\top \frac{\ddot{\boldsymbol{\Lambda}}}{\sqrt{n}} \\
&\qquad = \frac{1}{nT}\tilde{\boldsymbol{\Lambda}}\boldsymbol{F}^\top\boldsymbol{F}\tilde{\boldsymbol{\Lambda}}^\top\frac{\ddot{\boldsymbol{\Lambda}}}{\sqrt{n}} + \frac{1}{nT}\left(\tilde{\boldsymbol{\varepsilon}}\boldsymbol{F}\tilde{\boldsymbol{\Lambda}}^\top + \tilde{\boldsymbol{\Lambda}}\boldsymbol{F}^\top\tilde{\boldsymbol{\varepsilon}}^\top + \tilde{\boldsymbol{\varepsilon}}\tilde{\boldsymbol{\varepsilon}}^\top\right)\frac{\ddot{\boldsymbol{\Lambda}}}{\sqrt{n}} \\
&\qquad =: \frac{\ddot{\boldsymbol{\Lambda}}}{\sqrt{n}}\left(\frac{1}{T}\ddot{\boldsymbol{F}}^\top\ddot{\boldsymbol{F}}\right) + \boldsymbol{e}(\ddot{\boldsymbol{\Lambda}}, \tilde{\boldsymbol{\Lambda}}, \boldsymbol{F}, \tilde{\boldsymbol{\varepsilon}}).
\end{aligned}$$

Yet now, if the covariates used to construct $\boldsymbol{Q}_{\mathcal{X}}$ are strictly exogenous, under quite general conditions, $\|\tilde{\boldsymbol{\varepsilon}}\|_2 = \mathcal{O}_p(n^{\frac{1}{4}}\sqrt{T})$ from which it follows that, again, with suitable conditions on the factors and the loadings, the second term in the above is (uniformly) $o_p(1)$ as $n \to \infty$ with T fixed (or indeed $T \to \infty$). As a consequence, it is possible to estimate the columns of $\ddot{\boldsymbol{\Lambda}}$ as eigenvectors of $(nT)^{-1}(\tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{X}} \cdot \boldsymbol{\beta})(\tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{X}} \cdot \boldsymbol{\beta})^\top$. Although in general $\ddot{\boldsymbol{\Lambda}}$ will not equal $\tilde{\boldsymbol{\Lambda}}$, these two matrices will share the same column space which suffices to control for the term in estimation.

2.3 Objective Function

The transformed model (2.2) can be estimated by minimising the following least squares objective function:

$$\mathcal{Q}(\boldsymbol{\beta}, \tilde{\boldsymbol{\Lambda}}, \boldsymbol{F}) \coloneqq \frac{1}{nT}\operatorname{tr}\left(\left(\tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{X}} \cdot \boldsymbol{\beta} - \tilde{\boldsymbol{\Lambda}}\boldsymbol{F}^\top\right)^\top \left(\tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{X}} \cdot \boldsymbol{\beta} - \tilde{\boldsymbol{\Lambda}}\boldsymbol{F}^\top\right)\right). \tag{2.3}$$

Both the factors and the transformed loadings can be profiled out of (2.3), in which case one arrives at an objective function involving $\boldsymbol{\beta}$ alone

$$\mathcal{Q}(\boldsymbol{\beta}) \coloneqq \frac{1}{nT}\sum_{t=R+1}^{T}\mu_t\left(\left(\tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{X}} \cdot \boldsymbol{\beta}\right)^\top\left(\tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{X}} \cdot \boldsymbol{\beta}\right)\right), \tag{2.4}$$

[9] It is tacitly assumed that $\operatorname{rank}(\boldsymbol{\Lambda}\boldsymbol{F}^\top) = \operatorname{rank}(\tilde{\boldsymbol{\Lambda}}\boldsymbol{F}^\top)$.

that is, the profile objective function involves the sum of the $(T-R)$ smallest eigenvalues of the right-hand side matrix.[10] Using this, the TLS estimator $\hat{\boldsymbol{\beta}}$ can then be defined as

$$\hat{\boldsymbol{\beta}} \coloneqq \underset{\boldsymbol{\beta}\in\Theta_\beta}{\arg\min}\, \mathcal{Q}(\boldsymbol{\beta}), \tag{2.5}$$

where Θ_β denotes a suitable parameter space for $\boldsymbol{\beta}$.[11,12,13]

3 Asymptotic Properties

3.1 Consistency

Throughout the following both $\boldsymbol{\Lambda}$ and $\boldsymbol{F}$ are treated as fixed parameters in estimation and the subscript 0 is now introduced to distinguish true parameter values. The following assumptions are made.

Assumption MD.

(i) The parameter vector $\boldsymbol{\beta}_0$ lies in the interior of Θ_β, where Θ_β is a compact subset of $\mathbb{R}^K$.

(ii) The elements of $\boldsymbol{X}_k$, $\boldsymbol{\Lambda}_0$, and $\boldsymbol{F}_0$ have uniformly bounded fourth moments.

Assumption MD(ii) imposes standard conditions on the moments of the covariates, the factors, and the loadings. Let $\mathcal{C}_{nT}$ denote $\sigma(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_K)$, that is, the sigma-algebra generated by the covariates, and define $\boldsymbol{\Sigma}_{\mathcal{C}} \coloneqq \mathbb{E}[\text{vec}(\boldsymbol{\varepsilon})\text{vec}(\boldsymbol{\varepsilon})^\top|\mathcal{C}_{nT}]$.

Assumption EC. Conditional on $\mathcal{C}_{nT}$, ε_{it} are independent over i, with $\mathbb{E}[\varepsilon_{it}|\mathcal{C}_{nT}] = 0$, and $\mathbb{E}[\varepsilon_{it}^4|\mathcal{C}_{nT}]$ uniformly bounded. In addition, the eigenvalues of $\boldsymbol{\Sigma}_{\mathcal{C}}$ are uniformly bounded away from zero and from above by a constant.

Assumption EC allows for heteroskedasticity (conditional and unconditional) in both dimensions of the panel, as well as serial dependence. The conditions imposed are weaker

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

[11] Note that the objective function $\mathcal{Q}(\boldsymbol{\beta})$ need not be convex in $\boldsymbol{\beta}$, and therefore there may exist local minima. A practical consequence of this is that any optimisation routine should be initialised from multiple starting values.

[12] Notice that when $TK \geq n$, $\boldsymbol{Q}_{\mathbb{X}}\boldsymbol{Q}_{\mathbb{X}}^\top = \boldsymbol{I}_n$, and therefore the TLS estimator is equal to the LS estimator: see Appendix A.1.

[13] The author is grateful to a referee for pointing out that, with strictly exogenous regressors, the TLS estimator coincides with an iterative estimator described in Breitung and Hansen (2021), which the authors refer to as ALS*. This is based on an iterative procedure detailed in the appendix of Ahn et al. (2013) as a means to compute their estimator. Neither Breitung and Hansen (2021) nor Ahn et al. (2013) establish any asymptotic results for this estimator.

than the assumption that the errors (and typically also the covariates and the loadings) are identically distributed over i, as is frequently assumed in the literature on short panels with interactive effects (see, e.g., Ahn et al. (2013) and Robertson and Sarafidis (2015)). Moreover, it is not required that the error be independent of covariates, the factors, and the loadings as in, for instance, Bai (2009). In aid of the following let $\tilde{\boldsymbol{X}} \cdot \boldsymbol{\delta} \coloneqq \sum_{k=1}^{K} \delta_k \tilde{\boldsymbol{X}}_k$. Moreover, let $T_{\min} \coloneqq R_e + R_0 + 1$.

Assumption CS.

(i) $R_e \geq R_0 \coloneqq \operatorname{rank}(\boldsymbol{\Lambda}_0 \boldsymbol{F}_0^\top)$, where R_e denotes the number of factors used in estimation, and R_e and R_0 are constants that do not depend on sample size.

(ii) $\min_{\boldsymbol{\delta} \in \mathbb{R}^K : \|\boldsymbol{\delta}\|_2 = 1} \sum_{t=T_{\min}}^{T} \mu_t((nT)^{-1}(\tilde{\boldsymbol{X}} \cdot \boldsymbol{\delta})^\top (\tilde{\boldsymbol{X}} \cdot \boldsymbol{\delta})) \geq b > 0$ w.p.a.1 as $n \to \infty$, with $T \geq T_{\min}$ fixed or $T \to \infty$.

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_e, is no less than R_0. This assumption also formalises the core model assumption that the factor term $\boldsymbol{\Lambda}_0 \boldsymbol{F}_0^\top$ has a low (relative to sample size), fixed rank. Assumption CS(ii) is a multicollinearity condition. Notice that

$$
\begin{aligned}
&\min_{\boldsymbol{\delta} \in \mathbb{R}^K : \|\boldsymbol{\delta}\|_2 = 1} \sum_{t=T_{\min}}^{T} \mu_t \left(\frac{1}{nT} (\tilde{\boldsymbol{X}} \cdot \boldsymbol{\delta})^\top (\tilde{\boldsymbol{X}} \cdot \boldsymbol{\delta}) \right) \\
&= \min_{\tilde{\boldsymbol{\Lambda}} \in \mathbb{R}^{TK \times R_e},\ \boldsymbol{F} \in \mathbb{R}^{T \times R_0}} \mu_{\min} \left(\frac{1}{nT} \tilde{\boldsymbol{\mathcal{X}}}^\top (\boldsymbol{M}_{\boldsymbol{F}} \otimes \boldsymbol{M}_{\tilde{\boldsymbol{\Lambda}}}) \tilde{\boldsymbol{\mathcal{X}}} \right),
\end{aligned}
\tag{3.1}
$$

where $\tilde{\boldsymbol{\mathcal{X}}} \coloneqq (\operatorname{vec}(\tilde{\boldsymbol{X}}_1), \ldots, \operatorname{vec}(\tilde{\boldsymbol{X}}_K))$. Going further, it can be established that

$$
\begin{aligned}
&\min_{\boldsymbol{\delta} \in \mathbb{R}^K : \|\boldsymbol{\delta}\|_2 = 1} \sum_{t=T_{\min}}^{T} \mu_t \left(\frac{1}{nT} (\tilde{\boldsymbol{X}} \cdot \boldsymbol{\delta})^\top (\tilde{\boldsymbol{X}} \cdot \boldsymbol{\delta}) \right) \\
&\geq \min_{\boldsymbol{\Lambda} \in \mathbb{R}^{n \times R_e},\ \boldsymbol{F} \in \mathbb{R}^{T \times R_0}} \mu_{\min} \left(\frac{1}{nT} \boldsymbol{\mathcal{X}}^\top (\boldsymbol{M}_{\boldsymbol{F}} \otimes \boldsymbol{M}_{\boldsymbol{\Lambda}}) \boldsymbol{\mathcal{X}} \right),
\end{aligned}
\tag{3.2}
$$

where $\boldsymbol{\mathcal{X}}$ is defined analogously to $\tilde{\boldsymbol{\mathcal{X}}}$.[14] Hence, Assumption CS(ii) is satisfied as long as there remains a sufficient amount of variation in the regressors after having been projected orthogonal to arbitrary $T \times R_0$ factors and $n \times R_e$ loadings. This is analogous

[14] See Appendix A.2.

to Assumption A in Bai (2009) and Assumption NC in Moon and Weidner (2015).[15]

Proposition 1 (Consistency). *Under Assumptions MD, EC, and CS, $\hat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}_0$ as $n \to \infty$, with $T \geq T_{\min}$ fixed or $T \to \infty$.*

Proposition 1 demonstrates the TLS estimator is consistent as $n \to \infty$, regardless of whether T is fixed or $T \to \infty$. This result is obtained allowing for heteroskedasticity and serial 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.[16]

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 estimator. Their result establishes that

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2 = \mathcal{O}_p\left(\frac{1}{\sqrt{\min\{n, T\}}}\right),$$

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

$$\frac{\|\boldsymbol{\varepsilon}\|_2}{\sqrt{nT}} = \mathcal{O}_p\left(\frac{1}{\sqrt{\min\{n, T\}}}\right).^{17} \tag{3.3}$$

In similar fashion, the rate obtained in Proposition 1 can be attributed to the quantity $\|\tilde{\boldsymbol{\varepsilon}}\|_2$ which plays an analogous role in this paper. Under Assumption EC this can be shown to satisfy

$$\frac{\|\tilde{\boldsymbol{\varepsilon}}\|_2}{\sqrt{nT}} = \mathcal{O}_p\left(n^{-\frac{1}{4}}\right).$$

Recalling the discussion in Section 2.2, it is worth stressing again the importance of

[15] See Appendix C in Higgins and Martellosio (2023) for further discussion on the relation between these conditions.

[16] Using $\|\tilde{\boldsymbol{\varepsilon}}\|_2 = \mathcal{O}_p(T^{\frac{3}{4}})$, which is established in the proof of Lemma B.2(i), inspection of the proof of Proposition 1 reveals that as $n \to \infty$ with $T \geq T_{\min}$ fixed, the TLS is $\sqrt{n}$-consistent irrespective of the strength of the factors.

[17] Moreover, (3.3) also proves to be important for the asymptotic expansion of the objective function; see Section 3.2.

the difference between $\boldsymbol{\varepsilon}$ and $\tilde{\boldsymbol{\varepsilon}}$. To highlight this, consider the rudimentary example of identically and independently distributed errors, i.e., $\mathbb{E}[\varepsilon_{i_1t_1}\varepsilon_{i_2t_2}] = \sigma^2$ for $i_1 = i_2$ and $t_1 = t_2$, and is zero otherwise. In such a case,

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

as $n \to \infty$ with T fixed using $\operatorname{rank}(\boldsymbol{A})^{-\frac{1}{2}}\|\boldsymbol{A}\|_F \leq \|\boldsymbol{A}\|_2$. Therefore $(nT)^{-\frac{1}{2}}\|\boldsymbol{\varepsilon}\|_2$ cannot be $\mathcal{O}_p(1)$ with T fixed, provided σ is bounded from below by a constant.

3.2 Asymptotic Expansion

Typically the asymptotic distribution of an extremum estimator is obtained by expanding the objective function locally around the true parameter value. It is, however, difficult to obtain an expansion of the objective function (2.4) since this involves a summation over a certain number of eigenvalues of a matrix. Following Bai (2009), one 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. Key to this approach is demonstrating that the perturbation is asymptotically small, which in this case follows from Proposition 1, whereby $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2$ is small, and from $(nT)^{-\frac{1}{2}}\|\tilde{\boldsymbol{\varepsilon}}\|_2$ diminishing asymptotically. In light of the discussion in the previous section, the significance of transforming the errors is again highlighted as the expansion of the objective function remains valid only so long as $(nT)^{-\frac{1}{2}}\|\tilde{\boldsymbol{\varepsilon}}\|_2$ is asymptotically small. Since $\|\tilde{\boldsymbol{\varepsilon}}\|_2 \leq \|\boldsymbol{\varepsilon}\|_2$, $(nT)^{-\frac{1}{2}}\|\tilde{\boldsymbol{\varepsilon}}\|_2$ will be asymptotically small in situations where this will not be true of $(nT)^{-\frac{1}{2}}\|\boldsymbol{\varepsilon}\|_2$.[18] Let $\mathcal{D}_{nT} := \mathcal{C}_{nT} \vee \sigma(\tilde{\boldsymbol{\Lambda}}_0, \boldsymbol{F}_0)$.[19] The following assumption is made.

Assumption ED. Conditional on $\mathcal{D}_{nT}$, ε_{it} are independent over i, with $\mathbb{E}[\varepsilon_{it}|\mathcal{D}_{nT}] = 0$, $\mathbb{E}[\varepsilon_{it}^2|\mathcal{D}_{nT}] > 0$, and $\sup_{\|\boldsymbol{v}\|_2=1}\mathbb{E}[(\boldsymbol{v}^\top\boldsymbol{\varepsilon}_i)^4|\mathcal{D}_{nT}]$ uniformly bounded for $\mathcal{D}_{nT}$-measurable vectors $\boldsymbol{v}$.

Assumption ED strengthens EC to restrict dependence between the error and the factor term, and imposes more stringent conditions on serial dependence in the error.

[18] The inequality $\|\tilde{\boldsymbol{\varepsilon}}\|_2 \leq \|\boldsymbol{\varepsilon}\|_2$ is obtained by the submultiplicativity of the spectral norm and noting that $\|\boldsymbol{Q}_{\mathfrak{X}}\|_2 = 1$; see Appendix A.1.

[19] For two sigma-algebras $\mathcal{A}$ and $\mathcal{B}$, $\mathcal{A} \vee \mathcal{B}$ denotes the sigma-algebra generated by the union of both $\mathcal{A}$ and $\mathcal{B}$.

The last part of this assumption can be understood as a generalised bound on the fourth moment of the error, and is closely related to Assumption C(iv) in Bai (2009) which can be seen by noticing that

$$\frac{1}{T^2}\sum_{t_1=1}^{T}\sum_{t_2=1}^{T}\sum_{t_3=1}^{T}\sum_{t_4=1}^{T}\operatorname{cov}(\varepsilon_{it_1}\varepsilon_{it_2},\varepsilon_{it_3}\varepsilon_{it_4}) = \mathbb{E}[(\boldsymbol{v}^\top\boldsymbol{\varepsilon}_i)^4] - \mathbb{E}[(\boldsymbol{v}^\top\boldsymbol{\varepsilon}_i)^2]^2,$$

with $\boldsymbol{v} = \boldsymbol{\iota}_T/\sqrt{T}$. The more general condition arises since, unlike in Bai (2009), the errors are not assumed to be independent of the factors, the loadings, and the covariates.

Assumption AE.

(i) $R_e = R_0 = \operatorname{rank}(\tilde{\boldsymbol{\Lambda}}_0\boldsymbol{F}_0^\top)$.

(ii) $n^{-1}\tilde{\boldsymbol{\Lambda}}_0^\top\tilde{\boldsymbol{\Lambda}}_0 \xrightarrow{p} \boldsymbol{\Sigma}_{\tilde{\boldsymbol{\Lambda}}_0}$ as $n\to\infty$, with $T \geq R_0+1$ fixed or $T\to\infty$, where the eigenvalues of $\boldsymbol{\Sigma}_{\tilde{\boldsymbol{\Lambda}}_0}$ are bounded away from zero and from above by a constant.

(iii) For $T \geq R_0+1$ fixed, the eigenvalues of $\boldsymbol{F}_0^\top\boldsymbol{F}_0$ are bounded away from zero and from above by a constant, otherwise $T^{-1}\boldsymbol{F}_0^\top\boldsymbol{F}_0 \xrightarrow{p} \boldsymbol{\Sigma}_{\boldsymbol{F}_0}$ as $T\to\infty$, where the eigenvalues of $\boldsymbol{\Sigma}_{\boldsymbol{F}_0}$ are bounded away from zero and from above by a constant.

Assumption AE(i) strengthens CS(i) and imposes that the number of factors used in estimation equals to the true rank of the factor term as in Bai (2009) and Moon and Weidner (2017). A method to detect the true number of factors is discussed in Section 5.2. Assumptions AE(ii) and AE(iii) are similar in spirit to the strong factor assumption, Assumption B in Bai (2009) and SF in Moon and Weidner (2015). Assumption AE(ii) is, however, somewhat stronger since it could be that $n^{-1}\boldsymbol{\Lambda}_0^\top\boldsymbol{\Lambda}_0$ converges in probability to a positive definite matrix, while $n^{-1}\tilde{\boldsymbol{\Lambda}}_0^\top\tilde{\boldsymbol{\Lambda}}_0$ converges in probability to a singular matrix. The leading example of this is where some or all of the factor loadings are independent of the covariates. Suppose, for example, that $\boldsymbol{\lambda}_{0,i} \sim \operatorname{iid}(\mathbf{0},\boldsymbol{\Sigma}_\lambda)$, with $\mathbb{E}[\|\boldsymbol{\lambda}_{0,i}\|_2^4]$ uniformly bounded, and are independent of the covariates. Then as $n\to\infty$, with T fixed or $T\to\infty$ and $T/n\to\gamma\in[0,\infty)$,

$$\frac{1}{n}\tilde{\boldsymbol{\Lambda}}_0^\top\tilde{\boldsymbol{\Lambda}}_0 \xrightarrow{p} \min\{1,\gamma K\}\times\boldsymbol{\Sigma}_\lambda.$$

Thus, if $\gamma>0$ and $\boldsymbol{\Sigma}_\lambda \succ 0$, then Assumption AE(ii) would still be satisfied. However, if $\gamma = 0$ then this is no longer the case, and while the TLS estimator would remain consistent (Proposition 1 does not require Assumption AE(ii)), it would become more challenging to establish its asymptotic distribution. In this event, one may instead pursue a random effects approach; see, e.g., Section 5 in Hsiao (2018).

Proposition 2. *Assume $\boldsymbol{\beta} \xrightarrow{p} \boldsymbol{\beta}_0$ as $n \to \infty$, with $T \geq T_{\min}$ fixed or $T \to \infty$. Under Assumptions MD, ED, and AE, as $n \to \infty$, with $T \geq T_{\min}$ fixed or $T \to \infty$,*

$$\mathcal{Q}(\boldsymbol{\beta}) = \mathcal{Q}(\boldsymbol{\beta}_0) - \frac{2}{\sqrt{nT}}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)^\top \boldsymbol{d} + (\boldsymbol{\beta} - \boldsymbol{\beta}_0)^\top \boldsymbol{D}(\boldsymbol{\beta} - \boldsymbol{\beta}_0) + r(\boldsymbol{\beta}),$$

where $\boldsymbol{d} := \boldsymbol{c} + \boldsymbol{b}^{(1)} + \boldsymbol{b}^{(2)} + \boldsymbol{b}^{(3)}$ with

$$\begin{aligned}
D_{k_1k_2} &:= \frac{1}{nT}\mathrm{tr}(\tilde{\boldsymbol{X}}_{k_1}\boldsymbol{M}_{\boldsymbol{F}_0}\tilde{\boldsymbol{X}}_{k_2}^\top \boldsymbol{M}_{\tilde{\boldsymbol{\Lambda}}_0}) \\
c_k &:= \frac{1}{\sqrt{nT}}\mathrm{tr}(\tilde{\boldsymbol{X}}_k\boldsymbol{M}_{\boldsymbol{F}_0}\tilde{\boldsymbol{\varepsilon}}^\top \boldsymbol{M}_{\tilde{\boldsymbol{\Lambda}}_0}) \\
b_k^{(1)} &:= -\frac{1}{\sqrt{nT}}\mathrm{tr}\left(\boldsymbol{M}_{\boldsymbol{F}_0}\tilde{\boldsymbol{\varepsilon}}^\top \boldsymbol{M}_{\tilde{\boldsymbol{\Lambda}}_0}\tilde{\boldsymbol{X}}_k\boldsymbol{F}_0(\boldsymbol{F}_0^\top\boldsymbol{F}_0)^{-1}(\tilde{\boldsymbol{\Lambda}}_0^\top\tilde{\boldsymbol{\Lambda}}_0)^{-1}\tilde{\boldsymbol{\Lambda}}_0^\top\tilde{\boldsymbol{\varepsilon}}\right) \\
b_k^{(2)} &:= -\frac{1}{\sqrt{nT}}\mathrm{tr}\left(\boldsymbol{M}_{\boldsymbol{F}_0}\tilde{\boldsymbol{X}}_k^\top \boldsymbol{M}_{\tilde{\boldsymbol{\Lambda}}_0}\tilde{\boldsymbol{\varepsilon}}\boldsymbol{F}_0(\boldsymbol{F}_0^\top\boldsymbol{F}_0)^{-1}(\tilde{\boldsymbol{\Lambda}}_0^\top\tilde{\boldsymbol{\Lambda}}_0)^{-1}\tilde{\boldsymbol{\Lambda}}_0^\top\tilde{\boldsymbol{\varepsilon}}\right) \\
b_k^{(3)} &:= -\frac{1}{\sqrt{nT}}\mathrm{tr}\left(\boldsymbol{M}_{\boldsymbol{F}_0}\tilde{\boldsymbol{\varepsilon}}^\top \boldsymbol{M}_{\tilde{\boldsymbol{\Lambda}}_0}\tilde{\boldsymbol{\varepsilon}}\boldsymbol{F}_0(\boldsymbol{F}_0^\top\boldsymbol{F}_0)^{-1}(\tilde{\boldsymbol{\Lambda}}_0^\top\tilde{\boldsymbol{\Lambda}}_0)^{-1}\tilde{\boldsymbol{\Lambda}}_0^\top\tilde{\boldsymbol{X}}_k\right).
\end{aligned}$$

Moreover, $r(\boldsymbol{\beta})$ is $o_p((nT)^{-1}(1 + \sqrt{nT}\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_2)^2)$.

Proposition 2 establishes an expansion of the objective function around the true parameter value $\boldsymbol{\beta}_0$, from which the asymptotic distribution of the estimator can be obtained.

3.3 Asymptotic Distribution

In order to describe the asymptotic distribution of the estimator some additional notation is introduced. Let

$$\boldsymbol{V} := \frac{1}{nT}\tilde{\boldsymbol{\mathcal{X}}}^\top(\boldsymbol{M}_{\boldsymbol{F}_0}\otimes\boldsymbol{M}_{\tilde{\boldsymbol{\Lambda}}_0})\tilde{\boldsymbol{\Sigma}}_{\mathcal{D}}(\boldsymbol{M}_{\boldsymbol{F}_0}\otimes\boldsymbol{M}_{\tilde{\boldsymbol{\Lambda}}_0})\tilde{\boldsymbol{\mathcal{X}}},$$

$\boldsymbol{\Sigma}_{\mathcal{D}} := \mathbb{E}[\mathrm{vec}(\boldsymbol{\varepsilon})\mathrm{vec}(\boldsymbol{\varepsilon})^\top|\mathcal{D}_{nT}]$, and $\tilde{\boldsymbol{\Sigma}}_{\mathcal{D}} := (\boldsymbol{I}_T\otimes\boldsymbol{Q}_{\mathcal{X}}^\top)\boldsymbol{\Sigma}_{\mathcal{D}}(\boldsymbol{I}_T\otimes\boldsymbol{Q}_{\mathcal{X}})$.

Assumption AD.

(i) The elements of $\boldsymbol{M}_{\boldsymbol{P}_{\mathcal{X}}\boldsymbol{\Lambda}_0}\boldsymbol{X}_k$, $\boldsymbol{\Lambda}_0$, and $\boldsymbol{F}_0$ have uniformly bounded eighth moments.

(ii) There exist nonstochastic matrices $\mathbb{D}$ and $\mathbb{V}$ such that $\boldsymbol{D} \xrightarrow{p} \mathbb{D}$ and $\boldsymbol{V} \xrightarrow{p} \mathbb{V}$ as $n \to \infty$, with $T \geq T_{\min}$ fixed or $T \to \infty$, and the eigenvalues of $\mathbb{D}$ and $\mathbb{V}$ are bounded away from zero and from above by a constant.

The first part of Assumption AD(i) requires that after having been transformed by $\boldsymbol{M}_{\boldsymbol{P}_{\mathfrak{X}}\boldsymbol{\Lambda}_0}$, the covariates have finite eighth moments. Intuitively this can be thought of as applying a weighted demeaning to the data. If, for example, $\|\boldsymbol{P}_{\mathfrak{X}}\boldsymbol{\Lambda}_0\|_{\max}$, $\|\boldsymbol{\Lambda}_0\|_{\max}$, and $\|\boldsymbol{X}_k\|_{\max}$ are uniformly bounded, and

$$\mu_{\min}\left(\frac{1}{n}\tilde{\boldsymbol{\Lambda}}_0^\top \tilde{\boldsymbol{\Lambda}}_0\right) \geq c > 0,$$

then this can be shown to hold.

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

(i) *with* $T \geq T_{\min}$ *fixed or* $T \to \infty$ *and* $T/n \to 0$*,*

$$\sqrt{nT}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbb{D}^{-1}\mathbb{V}\mathbb{D}^{-1}),$$

(ii) *with* $T \to \infty$ *and* $T/n \to \gamma \in (0, \infty)$*,*

$$\sqrt{nT}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) + \boldsymbol{D}^{-1}(\boldsymbol{\psi}^{(1)} + \boldsymbol{\psi}^{(2)}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbb{D}^{-1}\mathbb{V}\mathbb{D}^{-1}),$$

where

$$\begin{aligned}
\psi_k^{(1)} &:= \frac{1}{\sqrt{nT}}\mathrm{tr}(\tilde{\boldsymbol{\Sigma}}_{\mathcal{D}}(\boldsymbol{I}_T \otimes \boldsymbol{M}_{\tilde{\boldsymbol{\Lambda}}_0}\tilde{\boldsymbol{X}}_k\boldsymbol{F}_0(\boldsymbol{F}_0^\top \boldsymbol{F}_0)^{-1}(\tilde{\boldsymbol{\Lambda}}_0^\top \tilde{\boldsymbol{\Lambda}}_0)^{-1}\tilde{\boldsymbol{\Lambda}}_0^\top)) \\
\psi_k^{(2)} &:= \frac{1}{\sqrt{nT}}\mathrm{tr}(\tilde{\boldsymbol{\Sigma}}_{\mathcal{D}}(\boldsymbol{F}_0(\boldsymbol{F}_0^\top \boldsymbol{F}_0)^{-1}(\tilde{\boldsymbol{\Lambda}}_0^\top \tilde{\boldsymbol{\Lambda}}_0)^{-1}\tilde{\boldsymbol{\Lambda}}_0^\top \tilde{\boldsymbol{X}}_k\boldsymbol{M}_{\boldsymbol{F}_0} \otimes \boldsymbol{I}_{TK})).
\end{aligned}$$

Theorem 1 establishes that the TLS estimator is asymptotically normally distributed. When $n \to \infty$ and T is fixed it is asymptotically unbiased. When $n, T \to \infty$ and $T/n \to 0$ it can be established that $\boldsymbol{\psi}^{(1)}$ and $\boldsymbol{\psi}^{(2)}$ are $\boldsymbol{o}_p(1)$, and therefore the TLS estimator remains asymptotically unbiased. However, when $n, T \to \infty$ and $T/n \to \gamma \in (0, \infty)$ the TLS estimator, like the LS estimator, may exhibit asymptotic bias.

3.4 Comparison with LS Estimator

The terms $\boldsymbol{\psi}^{(1)}$ and $\boldsymbol{\psi}^{(2)}$ that appear in Theorem 1 are near duplicates of the corresponding expressions described in Theorem 3 of Bai (2009), which provides a distributional result for the LS estimator. This result, translated in the present context, reads

$$\sqrt{nT}(\hat{\boldsymbol{\beta}}_{\mathrm{LS}} - \boldsymbol{\beta}_0) + \boldsymbol{D}_{\mathrm{LS}}^{-1}(\boldsymbol{\psi}_{\mathrm{LS}}^{(1)} + \boldsymbol{\psi}_{\mathrm{LS}}^{(2)}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbb{D}_{\mathrm{LS}}^{-1}\mathbb{V}_{\mathrm{LS}}\mathbb{D}_{\mathrm{LS}}^{-1}), \tag{3.4}$$

as $n, T \to \infty$ with $T/n \to \gamma \in (0, \infty)$, and where

$$\begin{aligned}
\psi_{\mathrm{LS},k}^{(1)} &:= \frac{1}{\sqrt{nT}}\mathrm{tr}(\boldsymbol{\Sigma}_{\mathcal{D}^*}(\boldsymbol{I}_T \otimes \boldsymbol{M}_{\boldsymbol{\Lambda}_0}\boldsymbol{X}_k\boldsymbol{F}_0(\boldsymbol{F}_0^\top\boldsymbol{F}_0)^{-1}(\boldsymbol{\Lambda}_0^\top\boldsymbol{\Lambda}_0)^{-1}\boldsymbol{\Lambda}_0^\top)) \\
\psi_{\mathrm{LS},k}^{(2)} &:= \frac{1}{\sqrt{nT}}\mathrm{tr}(\boldsymbol{\Sigma}_{\mathcal{D}^*}(\boldsymbol{F}_0(\boldsymbol{F}_0^\top\boldsymbol{F}_0)^{-1}(\boldsymbol{\Lambda}_0^\top\boldsymbol{\Lambda}_0)^{-1}\boldsymbol{\Lambda}_0^\top\boldsymbol{X}_k\boldsymbol{M}_{\boldsymbol{F}_0} \otimes \boldsymbol{I}_n)) \\
\boldsymbol{D}_{\mathrm{LS}} &:= \frac{1}{nT}\boldsymbol{\mathcal{X}}^\top(\boldsymbol{M}_{\boldsymbol{F}_0} \otimes \boldsymbol{M}_{\boldsymbol{\Lambda}_0})\boldsymbol{\mathcal{X}} \\
\boldsymbol{V}_{\mathrm{LS}} &:= \frac{1}{nT}\boldsymbol{\mathcal{X}}^\top(\boldsymbol{M}_{\boldsymbol{F}_0} \otimes \boldsymbol{M}_{\boldsymbol{\Lambda}_0})\boldsymbol{\Sigma}_{\mathcal{D}^*}(\boldsymbol{M}_{\boldsymbol{F}_0} \otimes \boldsymbol{M}_{\boldsymbol{\Lambda}_0})\boldsymbol{\mathcal{X}},
\end{aligned}$$

with $\mathbb{D}_{\mathrm{LS}}$ and $\mathbb{V}_{\mathrm{LS}}$ being nonstochastic matrices, such that $\boldsymbol{D}_{\mathrm{LS}} \xrightarrow{p} \mathbb{D}_{\mathrm{LS}}$ and $\boldsymbol{V}_{\mathrm{LS}} \xrightarrow{p} \mathbb{V}_{\mathrm{LS}}$ as $n, T \to \infty$ with $T/n \to \gamma \in (0, \infty)$, and $\boldsymbol{\Sigma}_{\mathcal{D}^*} := \mathbb{E}[\mathrm{vec}(\boldsymbol{\varepsilon})\mathrm{vec}(\boldsymbol{\varepsilon})^\top|\mathcal{D}_{nT}^*]$ with $\mathcal{D}_{nT}^* := \mathcal{C}_{nT} \vee \sigma(\boldsymbol{\Lambda}_0, \boldsymbol{F}_0)$.

Comparing (3.4) to Theorem 1, one may observe that when $n, T \to \infty$ and $T/n \to \gamma \in [K^{-1}, \infty)$ the LS estimator and the TLS estimator are asymptotically equivalent. This is so because with $TK \geq n$, $\boldsymbol{Q}_{\mathfrak{X}}\boldsymbol{Q}_{\mathfrak{X}}^\top = \boldsymbol{I}_n$, and therefore the action of the transformation is redundant. However, when $T/n \to \gamma \in [0, K^{-1})$ this is no longer the case. In particular, examining the order of the bias terms:

	$\boldsymbol{\psi}_\bullet^{(1)}$	$\boldsymbol{\psi}_\bullet^{(2)}$
LS Estimator	$\mathcal{O}_p\left(\sqrt{\frac{T}{n}}\right)$	$\mathcal{O}_p\left(\sqrt{\frac{n}{T}}\right)$
TLS Estimator	$\mathcal{O}_p\left(\sqrt{\frac{T}{n}}\right)$	$\mathcal{O}_p\left(\min\left\{\sqrt{\frac{n}{T}}, \sqrt{\frac{T}{n}}\right\}\right)$.

Therefore, while for the LS estimator $\boldsymbol{\psi}_{\mathrm{LS}}^{(2)}$ is explosive as $n, T \to \infty$ and $T/n \to 0$, for the TLS estimator both $\boldsymbol{\psi}^{(1)}$ and $\boldsymbol{\psi}^{(2)}$ are $\mathcal{O}_p(1)$ under the same asymptotic, and thereby the TLS estimator is asymptotically unbiased. Indeed, it is on account of this difference that the TLS estimator is consistent with T fixed, while the LS estimator is not.

A natural question arises as to the relative efficiency of the LS and TLS estimators. There is no clear ordering, in general. Nonetheless, insight can be gained from considering the particular case in which the errors are homoskedastic. In this scenario, the following result is obtained.

Proposition 3. *Assume* $\boldsymbol{\Sigma}_{\mathcal{D}} = \boldsymbol{\Sigma}_{\mathcal{D}^*} = \sigma_0^2 \boldsymbol{I}_{nT}$*, and there exist nonstochastic matrices* $\mathbb{D}$ *and* $\mathbb{D}_{\mathrm{LS}}$ *such that* $\boldsymbol{D} \xrightarrow{p} \mathbb{D}$ *and* $\boldsymbol{D}_{\mathrm{LS}} \xrightarrow{p} \mathbb{D}_{\mathrm{LS}}$ *as* $n, T \to \infty$ *with* $T/n \to \gamma \in (0, \infty)$*, and the eigenvalues of* $\mathbb{D}$ *and* $\mathbb{D}_{\mathrm{LS}}$ *are bounded away from zero and from above by a constant. Moreover, assume*

$$\operatorname{avar}(\sqrt{nT}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)) = \sigma_0^2 \mathbb{D}^{-1}$$
$$\operatorname{avar}(\sqrt{nT}(\hat{\boldsymbol{\beta}}_{\mathrm{LS}} - \boldsymbol{\beta}_0)) = \sigma_0^2 \mathbb{D}_{\mathrm{LS}}^{-1},$$

where $\operatorname{avar}(\cdot)$ *denotes asymptotic variance. Then*

$$\operatorname{avar}(\sqrt{nT}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)) \succeq \operatorname{avar}(\sqrt{nT}(\hat{\boldsymbol{\beta}}_{\mathrm{LS}} - \boldsymbol{\beta}_0)).$$

Proposition 3 establishes that the LS estimator will, in some instances, be more efficient than the TLS estimator. To appreciate the source of the inefficiency, notice that

$$\boldsymbol{D} - \boldsymbol{D}_{\mathrm{LS}} = \frac{1}{nT} \boldsymbol{\mathcal{X}}^\top (\boldsymbol{M}_{\boldsymbol{F}_0} \otimes (\boldsymbol{P}_{\boldsymbol{\Lambda}_0} - \boldsymbol{P}_{\boldsymbol{P}_{\mathfrak{X}} \boldsymbol{\Lambda}_0})) \boldsymbol{\mathcal{X}}. \tag{3.5}$$

The LS estimator implicitly estimates both the factors and the loadings simultaneously. In transforming the model, information about the original factor loadings is lost which, ultimately, may result in a larger variance. This information loss is manifest in (3.5). Of course, as remarked on previously, when $\gamma \in [K^{-1}, \infty)$ the two estimators are asymptotically equivalent, and therefore they achieve the same asymptotic efficiency. This is manifest in (3.5) as $\boldsymbol{P}_{\boldsymbol{P}_{\mathfrak{X}} \boldsymbol{\Lambda}_0} = \boldsymbol{P}_{\boldsymbol{\Lambda}_0}$ with $TK \geq n$.

One may ask whether the model could be transformed in an alternative way in order to minimise any efficiency loss. If, for example, $\boldsymbol{\Lambda}_0$ is a smooth function of $\mathfrak{X}$, one may consider using powers of $\mathfrak{X}$ in the manner of a series approximation to construct an $n \times d$

matrix $\boldsymbol{\mathcal{W}}$, and thereafter construct $\boldsymbol{Q}_{\boldsymbol{\mathcal{W}}}$ and proceed as previously. Alternatively, one may have access to additional external variables which can supplement the covariates in $\boldsymbol{\mathcal{W}}$ to achieve a better approximation of the column space of the factor loadings. Indeed, in the event that $\mathrm{col}((\boldsymbol{\mathcal{X}}, \boldsymbol{\Lambda}_0)) \subseteq \mathrm{col}(\boldsymbol{\mathcal{W}})$ there is no loss of information in transforming model through $\boldsymbol{Q}_{\boldsymbol{\mathcal{W}}}$, and therefore $\mathrm{avar}(\sqrt{nT}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)) = \mathrm{avar}(\sqrt{nT}(\hat{\boldsymbol{\beta}}_{\mathrm{LS}}-\boldsymbol{\beta}_0))$. However, there will typically be a cost to this in terms of asymptotic bias, since the analogue of $\boldsymbol{\psi}^{(2)}$ that appears in Theorem 1 would generally be of order $\min\{n, d\}(nT)^{-\frac{1}{2}}$. Therefore improving on the approximation at the expense of a larger d would typically result in greater bias.

Following on from this discussion, it is natural to compare a generalised least squares interactive fixed effects estimator (GLS) constructed as

$$
\begin{aligned}
\hat{\boldsymbol{\beta}}^*_{\mathrm{GLS}} = &\left(\boldsymbol{\mathcal{X}}^\top \left((\boldsymbol{M}_{\boldsymbol{F}_0} \otimes \boldsymbol{M}_{\boldsymbol{\Lambda}_0})\boldsymbol{\Sigma}_{\mathcal{D}^*}(\boldsymbol{M}_{\boldsymbol{F}_0} \otimes \boldsymbol{M}_{\boldsymbol{\Lambda}_0})\right)^+ \boldsymbol{\mathcal{X}}\right)^{-1} \\
&\times \boldsymbol{\mathcal{X}}^\top \left((\boldsymbol{M}_{\boldsymbol{F}_0} \otimes \boldsymbol{M}_{\boldsymbol{\Lambda}_0})\boldsymbol{\Sigma}_{\mathcal{D}^*}(\boldsymbol{M}_{\boldsymbol{F}_0} \otimes \boldsymbol{M}_{\boldsymbol{\Lambda}_0})\right)^+ \mathrm{vec}(\boldsymbol{Y}),
\end{aligned}
$$

and a corresponding generalised *transformed* least squares interactive fixed effects estimator (GTLS)

$$
\begin{aligned}
\hat{\boldsymbol{\beta}}^*_{\mathrm{GTLS}} = &\left(\tilde{\boldsymbol{\mathcal{X}}}^\top \left((\boldsymbol{M}_{\boldsymbol{F}_0} \otimes \boldsymbol{M}_{\tilde{\boldsymbol{\Lambda}}_0})\tilde{\boldsymbol{\Sigma}}_{\mathcal{D}}(\boldsymbol{M}_{\boldsymbol{F}_0} \otimes \boldsymbol{M}_{\tilde{\boldsymbol{\Lambda}}_0})\right)^+ \tilde{\boldsymbol{\mathcal{X}}}\right)^{-1} \\
&\times \tilde{\boldsymbol{\mathcal{X}}}^\top \left((\boldsymbol{M}_{\boldsymbol{F}_0} \otimes \boldsymbol{M}_{\tilde{\boldsymbol{\Lambda}}_0})\tilde{\boldsymbol{\Sigma}}_{\mathcal{D}}(\boldsymbol{M}_{\boldsymbol{F}_0} \otimes \boldsymbol{M}_{\tilde{\boldsymbol{\Lambda}}_0})\right)^+ \mathrm{vec}(\tilde{\boldsymbol{Y}}).
\end{aligned}
$$

The relative efficiency of these estimators is compared in the following result. Let

$$
\begin{aligned}
\boldsymbol{D}^* &:= \frac{1}{nT}\tilde{\boldsymbol{\mathcal{X}}}^\top \left((\boldsymbol{M}_{\boldsymbol{F}_0} \otimes \boldsymbol{M}_{\tilde{\boldsymbol{\Lambda}}_0})\tilde{\boldsymbol{\Sigma}}_{\mathcal{D}}(\boldsymbol{M}_{\boldsymbol{F}_0} \otimes \boldsymbol{M}_{\tilde{\boldsymbol{\Lambda}}_0})\right)^+ \tilde{\boldsymbol{\mathcal{X}}} \\
\boldsymbol{D}^*_{\mathrm{LS}} &:= \frac{1}{nT}\boldsymbol{\mathcal{X}}^\top \left((\boldsymbol{M}_{\boldsymbol{F}_0} \otimes \boldsymbol{M}_{\boldsymbol{\Lambda}_0})\boldsymbol{\Sigma}_{\mathcal{D}^*}(\boldsymbol{M}_{\boldsymbol{F}_0} \otimes \boldsymbol{M}_{\boldsymbol{\Lambda}_0})\right)^+ \boldsymbol{\mathcal{X}}.
\end{aligned}
$$

Proposition 4. *Assume $\boldsymbol{\Sigma}_{\mathcal{D}} = \boldsymbol{\Sigma}_{\mathcal{D}^*} = \boldsymbol{\Sigma}$, where $\boldsymbol{\Sigma}$ is nonstochastic, and there exist nonstochastic matrices $\mathbb{D}^*$ and $\mathbb{D}^*_{\mathrm{LS}}$, such that $\boldsymbol{D}^* \xrightarrow{p} \mathbb{D}^*$ and $\boldsymbol{D}^*_{\mathrm{LS}} \xrightarrow{p} \mathbb{D}^*_{\mathrm{LS}}$ as $n, T \to \infty$ with $T/n \to \gamma \in (0, \infty)$, and the eigenvalues of $\mathbb{D}^*$ and $\mathbb{D}^*_{\mathrm{LS}}$ are bounded away from zero* *and from above by a constant. Moreover, assume*

$$\operatorname{avar}(\sqrt{nT}(\hat{\boldsymbol{\beta}}^*_{\text{GTLS}} - \boldsymbol{\beta}_0)) = \mathbb{D}^{*-1}$$
$$\operatorname{avar}(\sqrt{nT}(\hat{\boldsymbol{\beta}}^*_{\text{GLS}} - \boldsymbol{\beta}_0)) = \mathbb{D}^{*-1}_{\text{LS}}.$$

Then

$$\operatorname{avar}(\sqrt{nT}(\hat{\boldsymbol{\beta}}^*_{\text{GTLS}} - \boldsymbol{\beta}_0)) \succeq \operatorname{avar}(\sqrt{nT}(\hat{\boldsymbol{\beta}}^*_{\text{GLS}} - \boldsymbol{\beta}_0)).$$

Intuitively, the source of the relative inefficiency of the GTLS estimator remains the loss of information through transforming the factor loadings.

4 Alternative Estimators

According to Theorem 1, when n is large and T is fixed the TLS estimator is consistent and asymptotically unbiased. This stands in contrast to the LS estimator which is generally inconsistent with T fixed. There are, however, alternative estimators that can be applied to estimate $\boldsymbol{\beta}_0$ when T is fixed. Of particular interest here are the FIVU estimator of Robertson and Sarafidis (2015) (RS), and the quasi-difference estimator of Ahn et al. (2013) (ALS). This section places the RS, ALS, and TLS estimators in a common framework and establishes connections between them. Since the focus of this section is on the large n, fixed T setting, it is assumed throughout the following that $\boldsymbol{\mathfrak{X}}$ has full column rank.

4.1 RS Estimator

Under strict exogeneity

$$\mathbb{E}\left[(\boldsymbol{I}_T \otimes \boldsymbol{\mathfrak{X}})^\top \operatorname{vec}(\boldsymbol{\varepsilon})\right] = \mathbb{E}\left[(\boldsymbol{I}_T \otimes \boldsymbol{\mathfrak{X}})^\top \operatorname{vec}(\boldsymbol{Y} - \boldsymbol{X} \cdot \boldsymbol{\beta}_0 - \boldsymbol{\Lambda}_0 \boldsymbol{F}_0^\top)\right] = \mathbf{0}.$$

If, in addition, one assumes the data $\{\boldsymbol{X}_i, \boldsymbol{\lambda}_{0,i}, \boldsymbol{\varepsilon}_i\}$ are identically distributed over i, and that the factors are fixed, then

$$\begin{aligned}\mathbb{E}\left[\mathfrak{X}^\top \boldsymbol{\Lambda}_0 \boldsymbol{F}_0^\top\right] &= \sum_{i=1}^{n} \mathbb{E}\left[\mathrm{vec}(\boldsymbol{X}_i)\boldsymbol{\lambda}_{0,i}^\top\right] \boldsymbol{F}_0^\top \\ &=: n\boldsymbol{\Psi}_0 \boldsymbol{F}_0^\top.\end{aligned}$$ [20]

This leads to the FIVU estimator of Robertson and Sarafidis (2015), which is based on the moment condition

$$\mathbb{E}\left[(\boldsymbol{I}_T \otimes \mathfrak{X})^\top \mathrm{vec}(\boldsymbol{Y} - \boldsymbol{X} \cdot \boldsymbol{\beta}_0) - n\mathrm{vec}(\boldsymbol{\Psi}_0 \boldsymbol{F}_0^\top)\right] = \boldsymbol{0}. \tag{4.1}$$

Though their approach is predicated on the factor loadings being random, one may instead adopt a fixed effect perspective treating these as additional parameters. In doing so one may reframe

$$\sum_{i=1}^{n} \mathrm{vec}(\boldsymbol{X}_i)\boldsymbol{\lambda}_{0,i}^\top = \mathfrak{X}^\top \boldsymbol{\Lambda}_0 = (\mathfrak{X}^\top \mathfrak{X})^{\frac{1}{2}} \tilde{\boldsymbol{\Lambda}}_0, \tag{4.2}$$

and so consider the alternate moment condition

$$\mathbb{E}\left[(\boldsymbol{I}_T \otimes \mathfrak{X})^\top \mathrm{vec}(\boldsymbol{Y} - \boldsymbol{X} \cdot \boldsymbol{\beta}_0) - \mathrm{vec}((\mathfrak{X}^\top \mathfrak{X})^{\frac{1}{2}} \tilde{\boldsymbol{\Lambda}}_0 \boldsymbol{F}_0^\top)\right] = \boldsymbol{0}, \tag{M-RS}$$

which, notice, does not rely on the factor loadings and the covariates being identically distributed over the cross-section. This latter moment condition is referred to as M-RS since, despite differing from (4.1), both conditions share a common conception.

M-RS does not, however, uniquely identify $\tilde{\boldsymbol{\Lambda}}_0$ nor $\boldsymbol{F}_0$ since the product $\tilde{\boldsymbol{\Lambda}}_0 \boldsymbol{F}_0^\top = \tilde{\boldsymbol{\Lambda}}_0 \boldsymbol{H}\boldsymbol{H}^{-1} \boldsymbol{F}_0^\top = \tilde{\boldsymbol{\Lambda}}_* \boldsymbol{F}_*^\top$ for any $R_0 \times R_0$ invertible matrix $\boldsymbol{H}$. This is a consequence of a fundamental indeterminacy inherent to the factor structure, and is typically dealt with by focusing on a particular factorisation of $\tilde{\boldsymbol{\Lambda}}_0 \boldsymbol{F}_0^\top$. There are multiple factorisations which may be plausibly imposed, but for concreteness the following scheme is adopted:

$$\boldsymbol{F}_{\mathrm{RS}} = \begin{pmatrix} \boldsymbol{I}_{R_0} \\ \boldsymbol{\Phi}_{\mathrm{RS}} \end{pmatrix}, \qquad \tilde{\boldsymbol{\Lambda}}_{\mathrm{RS}} \text{ is unrestricted,} \tag{R-RS}$$

[20] Where $\boldsymbol{X}_i$ is $T \times K$.

where $\mathbf{\Phi}_{\mathrm{RS}}$ is a $(T-R_0)\times R_0$ matrix of unrestricted parameters. Let $\Theta_{\tilde{\Lambda}}$ and Θ_F denote the parameter space of $\tilde{\mathbf{\Lambda}}$ and $\boldsymbol{F}$, respectively, and let $\bar{\Theta}_F$ denote the restricted parameter space of $\boldsymbol{F}$ under R-RS. The RS estimator is obtained as

$$(\hat{\boldsymbol{\beta}}_{\mathrm{RS}}, \hat{\tilde{\mathbf{\Lambda}}}_{\mathrm{RS}}, \hat{\boldsymbol{F}}_{\mathrm{RS}}) \coloneqq \underset{\boldsymbol{\beta}\in\Theta_\beta,\ \tilde{\mathbf{\Lambda}}\in\Theta_{\tilde{\Lambda}},\ \boldsymbol{F}\in\bar{\Theta}_F}{\arg\min} \mathcal{Q}_{\mathrm{RS}}(\boldsymbol{\beta}, \tilde{\mathbf{\Lambda}}, \boldsymbol{F}), \tag{4.3}$$

with

$$\begin{aligned}
\mathcal{Q}_{\mathrm{RS}}(\boldsymbol{\beta}, \tilde{\mathbf{\Lambda}}, \boldsymbol{F}) &\coloneqq \boldsymbol{\varphi}_{\mathrm{RS}}^\top(\boldsymbol{\beta}, \tilde{\mathbf{\Lambda}}, \boldsymbol{F})\boldsymbol{W}\boldsymbol{\varphi}_{\mathrm{RS}}(\boldsymbol{\beta}, \tilde{\mathbf{\Lambda}}, \boldsymbol{F})\\
\boldsymbol{\varphi}_{\mathrm{RS}}(\boldsymbol{\beta}, \tilde{\mathbf{\Lambda}}, \boldsymbol{F}) &\coloneqq \frac{1}{nT}\left((\boldsymbol{I}_T\otimes\boldsymbol{\mathcal{X}})^\top \mathrm{vec}(\boldsymbol{Y}-\boldsymbol{X}\cdot\boldsymbol{\beta}) - \mathrm{vec}((\boldsymbol{\mathcal{X}}^\top\boldsymbol{\mathcal{X}})^{\frac{1}{2}}\tilde{\mathbf{\Lambda}}\boldsymbol{F}^\top)\right),
\end{aligned}$$

where $\boldsymbol{W}$ denotes a positive definite weighting matrix.

4.2 ALS Estimator

A different moment condition is studied by Ahn et al. (2013) which takes the form

$$\mathbb{E}\left[(\boldsymbol{\mathcal{V}}_0\otimes\boldsymbol{\mathcal{X}})^\top \mathrm{vec}(\boldsymbol{Y}-\boldsymbol{X}\cdot\boldsymbol{\beta}_0)\right] = \mathbf{0}, \tag{M-ALS}$$

where the $T\times(T-R_0)$ matrix $\boldsymbol{\mathcal{V}}_0$ forms a basis for the left null space of $\boldsymbol{F}_0$. As previously, M-ALS fails to uniquely identify $\boldsymbol{\mathcal{V}}_0$, as $\boldsymbol{\mathcal{V}}_0\boldsymbol{H}$ forms a basis for the left null space of $\boldsymbol{F}_0$ for any $(T-R_0)\times(T-R_0)$ invertible matrix $\boldsymbol{H}$. As a consequence, additional restrictions are adopted. Ahn et al. (2013) consider the following restriction:

$$\boldsymbol{\mathcal{V}} = \begin{pmatrix}\mathbf{\Phi}_{\mathrm{ALS}}\\ -\boldsymbol{I}_{T-R_0}\end{pmatrix}, \tag{R-ALS}$$

where $\mathbf{\Phi}_{\mathrm{ALS}}$ is an $R_0\times(T-R_0)$ matrix of unrestricted parameters. Let $\bar{\mathbf{\Theta}}_{\mathcal{V}}$ denote the restricted parameter space of $\boldsymbol{\mathcal{V}}$ under R-ALS. The ALS estimator is obtained as

$$(\hat{\boldsymbol{\beta}}_{\mathrm{ALS}}, \hat{\boldsymbol{\mathcal{V}}}_{\mathrm{ALS}}) \coloneqq \underset{\boldsymbol{\beta}\in\Theta_\beta,\ \boldsymbol{\mathcal{V}}\in\bar{\Theta}_{\mathcal{V}}}{\arg\min} \mathcal{Q}_{\mathrm{ALS}}(\boldsymbol{\beta}, \boldsymbol{\mathcal{V}}), \tag{4.4}$$

with

$$
\begin{aligned}
\mathcal{Q}_{\mathrm{ALS}}(\boldsymbol{\beta}, \boldsymbol{\mathcal{V}}) &\coloneqq \boldsymbol{\varphi}_{\mathrm{ALS}}^{\top}(\boldsymbol{\beta}, \boldsymbol{\mathcal{V}}) \boldsymbol{W} \boldsymbol{\varphi}_{\mathrm{ALS}}(\boldsymbol{\beta}, \boldsymbol{\mathcal{V}}) \\
\boldsymbol{\varphi}_{\mathrm{ALS}}(\boldsymbol{\beta}, \boldsymbol{\mathcal{V}}) &\coloneqq \frac{1}{nT} \left((\boldsymbol{\mathcal{V}} \otimes \boldsymbol{\mathcal{X}})^{\top} \mathrm{vec}(\boldsymbol{Y} - \boldsymbol{X} \cdot \boldsymbol{\beta}) \right),
\end{aligned}
$$

where $\boldsymbol{W}$ denotes a positive definite weighting matrix.

4.3 TLS Estimator

The TLS estimator can be obtained from the moment condition

$$
\mathbb{E}\left[(\boldsymbol{\mathcal{V}}_0 \otimes \boldsymbol{Q}_{\boldsymbol{\mathcal{X}}})^{\top} \mathrm{vec}(\boldsymbol{Y} - \boldsymbol{X} \cdot \boldsymbol{\beta}_0) \right] = \mathbf{0}, \tag{M-TLS}
$$

which holds under Assumption ED. This is similar to M-ALS, however, the TLS estimator can be understood to utilise an alternative restriction to R-ALS which takes the form

$$
\boldsymbol{\mathcal{V}}^{\top} \boldsymbol{\mathcal{V}} = \boldsymbol{I}_{T-R_0}. \tag{R-TLS}
$$

Under R-TLS it is possible to profile $\boldsymbol{\mathcal{V}}$ out of the objective function to obtain

$$
\begin{aligned}
\hat{\boldsymbol{\beta}} &\coloneqq \underset{\boldsymbol{\beta} \in \Theta_{\beta}}{\arg\min} \left(\min_{\boldsymbol{\mathcal{V}} \in \tilde{\Theta}_{\mathcal{V}}} \mathcal{Q}_{\mathrm{TLS}}(\boldsymbol{\beta}, \boldsymbol{\mathcal{V}}) \right) \\
&= \underset{\boldsymbol{\beta} \in \Theta_{\beta}}{\arg\min} \ \mathcal{Q}(\boldsymbol{\beta}),
\end{aligned}
$$

where $\tilde{\Theta}_{\mathcal{V}}$ denotes the restricted parameter space for $\boldsymbol{\mathcal{V}}$ under R-TLS, and

$$
\begin{aligned}
\mathcal{Q}_{\mathrm{TLS}}(\boldsymbol{\beta}, \boldsymbol{\mathcal{V}}) &\coloneqq \boldsymbol{\varphi}_{\mathrm{TLS}}^{\top}(\boldsymbol{\beta}, \boldsymbol{\mathcal{V}}) \boldsymbol{\varphi}_{\mathrm{TLS}}(\boldsymbol{\beta}, \boldsymbol{\mathcal{V}}) \\
\boldsymbol{\varphi}_{\mathrm{TLS}}(\boldsymbol{\beta}, \boldsymbol{\mathcal{V}}) &\coloneqq \frac{1}{\sqrt{nT}} (\boldsymbol{\mathcal{V}} \otimes \boldsymbol{Q}_{\boldsymbol{\mathcal{X}}})^{\top} \mathrm{vec}(\boldsymbol{Y} - \boldsymbol{X} \cdot \boldsymbol{\beta}).
\end{aligned}
$$

4.4 Asymptotic Comparisons

Though the TLS estimator uses an identity weighting matrix, notice that if the errors $\varepsilon_{it} \sim \mathrm{iid}(0, \sigma_0^2)$ conditional on $\mathcal{D}_{nT}$, then the optimal weighting matrix associated with

M-TLS is (up to scale)

$$\boldsymbol{W}^{*}=\frac{1}{\sigma_0^2}(\boldsymbol{\mathcal{V}}_0^\top\boldsymbol{\mathcal{V}}_0\otimes\boldsymbol{Q}_{\mathfrak{X}}^\top\boldsymbol{Q}_{\mathfrak{X}})^{-1}=\frac{1}{\sigma_0^2}\boldsymbol{I}_{(T-R_0)TK}.$$

Thus, transforming the model by $\boldsymbol{Q}_{\mathfrak{X}}$ tacitly imposes the optimal weighting matrix under homoskedasticity. This, as it turns out, is important to ensure that the estimator remains consistent when $T\to\infty$. Intuitively, as $n,T\to\infty$ and $T/n\to K^{-1}$ the TLS estimator approaches the LS estimator. Since the LS estimator is known to be consistent when both n and T are large, this closeness is desirable, and is a mirror to the relationship between the within estimator and the optimal GMM estimator discussed in Alvarez and Arellano (2003). The aim of this section is to draw comparisons between the RS, ALS, and TLS estimators. However, since the TLS estimator cannot be separated from the way in which it is weighted, comparable one-step ALS and RS estimators are studied. For the RS estimator this amounts to setting $\boldsymbol{W}=(\boldsymbol{I}_T\otimes(\mathfrak{X}^\top\mathfrak{X})^{-1})$, and for the ALS estimator setting $\boldsymbol{W}=(\boldsymbol{I}_{T-R_0}\otimes(\mathfrak{X}^\top\mathfrak{X})^{-1})$. Though insightful, to be clear, the results obtained in this section are specific to this choice of weighting matrix and would not necessarily apply under alternative weighting schemes. In the first result, Proposition 5 establishes that a one-step RS estimator that imposes R-RS is asymptotically equivalent to the TLS estimator.

Proposition 5. *Assume it is possible to decompose $\tilde{\boldsymbol{\Lambda}}_0\boldsymbol{F}_0^\top=\tilde{\boldsymbol{\Lambda}}_*\boldsymbol{F}_*^\top$ such that $\boldsymbol{F}_*\in\bar{\Theta}_F$. Set $\boldsymbol{W}=(\boldsymbol{I}_T\otimes(\mathfrak{X}^\top\mathfrak{X})^{-1})$ and let $\boldsymbol{\theta}:=(\boldsymbol{\beta};\mathrm{vec}(\boldsymbol{\Phi}))$ and $\boldsymbol{\theta}_0:=(\boldsymbol{\beta}_0;\mathrm{vec}(\boldsymbol{\Phi}_*))$, where $\boldsymbol{\Phi}_*$ is the unrestricted block of $\boldsymbol{F}_*$. Let $\Theta_{\theta,\mathrm{RS}}$ denote the parameter space of $\boldsymbol{\theta}$. Assume that as $n\to\infty$ with $T\geq R_0+1$ fixed, the GMM estimator defined by*

$$\hat{\boldsymbol{\theta}}_{\mathrm{RS}}:=\operatorname*{arg\,min}_{\boldsymbol{\theta}\in\Theta_{\theta,\mathrm{RS}}}\left(\min_{\tilde{\boldsymbol{\Lambda}}\in\Theta_{\tilde{\Lambda}}}\mathcal{Q}_{\mathrm{RS}}(\boldsymbol{\beta},\tilde{\boldsymbol{\Lambda}},\boldsymbol{F}(\boldsymbol{\Phi}))\right),^{21}$$

satisfies

$$\sqrt{nT}(\hat{\boldsymbol{\theta}}_{\mathrm{RS}}-\boldsymbol{\theta}_0)\xrightarrow{d}\mathcal{N}\left(\mathbf{0},\mathbb{D}_{\mathrm{RS}}^{-1}\mathbb{V}_{\mathrm{RS}}\mathbb{D}_{\mathrm{RS}}^{-1}\right),$$

with $\mathbb{D}_{\mathrm{RS}}:=\mathrm{plim}_{n\to\infty}\,\mathbb{G}_{\mathrm{RS}}^\top\boldsymbol{W}\mathbb{G}_{\mathrm{RS}}$ and $\mathbb{V}_{\mathrm{RS}}:=\mathrm{plim}_{n\to\infty}\,nT\times\mathbb{G}_{\mathrm{RS}}^\top\boldsymbol{W}\mathbb{H}_{\mathrm{RS}}\boldsymbol{W}\mathbb{G}_{\mathrm{RS}}$, and

[21] Since $\tilde{\boldsymbol{\Lambda}}$ is unrestricted under R-RS, with $\boldsymbol{W}=(\boldsymbol{I}_T\otimes(\mathfrak{X}^\top\mathfrak{X})^{-1})$ it proves convenient to profile $\tilde{\boldsymbol{\Lambda}}$ out of the objective function.

where

$$\mathbb{G}_{\mathrm{RS}} \coloneqq \mathbb{E}[\nabla \boldsymbol{\varphi}_{\mathrm{RS}}(\boldsymbol{\beta}_0, \boldsymbol{F}(\boldsymbol{\Phi}_*))|\mathcal{D}_{nT}]$$
$$\mathbb{H}_{\mathrm{RS}} \coloneqq \mathbb{E}[\boldsymbol{\varphi}_{\mathrm{RS}}(\boldsymbol{\beta}_0, \boldsymbol{F}(\boldsymbol{\Phi}_*))\boldsymbol{\varphi}_{\mathrm{RS}}^\top(\boldsymbol{\beta}_0, \boldsymbol{F}(\boldsymbol{\Phi}_*))|\mathcal{D}_{nT}],$$

and $\boldsymbol{\varphi}_{\mathrm{RS}}(\boldsymbol{\beta}_0, \boldsymbol{F}(\boldsymbol{\Phi}_*))$ *is the moment condition obtained upon profiling out* $\tilde{\boldsymbol{\Lambda}}$*; see the proof details. Under Assumptions ED and AE, as* $n \to \infty$ *with* $T \geq R_0 + 1$ *fixed,*

$$\sqrt{nT}(\hat{\boldsymbol{\beta}}_{\mathrm{RS}} - \boldsymbol{\beta}_0) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbb{D}^{-1}\mathbb{V}\mathbb{D}^{-1}).$$

Proposition 5 establishes that the TLS and one-step RS estimators will share the same asymptotic distribution, despite the RS estimator imposing a restriction of a different nature. This is so because, while R-RS restricts the factors, these will only feature in the asymptotic distribution of $\hat{\boldsymbol{\beta}}$ through the projector $\boldsymbol{M}_{\boldsymbol{F}_*}$ which, provided that it is indeed possible to decompose $\tilde{\boldsymbol{\Lambda}}_0\boldsymbol{F}_0^\top = \tilde{\boldsymbol{\Lambda}}_*\boldsymbol{F}_*^\top$ such that $\boldsymbol{F}_* \in \bar{\Theta}_F$, equals to $\boldsymbol{M}_{\boldsymbol{F}_0}$. This latter point is, however, an important caveat. If this condition does not hold, then M-RS in combination with R-RS may fail to identify $\boldsymbol{\beta}_0$, though it may still be identified under alternative restrictions, such as R-TLS. Identification failures of this kind have been remarked on previously; see, e.g., Hayakawa (2016). In this sense restrictions imposed on the factor term for the purposes of identification are only without loss of generality if these are indeed compatible with the true, unknown factor term.

An obvious question arises as to whether this asymptotic equivalence also holds for the corresponding one-step ALS estimator. Proposition 6 below establishes that this is not necessarily the case.

Proposition 6. *Assume it is possible to decompose* $\tilde{\boldsymbol{\Lambda}}_0\boldsymbol{F}_0^\top = \tilde{\boldsymbol{\Lambda}}_*\boldsymbol{F}_*^\top$ *such that* $\boldsymbol{F}_* \in \bar{\Theta}_F$*. Set* $\boldsymbol{W} = (\boldsymbol{I}_{T-R_0} \otimes (\boldsymbol{\mathfrak{X}}^\top\boldsymbol{\mathfrak{X}})^{-1})$ *and let* $\boldsymbol{\theta} \coloneqq (\boldsymbol{\beta}; \mathrm{vec}(\boldsymbol{\Phi}))$ *and* $\boldsymbol{\theta}_0 \coloneqq (\boldsymbol{\beta}_0; \mathrm{vec}(\boldsymbol{\Phi}_*))$*, where* $\boldsymbol{\Phi}_*$ *is the unrestricted block of* $\boldsymbol{F}_*$*. Let* $\Theta_{\theta,\mathrm{ALS}}$ *denote the parameter space of* $\boldsymbol{\theta}$*. Moreover, let* $\boldsymbol{\mathcal{V}}_* \coloneqq (\boldsymbol{\Phi}_*; -\boldsymbol{I}_{T-R_0})$ *and assume that as* $n \to \infty$ *with* $T \geq R_0 + 1$ *fixed, the GMM estimator defined by*

$$\hat{\boldsymbol{\theta}}_{\mathrm{ALS}} = \operatorname*{arg\,min}_{\boldsymbol{\theta} \in \Theta_{\theta,\mathrm{ALS}}} \mathcal{Q}_{\mathrm{ALS}}(\boldsymbol{\beta}, \boldsymbol{\mathcal{V}}(\boldsymbol{\Phi})),$$

satisfies

$$\sqrt{nT}(\hat{\boldsymbol{\theta}}_{\mathrm{ALS}} - \boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \mathbb{D}_{\mathrm{ALS}}^{-1}\mathbb{V}_{\mathrm{ALS}}\mathbb{D}_{\mathrm{ALS}}^{-1}\right),$$

with $\mathbb{D}_{\mathrm{ALS}} \coloneqq \mathrm{plim}_{n\to\infty}\, \mathbb{G}_{\mathrm{ALS}}^{\top}\boldsymbol{W}\mathbb{G}_{\mathrm{ALS}}$ *and* $\mathbb{V}_{\mathrm{ALS}} \coloneqq \mathrm{plim}_{n\to\infty}\, nT \times \mathbb{G}_{\mathrm{ALS}}^{\top}\boldsymbol{W}\mathbb{H}_{\mathrm{ALS}}\boldsymbol{W}\mathbb{G}_{\mathrm{ALS}}$, *and where*

$$\begin{aligned}
\mathbb{G}_{\mathrm{ALS}} &\coloneqq \mathbb{E}[\nabla\boldsymbol{\varphi}_{\mathrm{ALS}}(\boldsymbol{\beta}_0, \boldsymbol{\mathcal{V}}(\boldsymbol{\Phi}_*))|\mathcal{D}_{nT}] \\
\mathbb{H}_{\mathrm{ALS}} &\coloneqq \mathbb{E}[\boldsymbol{\varphi}_{\mathrm{ALS}}(\boldsymbol{\beta}_0, \boldsymbol{\mathcal{V}}(\boldsymbol{\Phi}_*))\boldsymbol{\varphi}_{\mathrm{ALS}}^{\top}(\boldsymbol{\beta}_0, \boldsymbol{\mathcal{V}}(\boldsymbol{\Phi}_*))|\mathcal{D}_{nT}].
\end{aligned}$$

Under Assumptions ED and AE, as $n \to \infty$ *with* $T \geq R_0 + 1$ *fixed,*

$$\sqrt{nT}(\hat{\boldsymbol{\beta}}_{\mathrm{ALS}} - \boldsymbol{\beta}_0) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbb{D}_*^{-1}\mathbb{V}_*\mathbb{D}_*^{-1}), \tag{4.5}$$

where

$$\begin{aligned}
\mathbb{D}_* &\coloneqq \operatorname*{plim}_{n\to\infty} \frac{1}{nT}\tilde{\boldsymbol{\mathcal{X}}}^{\top}(\boldsymbol{\mathcal{V}}_*\boldsymbol{\mathcal{V}}_*^{\top} \otimes \boldsymbol{M}_{\tilde{\boldsymbol{\Lambda}}_0})\tilde{\boldsymbol{\mathcal{X}}} \\
\mathbb{V}_* &\coloneqq \operatorname*{plim}_{n\to\infty} \frac{1}{nT}\tilde{\boldsymbol{\mathcal{X}}}^{\top}(\boldsymbol{\mathcal{V}}_*\boldsymbol{\mathcal{V}}_*^{\top} \otimes \boldsymbol{M}_{\tilde{\boldsymbol{\Lambda}}_0})\boldsymbol{\Sigma}_{\mathcal{D}}(\boldsymbol{\mathcal{V}}_*\boldsymbol{\mathcal{V}}_*^{\top} \otimes \boldsymbol{M}_{\tilde{\boldsymbol{\Lambda}}_0})\tilde{\boldsymbol{\mathcal{X}}}.
\end{aligned}$$

The asymptotic distribution of the one-step ALS estimator will generally not coincide with that of the TLS estimator (nor indeed the one-step RS estimator under R-RS), unless $\boldsymbol{\mathcal{V}}_*^{\top}\boldsymbol{\mathcal{V}}_* = \boldsymbol{I}_{T-R_0}$, i.e. unless $\boldsymbol{\mathcal{V}}_*$ forms an *orthonormal* basis for the left null space of the factors.[22] Notice that Proposition 6 also assumes that it is possible to decompose the factor term in the manner of R-RS. This is because the existence of a matrix $\boldsymbol{\mathcal{V}} \in \bar{\Theta}_{\mathcal{V}}$ such that $\boldsymbol{\mathcal{V}}^{\top}\boldsymbol{F}_0 = \mathbf{0}$ is equivalent to the assumption that it is possible to decompose $\tilde{\boldsymbol{\Lambda}}_0\boldsymbol{F}_0^{\top} = \tilde{\boldsymbol{\Lambda}}_*\boldsymbol{F}_*^{\top}$ such that $\mathrm{vec}(\boldsymbol{F}_*) \in \bar{\Theta}_F$.[23]

Overall the results of this section illustrate the importance of both the moment condition *and* any adopted normalisation in determining the properties of the estimator. The TLS and ALS estimators consist of a moment condition and a particular normalisation. Although the moment conditions are similar, the difference in their normalisations is important and ultimately may result in these estimators having different asymptotic distributions (Proposition 6). The RS estimator is not coupled with a particular normal-

[22] If $T - R_0 = 1$ then $\boldsymbol{\mathcal{V}}_*^{\top}\boldsymbol{\mathcal{V}}_*$ will be a scalar and the covariance matrix will equal to that of the TLS estimator.

[23] See Appendix A.3.

isation, however, if one considers a normalisation which is, to some extent, equivalent to that used by ALS, the asymptotic distribution of the resultant RS estimator coincides with that of the TLS estimator and not the ALS estimator (Proposition 5).[24] As a final observation, notice that even under homoskedasticity the asymptotic variance of the ALS estimator would not collapse to $\mathbb{D}_*^{-1}$ unless $\boldsymbol{\mathcal{V}}_*$ is orthonormal, and therefore the TLS estimator will be more efficient. This is formalised in the following result.

Proposition 7. *Assume $\boldsymbol{\Sigma}_{\mathcal{D}} = \sigma_0^2 \boldsymbol{I}_{nT}$, and there exist nonstochastic matrices $\mathbb{D}$, $\mathbb{D}_*$, $\mathbb{V}$, and $\mathbb{V}_*$, such that $\boldsymbol{D} \xrightarrow{p} \mathbb{D}$, $\boldsymbol{D}_* \xrightarrow{p} \mathbb{D}_*$, $\boldsymbol{V} \xrightarrow{p} \mathbb{V}$, and $\boldsymbol{V}_* \xrightarrow{p} \mathbb{V}_*$ as $n \to \infty$ with $T \geq R_0 + 1$ fixed, and the eigenvalues of $\mathbb{D}$, $\mathbb{D}_*$, $\mathbb{V}$, and $\mathbb{V}_*$ are bounded away from zero and from above by a constant. Moreover, assume*

$$\mathrm{avar}(\sqrt{nT}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)) = \mathbb{D}^{-1}\mathbb{V}\mathbb{D}^{-1}$$
$$\mathrm{avar}(\sqrt{nT}(\hat{\boldsymbol{\beta}}_{\mathrm{ALS}} - \boldsymbol{\beta}_0)) = \mathbb{D}_*^{-1}\mathbb{V}_*\mathbb{D}_*^{-1}.$$

Then

$$\mathrm{avar}(\sqrt{nT}(\hat{\boldsymbol{\beta}}_{\mathrm{ALS}} - \boldsymbol{\beta}_0)) \succeq \mathrm{avar}(\sqrt{nT}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)),$$

where avar$(\cdot)$ *denotes asymptotic variance.*

5 Further Matter

5.1 Inference

This subsection describes how to proceed with asymptotic inference. Two settings are considered, in turn. The first case corresponds to the large n, fixed T setting, while the second case corresponds to the large n, large T setting. In the interest of space, a single set of conditions is presented under which the results in the following subsections can be obtained. Let $\Gamma_{b_T}(\boldsymbol{A}) := \boldsymbol{A} \odot (\boldsymbol{\Omega} \otimes \boldsymbol{I}_n)$ for an $nT \times nT$ matrix $\boldsymbol{A}$, with $\boldsymbol{\Omega}$ being a $T \times T$ matrix with elements $\omega_{t_1 t_2} = 1\{|t_1 - t_2| < b_T\}$, and b_T is a positive integer-valued sequence.

[24] Propositions 5, 6, and 7 establish asymptotic results for ALS and RS estimators that do not employ optimal weighting matrices. If, instead, one considers optimal GMM estimators, then the RS and ALS estimators are asymptotically equivalent which follows from Theorem 4 of Robertson and Sarafidis (2015). Further, one can show that these are, in turn, equivalent to the GTLS estimator described in Section 3.4.

Assumption IF.

(i) The elements of $\boldsymbol{X}_k$ and $\boldsymbol{M}_{\boldsymbol{P}_{\mathcal{X}}\boldsymbol{\Lambda}_0}\boldsymbol{X}_k\boldsymbol{M}_{\boldsymbol{F}_0}$ have uniformly bounded eighth moments.

(ii) $n^{-1}\boldsymbol{\Lambda}_0^\top\boldsymbol{\Lambda}_0 \xrightarrow{p} \boldsymbol{\Sigma}_{\boldsymbol{\Lambda}_0}$ as $n \to \infty$, where the eigenvalues of $\boldsymbol{\Sigma}_{\boldsymbol{\Lambda}_0}$ are bounded from above by a constant.

(iii) Conditional on $\mathcal{D}_{nT}$, ε_{it} are independent over i, with $\mathbb{E}[\varepsilon_{it}|\mathcal{D}_{nT}] = 0$, $\mathbb{E}[\varepsilon_{it}^2|\mathcal{D}_{nT}] > 0$, and $\sup_{\|\boldsymbol{v}\|_2=1}\mathbb{E}[(\boldsymbol{v}^\top\boldsymbol{\varepsilon}_i)^{16}|\mathcal{D}_{nT}]$ uniformly bounded for $\mathcal{D}_{nT}$-measurable vectors $\boldsymbol{v}$.

(iv) $\|\Gamma_{b_T}(\boldsymbol{\Sigma}_{\mathcal{D}}) - \boldsymbol{\Sigma}_{\mathcal{D}}\|_2 = \mathcal{O}_p(1)$ as $n, T, b_T \to \infty$ with $T/n \to \gamma \in (0, \infty)$ and $b_T^8/n \to 0$.

Assumption IF imposes more stringent conditions on several variables. Weaker restrictions can be imposed upon the errors, at the expense of more restrictive conditions on the factors and loadings; for example, that these are uniformly bounded as in Moon and Weidner (2017). Appendix A.4 verifies IF(iii) and IF(iv) under first order serial correlation as an illustration.

5.1.1 Inference under Large n, Fixed T

When T is fixed the TLS estimator is asymptotically unbiased. Therefore, in order to proceed with asymptotic inference, one need only obtain consistent estimators of the asymptotic covariance matrices. Notice, however, that minimisers of the objective function (2.3) with respect to the transformed factor loadings and the factors are not unique. In order to resolve this indeterminacy, estimates of the transformed factor loadings and the factors may be defined in the following manner. Consider a singular value decomposition $(\tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{X}} \cdot \hat{\boldsymbol{\beta}}) =: \sum_{t=1}^T s_t \boldsymbol{u}_t \boldsymbol{v}_t^\top$ with singular values $s_T \leq \ldots \leq s_1$. Define $\hat{\tilde{\boldsymbol{\Lambda}}} := \sqrt{n}(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_{R_0})$ and $\hat{\boldsymbol{F}} := (s_1\boldsymbol{v}_1, \ldots, s_{R_0}\boldsymbol{v}_{R_0})/\sqrt{n}$. Although these estimators will not, in general, be consistent for $\boldsymbol{F}_0$ and $\tilde{\boldsymbol{\Lambda}}_0$ themselves, they will produce consistent estimators of projectors $\boldsymbol{M}_{\boldsymbol{F}_0}$ and $\boldsymbol{M}_{\tilde{\boldsymbol{\Lambda}}_0}$. Let

$$\hat{\boldsymbol{D}} := \frac{1}{nT}\tilde{\boldsymbol{\mathcal{X}}}^\top(\boldsymbol{M}_{\hat{\boldsymbol{F}}} \otimes \boldsymbol{M}_{\hat{\tilde{\boldsymbol{\Lambda}}}})\tilde{\boldsymbol{\mathcal{X}}}$$
$$\hat{\boldsymbol{V}} := \frac{1}{nT}\tilde{\boldsymbol{\mathcal{X}}}^\top(\boldsymbol{M}_{\hat{\boldsymbol{F}}} \otimes \boldsymbol{M}_{\hat{\tilde{\boldsymbol{\Lambda}}}})\hat{\tilde{\boldsymbol{\Sigma}}}(\boldsymbol{M}_{\hat{\boldsymbol{F}}} \otimes \boldsymbol{M}_{\hat{\tilde{\boldsymbol{\Lambda}}}})\tilde{\boldsymbol{\mathcal{X}}},$$

where $\hat{\hat{\boldsymbol{\Sigma}}} := (\boldsymbol{I}_T \otimes \boldsymbol{Q}_{\mathcal{X}}^\top)\Gamma_T(\text{vec}(\hat{\boldsymbol{e}})\text{vec}(\hat{\boldsymbol{e}})^\top)(\boldsymbol{I}_T \otimes \boldsymbol{Q}_{\mathcal{X}})$ and $\hat{\boldsymbol{e}} := (\boldsymbol{Y} - \boldsymbol{X} \cdot \hat{\boldsymbol{\beta}})\boldsymbol{M}_{\hat{\boldsymbol{F}}}$.

Proposition 8. *Under Assumptions MD, CS, AE, AD, and IF, as $n \to \infty$ with $T \geq T_{\min}$ fixed,*

$$\|\hat{\boldsymbol{D}} - \boldsymbol{D}\|_2 = o_p(1)$$
$$\|\hat{\boldsymbol{V}} - \boldsymbol{V}\|_2 = o_p(1).$$

5.1.2 Inference under Large n, Large T

Under an asymptotic where $n, T \to \infty$ and $T/n \to \gamma \in (0, \infty)$, the TLS estimator is asymptotically biased. This subsection shows that this bias can be consistently estimated and a bias-corrected TLS estimator can be constructed. This, in combination with consistent estimators of the asymptotic covariance matrices, paves the way for asymptotically valid inference. Consider a singular value decomposition $(\boldsymbol{Y} - \boldsymbol{X} \cdot \hat{\boldsymbol{\beta}}) =: \sum_{t=1}^{T} s_t \boldsymbol{u}_t \boldsymbol{v}_t^\top$ with singular values $s_T \leq \ldots \leq s_1$. Then define $\check{\boldsymbol{F}} := (s_1 \boldsymbol{v}_1, \ldots, s_{R_0} \boldsymbol{v}_{R_0})/\sqrt{n}$,

$$\begin{aligned}
\hat{\boldsymbol{D}} &:= \frac{1}{nT} \tilde{\boldsymbol{\mathcal{X}}}^\top (\boldsymbol{M}_{\hat{\boldsymbol{F}}} \otimes \boldsymbol{M}_{\hat{\hat{\boldsymbol{\Lambda}}}}) \tilde{\boldsymbol{\mathcal{X}}} \\
\hat{\boldsymbol{V}} &:= \frac{1}{nT} \tilde{\boldsymbol{\mathcal{X}}}^\top (\boldsymbol{M}_{\hat{\boldsymbol{F}}} \otimes \boldsymbol{M}_{\hat{\hat{\boldsymbol{\Lambda}}}}) \check{\hat{\boldsymbol{\Sigma}}} (\boldsymbol{M}_{\hat{\boldsymbol{F}}} \otimes \boldsymbol{M}_{\hat{\hat{\boldsymbol{\Lambda}}}}) \tilde{\boldsymbol{\mathcal{X}}} \\
\hat{\psi}_k^{(1)} &:= \frac{1}{\sqrt{nT}} \text{tr}(\check{\hat{\boldsymbol{\Sigma}}}(\boldsymbol{I}_T \otimes \boldsymbol{M}_{\hat{\hat{\boldsymbol{\Lambda}}}} \tilde{\boldsymbol{X}}_k \hat{\boldsymbol{F}} (\hat{\boldsymbol{F}}^\top \hat{\boldsymbol{F}})^{-1} (\hat{\hat{\boldsymbol{\Lambda}}}^\top \hat{\hat{\boldsymbol{\Lambda}}})^{-1} \hat{\hat{\boldsymbol{\Lambda}}}^\top)) \\
\hat{\psi}_k^{(2)} &:= \frac{1}{\sqrt{nT}} \text{tr}(\check{\hat{\boldsymbol{\Sigma}}}(\hat{\boldsymbol{F}} (\hat{\boldsymbol{F}}^\top \hat{\boldsymbol{F}})^{-1} (\hat{\hat{\boldsymbol{\Lambda}}}^\top \hat{\hat{\boldsymbol{\Lambda}}})^{-1} \hat{\hat{\boldsymbol{\Lambda}}}^\top \tilde{\boldsymbol{X}}_k \boldsymbol{M}_{\hat{\boldsymbol{F}}} \otimes \boldsymbol{I}_{TK})),
\end{aligned}$$

where $\check{\hat{\boldsymbol{\Sigma}}} := (\boldsymbol{I}_T \otimes \boldsymbol{Q}_{\mathcal{X}}^\top)\Gamma_{b_T}(\text{vec}(\check{\boldsymbol{e}})\text{vec}(\check{\boldsymbol{e}})^\top)(\boldsymbol{I}_T \otimes \boldsymbol{Q}_{\mathcal{X}})$ and $\check{\boldsymbol{e}} := (\boldsymbol{Y} - \boldsymbol{X} \cdot \hat{\boldsymbol{\beta}})\boldsymbol{M}_{\check{\boldsymbol{F}}}$.

Proposition 9. *Under Assumptions MD, CS, AE, AD, and IF, as $n, T \to \infty$ with $T/n \to \gamma \in (0, \infty)$,*

$$\|\hat{\boldsymbol{\psi}}^{(1)} - \boldsymbol{\psi}^{(1)}\|_2 = o_p(1)$$
$$\|\hat{\boldsymbol{\psi}}^{(2)} - \boldsymbol{\psi}^{(2)}\|_2 = o_p(1)$$
$$\|\hat{\boldsymbol{D}} - \boldsymbol{D}\|_2 = o_p(1)$$
$$\|\hat{\boldsymbol{V}} - \boldsymbol{V}\|_2 = o_p(1).$$

5.2 Estimating the Number of Factors

The result established in Section 3.1 demonstrates that the TLS estimator will remain consistent with the number of factors overestimated. However, the asymptotic distribution of the estimator is characterised under the assumption that $R_e = R_0$, and therefore where the true number of factors is not known in advance, it is necessary to estimate this from the data before proceeding with inference. Moreover, overestimation of the number of factors will typically lead to a loss of efficiency in finite samples and therefore it is still desirable to input the correct number of factors even if interest lies primarily in point estimation.[25] 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 this method can be applied to detect the number of factors in the present context.

Let ϱ_n be a sequence depending on n (and possibly also T) that tends towards zero. Define

$$\mu_r^* := \mu_r \left(\frac{1}{nT} \left(\tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{X}} \cdot \hat{\boldsymbol{\beta}} \right)^\top \left(\tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{X}} \cdot \hat{\boldsymbol{\beta}} \right) + \varrho_n^2 \boldsymbol{I}_T \right), \tag{5.1}$$

that is, μ_r^* is the r-th largest eigenvalue of the bracketed matrix on the right. Thereafter let

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

The main departure from the original test described by Ahn and Horenstein (2013) is the addition of the matrix $\varrho_n^2 \boldsymbol{I}_T$. This is used to control the rate at which the eigenvalues of the matrix in (5.1) can approach zero. Intuitively, the idea behind the eigenvalue ratio test is that the first R_0 eigenvalues of the matrix $(nT)^{-1}(\tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{X}} \cdot \hat{\boldsymbol{\beta}})^\top(\tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{X}} \cdot \hat{\boldsymbol{\beta}})$ should be of a similar magnitude and large, while the remaining $(T - R_0)$ should be of a similar magnitude but small. As such, the largest ratio between consecutive eigenvalues should

[25] In particular Lu and Su (2016) provide further discussion of this issue, as well as an alternative way to proceed.

reveal the true number of factors. Placed into the present context, Ahn and Horenstein (2013) study the model with $\boldsymbol{\beta}_0 = \mathbf{0}$, i.e.,

$$\boldsymbol{Y} = \boldsymbol{\Lambda}_0 \boldsymbol{F}_0^\top + \boldsymbol{\varepsilon},$$

where the properties of the $(T - R_0)$ smallest eigenvalues of the matrix $(nT)^{-1}\boldsymbol{Y}^\top\boldsymbol{Y}$ can be deduced through the properties of the singular values of the error $\boldsymbol{\varepsilon}$. In particular, under certain conditions, the smallest eigenvalues can be shown to be similar in magnitude. However, in this context the constructed factor model takes the form

$$\tilde{\boldsymbol{Y}} = \tilde{\boldsymbol{\Lambda}}_0 \boldsymbol{F}_0^\top + \tilde{\boldsymbol{e}} \qquad \text{with} \qquad \tilde{\boldsymbol{e}} := \tilde{\boldsymbol{X}} \cdot (\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}}) + \tilde{\boldsymbol{\varepsilon}}.$$

It is more difficult to establish the properties of the singular values of the transformed error $\tilde{\boldsymbol{\varepsilon}}$ without imposing further restrictions, in addition to which the term $\tilde{\boldsymbol{X}} \cdot (\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}})$ now also appears.[26] Instead, the sequence ϱ_n acts to regularise the eigenvalues and thereby avoids the need to impose more stringent conditions. In aid of the following let $(nT)^{-\frac{1}{2}}\|\tilde{\boldsymbol{\varepsilon}}\|_2 = \mathcal{O}_p(r_{nT})$.

Proposition 10. *Assume* $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2 = \mathcal{O}_p(r_{nT})$ *and* $r_{nT}, \varrho_n \to 0$ *with* $\varrho_n^{-1} r_{nT} = \mathcal{O}(1)$ *as* $n \to \infty$*, with* $T \geq T_{\min}$ *fixed or* $T \to \infty$*. Moreover, assume* $R_0 \geq 1$*. Under Assumptions MD and AE, as* $n \to \infty$*, with* $T \geq T_{\min}$ *fixed or* $T \to \infty$*,*

$$\Pr\left(\max_{1 \leq r \leq T-1} \mathrm{EigR}(r) = R_0\right) \to 1. \tag{5.2}$$

Following Ahn and Horenstein (2013), the possibility of $R_0 = 0$ can be accommodated by introducing a 'mock eigenvalue'. For example, set $\mu_0^* := \varrho_n$. If $R_0 > 0$ then $\mu_0^*/\mu_1^* = o_p(1)$, while if $R_0 = 0$, $\mu_0^*/\mu_1^* = \mathcal{O}_p(\varrho_n^{-1})$.[27] Notice also that $\varrho_n^{-1} r_{nT} = \mathcal{O}(1)$ allows for $\varrho_n^{-1} r_{nT} = o(1)$ and therefore Proposition 10 only requires that ϱ_n diminish no faster than r_{nT}. For example, one can establish $\|\tilde{\boldsymbol{\varepsilon}}\|_2 = \mathcal{O}_p(T^{\frac{3}{4}})$ under Assumption EC or ED, and therefore $\varrho_n = T^{\frac{1}{4}}/\sqrt{n}$ would be permissible provided $T = o(n^2)$, which includes both where T is fixed, and where $n, T \to \infty$ and $T/n \to \gamma \in [0, \infty)$.

[26] For example, if $\boldsymbol{\varepsilon}$ is assumed to be subgaussian then the results of Vershynin (2012) can be applied to establish properties of the singular values of $\tilde{\boldsymbol{\varepsilon}}$.

[27] An alternative means of detecting the correct number of factors can be based on the J-statistic described in Higgins (2023).

5.3 Dynamic Model

The TLS estimator can also be applied to estimate models in which lagged outcomes appear as regressors. To formalise this, consider the model

$$\begin{aligned}\boldsymbol{Y} &= \alpha_0 \boldsymbol{Y}_{-1} + \boldsymbol{X} \cdot \boldsymbol{\beta}_0 + \boldsymbol{\Lambda}_0 \boldsymbol{F}_0^\top + \boldsymbol{\varepsilon} \\ &=: \boldsymbol{Z} \cdot \boldsymbol{\theta}_0 + \boldsymbol{\Lambda}_0 \boldsymbol{F}_0^\top + \boldsymbol{\varepsilon},\end{aligned}$$

where $\theta_{0,1} \coloneqq \alpha_0$, $\boldsymbol{Z}_1 \coloneqq \boldsymbol{Y}_{-1} \coloneqq (\boldsymbol{y}_0, \ldots, \boldsymbol{y}_{T-1})$, and $\theta_{0,k+1} = \beta_{0,k}$, $\boldsymbol{Z}_{k+1} = \boldsymbol{X}_k$ for $k = 1, \ldots, K$. Let $\mathcal{E}_{nT}$ denote $\sigma(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_K, \tilde{\boldsymbol{y}}_0, \tilde{\boldsymbol{\Lambda}}_0, \boldsymbol{F}_0)$. Moreover, let $\boldsymbol{\pi}_J$ be a $J \times 1$ vector of all zeros, except the first element which equals to 1, and $\boldsymbol{W}$ be a $T \times T$ shift matrix which consists of zeros everywhere, except those elements directly above the main diagonal which equal to 1, $\boldsymbol{G}(\alpha) \coloneqq (\boldsymbol{I}_T - \alpha \boldsymbol{W})^{-1} \boldsymbol{W}$, $\boldsymbol{G} \coloneqq \boldsymbol{G}(\alpha_0)$, $\boldsymbol{g}(\alpha) \coloneqq (\boldsymbol{I}_T + \alpha \boldsymbol{G}(\alpha))^\top \boldsymbol{\pi}_T$, $\boldsymbol{g} \coloneqq \boldsymbol{g}(\alpha_0)$, $\boldsymbol{H}_1 \coloneqq \boldsymbol{X} \cdot \boldsymbol{\beta}_0 + \boldsymbol{y}_0 \boldsymbol{g}^\top$, and $\boldsymbol{H}_{k+1} \coloneqq \boldsymbol{X}_k$ for $k = 1, \ldots, K$. $\boldsymbol{\Sigma}_{\mathcal{E}} \coloneqq \mathbb{E}[\mathrm{vec}(\boldsymbol{\varepsilon})\mathrm{vec}(\boldsymbol{\varepsilon})^\top | \mathcal{E}_{nT}]$, and $\tilde{\boldsymbol{\Sigma}}_{\mathcal{E}} \coloneqq (\boldsymbol{I}_T \otimes \boldsymbol{Q}_{\mathcal{X}}^\top) \boldsymbol{\Sigma}_{\mathcal{E}} (\boldsymbol{I}_T \otimes \boldsymbol{Q}_{\mathcal{X}})$.

In order to obtain the following result, Assumptions MD, ED, CS, AE, and AD need to be extended to accommodate lagged outcomes appearing as regressors. These are referred to as MD*, ED*, CS*, AE*, and AD*. For the sake of brevity, these are presented in Appendix E. These extended assumptions are, for the most part, quite minor modifications. However, it is worth remarking further on Assumption CS*(ii) which requires

$$\min_{\boldsymbol{\delta} \in \mathbb{R}^{K+1} : \|\boldsymbol{\delta}\|_2 = 1} \sum_{t=T_{\min}}^{T} \mu_t \left(\frac{1}{nT} (\tilde{\boldsymbol{Z}} \cdot \boldsymbol{\delta})^\top (\tilde{\boldsymbol{Z}} \cdot \boldsymbol{\delta}) \right) \geq b > 0,$$

w.p.a.1 as $n \to \infty$, with $T \geq T_{\min}$ fixed or $T \to \infty$, and where $\tilde{\boldsymbol{Z}} \cdot \boldsymbol{\delta} \coloneqq \sum_{k=1}^{K+1} \delta_k \tilde{\boldsymbol{Z}}_k$. An important implication of this is that at least one element of $\boldsymbol{\beta}_0$ must be nonzero to ensure that the covariates can be used to construct a valid set of instruments for the lagged outcome.

Theorem 2. *Assume $\|\boldsymbol{c}_+\|_2 = \mathcal{O}_p(1)$.*[28] *Under Assumptions MD*, ED*, CS*, AE*, and AD*, as $n \to \infty$,*

[28] $\boldsymbol{c}_+$ is defined in Appendix E.

(i) *with* $T \geq T_{\min}$ *fixed or* $T \to \infty$ *and* $T/n \to 0$,

$$\sqrt{nT}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbb{D}^{-1}\mathbb{V}\mathbb{D}^{-1}),$$

(ii) *with* $T \to \infty$ *and* $T/n \to \gamma \in (0, \infty)$,

$$\sqrt{nT}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) + \boldsymbol{D}_+^{-1}(\boldsymbol{\psi}^{(0)} + \boldsymbol{\psi}^{(1)} + \boldsymbol{\psi}^{(2)}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbb{D}_+^{-1}\mathbb{V}_+\mathbb{D}_+^{-1}),$$

where

$$\begin{aligned}
\boldsymbol{\psi}^{(0)} &:= \frac{1}{\sqrt{nT}}\mathrm{tr}(\tilde{\boldsymbol{\Sigma}}_{\mathcal{E}}((\boldsymbol{P}_{\boldsymbol{F}_0}\boldsymbol{G}\boldsymbol{M}_{\boldsymbol{F}_0} + \boldsymbol{G}\boldsymbol{P}_{\boldsymbol{F}_0}) \otimes \boldsymbol{I}_{TK}))\boldsymbol{\pi}_{K+1} \\
\psi_k^{(1)} &:= \frac{1}{\sqrt{nT}}\mathrm{tr}(\tilde{\boldsymbol{\Sigma}}_{\mathcal{E}}(\boldsymbol{I}_T \otimes \boldsymbol{M}_{\tilde{\boldsymbol{\Lambda}}_0}\tilde{\boldsymbol{H}}_k\boldsymbol{F}_0(\boldsymbol{F}_0^\top\boldsymbol{F}_0)^{-1}(\tilde{\boldsymbol{\Lambda}}_0^\top\tilde{\boldsymbol{\Lambda}}_0)^{-1}\tilde{\boldsymbol{\Lambda}}_0^\top)) \\
\psi_k^{(2)} &:= \frac{1}{\sqrt{nT}}\mathrm{tr}(\tilde{\boldsymbol{\Sigma}}_{\mathcal{E}}(\boldsymbol{F}_0(\boldsymbol{F}_0^\top\boldsymbol{F}_0)^{-1}(\tilde{\boldsymbol{\Lambda}}_0^\top\tilde{\boldsymbol{\Lambda}}_0)^{-1}\tilde{\boldsymbol{\Lambda}}_0^\top\tilde{\boldsymbol{H}}_k\boldsymbol{M}_{\boldsymbol{F}_0} \otimes \boldsymbol{I}_{TK})),
\end{aligned}$$

and $\mathbb{D}$, $\mathbb{D}_+$, $\mathbb{V}$, *and* $\mathbb{V}_+$ *are defined in Assumption AD*(ii).*

The most significant difference when comparing Theorem 2 to Theorem 1 is the additional bias term $\boldsymbol{\psi}^{(0)}$. This arises due to the presence of the lagged outcome and is the analogue of the bias characterised in Moon and Weidner (2017) for the LS estimator. Comparing the order of the biases:

	$\boldsymbol{\psi}_\bullet^{(0)}$	$\boldsymbol{\psi}_\bullet^{(1)}$	$\boldsymbol{\psi}_\bullet^{(2)}$
LS Estimator	$\mathcal{O}_p\left(\sqrt{\frac{n}{T}}\right)$	$\mathcal{O}_p\left(\sqrt{\frac{T}{n}}\right)$	$\mathcal{O}_p\left(\sqrt{\frac{n}{T}}\right)$
TLS Estimator	$\mathcal{O}_p\left(\min\left\{\sqrt{\frac{n}{T}}, \sqrt{\frac{T}{n}}\right\}\right)$	$\mathcal{O}_p\left(\sqrt{\frac{T}{n}}\right)$	$\mathcal{O}_p\left(\min\left\{\sqrt{\frac{n}{T}}, \sqrt{\frac{T}{n}}\right\}\right)$.

Thus, it remains the case that the TLS estimator is asymptotically unbiased when $n, T \to \infty$ and $T/n \to 0$, unlike the LS estimator. It is, however, important to recognise that the nature of the bias $\boldsymbol{\psi}^{(0)}$ is somewhat different from that of the other two biases $\boldsymbol{\psi}^{(1)}$ and $\boldsymbol{\psi}^{(2)}$. To appreciate this, consider the case in which $\varepsilon_{it} \sim \mathrm{iid}(0, \sigma_0^2)$, and the true factors and loadings take the form of individual effects, that is,

$$\boldsymbol{\lambda}_0 := \begin{pmatrix} \lambda_{0,1} \\ \vdots \\ \lambda_{0,n} \end{pmatrix}, \qquad \boldsymbol{f}_0 := \boldsymbol{\iota}_T.$$

In this case $\boldsymbol{\psi}^{(1)} = \boldsymbol{\psi}^{(2)} = \mathbf{0}$ since $\boldsymbol{\Sigma}_{\mathcal{E}} \propto \boldsymbol{I}_{nT}$, leaving the only remaining bias as $\boldsymbol{\psi}^{(0)}$. The fact that $\boldsymbol{\psi}^{(0)}$ remains nonzero highlights that this bias originates from correlation between the errors and the lagged outcomes, unlike $\boldsymbol{\psi}^{(1)}$ and $\boldsymbol{\psi}^{(2)}$ which arise due to the properties of the error term itself.

It is also interesting to notice that when the true factors and factor loadings take the form of individual effects the expression for $\psi_1^{(0)}$ collapses to

$$\psi_1^{(0)} := \frac{\sigma_0^2}{\sqrt{nT}} \frac{1}{T} \mathrm{tr}(\boldsymbol{P}_{\boldsymbol{\mathcal{X}}}) \mathrm{tr}(\boldsymbol{G} \boldsymbol{\iota}_T \boldsymbol{\iota}_T^\top).$$

A bit of algebra reveals that

$$\psi_1^{(0)} = \min\left\{ \sqrt{\frac{n}{T}}, K\sqrt{\frac{T}{n}} \right\} \frac{\sigma_0^2}{(1-\alpha_0)} \left(1 - \frac{1}{T}\frac{(1-\alpha_0^T)}{1-\alpha_0}\right), \tag{5.3}$$

which follows because $\mathrm{tr}(\boldsymbol{P}_{\boldsymbol{\mathcal{X}}}) = \min\{n, TK\}$. This again highlights the significance of the transformation $\boldsymbol{Q}_{\boldsymbol{\mathcal{X}}}$. Without this

$$\psi_1^{(0)} = \sqrt{\frac{n}{T}} \frac{\sigma_0^2}{(1-\alpha_0)} \left(1 - \frac{1}{T}\frac{(1-\alpha_0^T)}{1-\alpha_0}\right), \tag{5.4}$$

which matches (up to scale) the familiar expression derived in Nickell (1981).

Similar to the result in Proposition 3, it is also insightful to compare the asymptotic variance of the LS and TLS estimators under homoskedasticity of the errors. Let $\bar{\boldsymbol{H}}_1 := \boldsymbol{X} \cdot \boldsymbol{\beta}_0$, $\bar{\boldsymbol{H}}_{k+1} := \boldsymbol{X}_k$ for $k = 1, \ldots, K$, $\bar{\boldsymbol{\mathcal{H}}} := (\mathrm{vec}(\bar{\boldsymbol{H}}_1), \ldots, \mathrm{vec}(\bar{\boldsymbol{H}}_{K+1}))$, and define

$$\begin{aligned} \bar{\boldsymbol{D}}_+ &:= \frac{1}{nT} \tilde{\bar{\boldsymbol{\mathcal{H}}}}^\top (\boldsymbol{M}_{\boldsymbol{F}_0} \otimes \boldsymbol{M}_{\tilde{\boldsymbol{\Lambda}}_0}) \tilde{\bar{\boldsymbol{\mathcal{H}}}} + \frac{1}{nT} \mathrm{tr}(\tilde{\boldsymbol{\Sigma}}_{\mathcal{E}} (\boldsymbol{G}\boldsymbol{G}^\top \otimes \boldsymbol{I}_{TK})) \boldsymbol{\pi}_{K+1} \boldsymbol{\pi}_{K+1}^\top, \\ \bar{\boldsymbol{D}}_{+,\mathrm{LS}} &:= \frac{1}{nT} \bar{\boldsymbol{\mathcal{H}}}^\top (\boldsymbol{M}_{\boldsymbol{F}_0} \otimes \boldsymbol{M}_{\boldsymbol{\Lambda}_0}) \bar{\boldsymbol{\mathcal{H}}} + \frac{1}{nT} \mathrm{tr}(\boldsymbol{\Sigma}_{\mathcal{E}^*} (\boldsymbol{G}\boldsymbol{G}^\top \otimes \boldsymbol{I}_n)) \boldsymbol{\pi}_{K+1} \boldsymbol{\pi}_{K+1}^\top, \end{aligned}$$

with $\mathcal{E}_{nT}^* := \sigma(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_K, \boldsymbol{y}_0, \boldsymbol{\Lambda}_0, \boldsymbol{F}_0)$ and $\boldsymbol{\Sigma}_{\mathcal{E}^*} := \mathbb{E}[\mathrm{vec}(\boldsymbol{\varepsilon})\mathrm{vec}(\boldsymbol{\varepsilon})^\top | \mathcal{E}_{nT}^*]$.

Proposition 11. *Assume* $\boldsymbol{\Sigma}_{\mathcal{E}} = \boldsymbol{\Sigma}_{\mathcal{E}^*} = \sigma_0^2 \boldsymbol{I}_{nT}$, $\|\boldsymbol{y}_0\|_2 = \mathcal{O}_p(\sqrt{n})$, *and there exist* *nonstochastic matrices* $\bar{\mathbb{D}}_+$ *and* $\bar{\mathbb{D}}_{+,\mathrm{LS}}$*, such that* $\bar{\boldsymbol{D}}_+ \xrightarrow{p} \bar{\mathbb{D}}_+$ *and* $\bar{\boldsymbol{D}}_{+,\mathrm{LS}} \xrightarrow{p} \bar{\mathbb{D}}_{+,\mathrm{LS}}$ *as* $n, T \to \infty$ *with* $T/n \to \gamma \in (0, \infty)$*, and the eigenvalues of* $\bar{\mathbb{D}}_+$ *and* $\bar{\mathbb{D}}_{+,\mathrm{LS}}$ *are bounded away from zero and from above by a constant. Moreover, assume*

$$\begin{aligned}\operatorname{avar}(\sqrt{nT}(\hat{\boldsymbol{\theta}}_{\mathrm{TLS}} - \boldsymbol{\theta}_0)) &= \sigma_0^2 \bar{\mathbb{D}}_+^{-1} \\ \operatorname{avar}(\sqrt{nT}(\hat{\boldsymbol{\theta}}_{\mathrm{LS}} - \boldsymbol{\theta}_0)) &= \sigma_0^2 \bar{\mathbb{D}}_{+,\mathrm{LS}}^{-1},\end{aligned}$$

where $\operatorname{avar}(\cdot)$ *denotes asymptotic variance. Then*

$$\operatorname{avar}(\sqrt{nT}(\hat{\boldsymbol{\theta}}_{\mathrm{TLS}} - \boldsymbol{\theta}_0)) \succeq \operatorname{avar}(\sqrt{nT}(\hat{\boldsymbol{\theta}}_{\mathrm{LS}} - \boldsymbol{\theta}_0)).$$

In this case one can decompose

$$\begin{aligned}\bar{\boldsymbol{D}}_{+,\mathrm{LS}} - \bar{\boldsymbol{D}}_+ = \frac{1}{nT} \bar{\boldsymbol{\mathcal{H}}}^\top (\boldsymbol{M}_{\boldsymbol{F}_0} \otimes (\boldsymbol{P}_{\boldsymbol{\Lambda}_0} - \boldsymbol{P}_{\boldsymbol{P}_{\mathcal{X}} \boldsymbol{\Lambda}_0})) \bar{\boldsymbol{\mathcal{H}}} \\ + \frac{\sigma_0^2}{nT} \operatorname{tr}((\boldsymbol{G}\boldsymbol{G}^\top \otimes \boldsymbol{M}_{\mathcal{X}})) \boldsymbol{\pi}_{K+1} \boldsymbol{\pi}_{K+1}^\top. \end{aligned} \tag{5.5}$$

The first term in the above is analogous to (3.5) and reflects the information lost in transforming the factor loadings. The second term in (5.5) reflects information in the covariates that is lost by transforming the model through $\boldsymbol{Q}_{\mathcal{X}}$. For the covariates $\boldsymbol{X}_k$ which are strictly exogenous, there is no loss of information. For lagged outcomes on the other hand, information may be lost. Indeed, one can establish that

$$\frac{\sigma_0^2}{nT} \operatorname{tr}((\boldsymbol{G}\boldsymbol{G}^\top \otimes \boldsymbol{M}_{\mathcal{X}})) \boldsymbol{\pi}_{K+1} \boldsymbol{\pi}_{K+1}^\top \succeq 0.$$

When the model includes lagged outcomes it becomes more difficult to compare the efficiency of the RS, ALS and TLS estimators in the manner of Section 4. In this event the TLS estimator uses only a subset of the available moment conditions. One would, therefore, expect there to be cases where the TLS estimator is less efficient than the RS and ALS estimators.

6 Simulations

6.1 Short Panel Exercise

This first exercise compares the performance of the TLS estimator with alternatives in the context of a short panel. Outcomes are generated according to

$$\boldsymbol{Y} = \beta_{0,1}\boldsymbol{X}_1 + \beta_{0,2}\boldsymbol{X}_2 + \boldsymbol{\Lambda}_0\boldsymbol{F}_0^\top + \boldsymbol{\varepsilon},$$

with $\beta_{0,1} = 1$ and $\beta_{0,2} = -1$. The covariate $\boldsymbol{X}_1$ is generated with elements drawn from the standard normal distribution. The covariate $\boldsymbol{X}_2 = \boldsymbol{\Lambda}_0\boldsymbol{F}_0^\top + \boldsymbol{\epsilon}$, where $R_0 = 2$, and $\lambda_{0,ir}$, $f_{0,tr}$, and ϵ_{it} are drawn independently from the standard normal distribution. Note that the factors and loadings which enter into $\boldsymbol{X}_2$ are the same as those that appear in the outcome equation. For the error in the outcome equation, first a variable u_{it} is generated as $u_{it} \coloneqq u_{it}^* \times \|\boldsymbol{f}_{0,t}\|_2$ where u_{it}^* are independent over i and t and normally distributed with a mean of zero and variance drawn uniformly from the interval $[0.5, 1.5]$. Thereafter, the errors are generated according to $\varepsilon_{it} = \phi\varepsilon_{i,t-1} + u_{it}$ with $\phi = 0.5$ and $\varepsilon_{i0} = 0$. In this way the errors exhibit both conditional and unconditional heteroskedasticity, as well as serial correlation.

Table 1 below displays the empirical bias and empirical standard error of the LS estimator, the TLS estimator, and the one-step ALS estimator described in Section 4.2. Throughout $R_e = R_0$ and the number of draws is $10,000$.[29]

[29] Specifically, the empirical bias is computed by averaging $\sqrt{nT}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$ over Monte Carlo draws. The empirical standard error is computed by taking the standard deviation of $\sqrt{nT}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$ over Monte Carlo draws.

Table 1: Empirical Bias (Empirical Standard Error)

		β_1			β_2		
	$n \setminus T$	6	9	12	6	9	12
LS	100	**0.004**	**-0.002**	**0.017**	**6.417**	**6.185**	**4.395**
		(1.393)	(1.377)	(1.356)	(5.405)	(6.047)	(5.563)
	250	**0.019**	**0.004**	**-0.003**	**9.789**	**8.962**	**5.382**
		(1.358)	(1.334)	(1.338)	(8.091)	(8.757)	(6.915)
	500	**-0.013**	**-0.019**	**-0.023**	**13.782**	**12.008**	**6.614**
		(1.349)	(1.347)	(1.329)	(11.292)	(11.849)	(8.171)
TLS	100	**0.010**	**-0.004**	**0.021**	**0.349**	**0.232**	**0.214**
		(1.474)	(1.396)	(1.369)	(1.933)	(1.545)	(1.450)
	250	**0.014**	**-0.004**	**0.009**	**0.088**	**0.053**	**0.034**
		(1.454)	(1.373)	(1.364)	(1.522)	(1.392)	(1.370)
	500	**-0.011**	**-0.019**	**-0.021**	**0.030**	**0.028**	**0.013**
		(1.455)	(1.385)	(1.353)	(1.477)	(1.395)	(1.355)
ALS	100	**0.031**	**0.003**	**0.019**	**0.356**	**0.173**	**0.191**
		(2.164)	(2.531)	(2.928)	(2.637)	(2.709)	(3.117)
	250	**0.016**	**0.019**	**-0.012**	**0.093**	**0.032**	**0.028**
		(2.126)	(2.548)	(2.870)	(2.224)	(2.528)	(2.868)
	500	**-0.012**	**-0.008**	**-0.014**	**0.025**	**0.041**	**-0.009**
		(2.140)	(2.516)	(2.839)	(2.140)	(2.546)	(2.884)

Since the first regressor is uncorrelated with the factors and the loadings, all three of the estimators for β_1 display small bias. On the other hand, for β_2 the LS estimator exhibits substantially more bias than the TLS and ALS estimators, with this becoming more acute with T fixed and n increasing. This is unsurprising, since the LS estimator is known to suffer from asymptotic bias of order n/T. When comparing the TLS and ALS estimators, their bias is small and overall quite similar. Turning to the standard error, for β_1 the LS estimator has a smaller standard error than the TLS estimator, though their values become more similar for larger values of n and T. This might be expected given Proposition 3. For β_2 they are markedly dissimilar, and indeed the standard error of the LS estimator becomes increasingly large with T fixed and n increasing. Comparing the TLS estimator with the ALS estimator, the latter consistently exhibits a larger standard error. This supports the findings of Proposition 7. Table 2 below presents empirical coverage probabilities of a 95% two-sided confidence interval. For the TLS estimator these are computed using the fixed-T covariance matrix estimator described in Section 5.1.1. For the LS estimator these are computed using the covariance matrix estimator described in Moon and Weidner (2017) with the bandwidth parameter set equal to $\lfloor \log(T) \rfloor$.

Table 2: Empirical Coverage Probability of a 95% Confidence Interval

		β_1			β_2		
	$n \setminus T$	6	9	12	6	9	12
LS	100	0.887	0.910	0.924	0.221	0.303	0.456
	250	0.901	0.928	0.928	0.122	0.187	0.347
	500	0.903	0.924	0.928	0.067	0.101	0.221
TLS	100	0.942	0.941	0.944	0.892	0.923	0.928
	250	0.946	0.950	0.946	0.940	0.948	0.947
	500	0.948	0.949	0.950	0.945	0.947	0.952
ALS	100	0.940	0.947	0.944	0.910	0.937	0.941
	250	0.948	0.945	0.948	0.943	0.947	0.947
	500	0.947	0.949	0.951	0.946	0.945	0.950

As one would expect, the coverage probabilities of the LS estimator are poor when T is small, and only start to improve when T increases. Indeed, for β_2 the coverage probabilities quickly decline with T fixed and n increasing. Both the ALS and TLS estimators exhibit good coverage, with these quickly approaching their nominal value with T fixed and n increasing.

Table 3 below shows the percentage of times the eigenvalue ratio test described in Section 5.2 correctly detects the true number of factors in the model. This entails first estimating the model with the number of factors overestimated ($R_e = 5$ is used in simulations), after which a pure factor model can be constructed, to which the test can then be applied. The parameter ϱ_n is set equal to $T^{\frac{1}{4}}/\sqrt{n}$.[30]

Table 3: Percentage of Estimated R equal to R_0

$n \setminus T$	9	12	15
100	75.84	83.59	87.33
250	89.51	95.63	97.37
500	95.03	98.26	99.38

[30]The values of T are slightly increased to ensure the condition $T \geq T_{\min}$ is satisfied.

Overall the test performs well, with the error rate decreasing as either n or T increases.

6.2 Large Panel Exercise

This second exercise compares the LS and TLS estimators, as well as their bias-corrected counterparts, in a setting where both n and T are large. Outcomes are generated according to

$$\boldsymbol{Y} = \alpha_0 \boldsymbol{Y}_{-1} + \beta_{0,1}\boldsymbol{X}_1 + \beta_{0,2}\boldsymbol{X}_2 + \boldsymbol{\Lambda}_0 \boldsymbol{F}_0^\top + \boldsymbol{\varepsilon},$$

with $\alpha_0 = 0.5$, $\beta_{0,1} = 1$ and $\beta_{0,2} = -1$. The regressors, the factors, the loadings, and the covariates are generated in the same manner as in the previous design. The error is also generated as previously, but with the autoregressive parameter $\phi = 0$. Table 4 below presents empirical bias and empirical standard error for the LS estimator (LS), the bias-corrected LS estimator (LS-BC), the TLS estimator (TLS), and the bias-corrected TLS estimator (TLS-BC). The bandwidth parameter for both the LS and TLS estimators is set equal to 1.

Table 4: Empirical Bias (Empirical Standard Error)

		α			β_1			β_2		
	$n \setminus T$	10	25	50	10	25	50	10	25	50
LS	100	**-0.031**	**-0.006**	**0.003**	**-0.026**	**-0.012**	**-0.007**	**0.125**	**0.050**	**0.021**
		(0.730)	(0.559)	(0.529)	(1.143)	(1.048)	(1.027)	(1.152)	(1.063)	(1.046)
	250	**-0.054**	**-0.008**	**-0.002**	**-0.018**	**0.000**	**0.001**	**0.106**	**0.039**	**0.018**
		(0.909)	(0.583)	(0.534)	(1.124)	(1.050)	(1.025)	(1.117)	(1.064)	(1.026)
	500	**-0.057**	**-0.012**	**-0.012**	**-0.063**	**0.002**	**0.006**	**0.111**	**0.030**	**0.023**
		(1.150)	(0.647)	(0.544)	(1.138)	(1.032)	(1.014)	(1.132)	(1.055)	(1.021)
LS-BC	100	**-0.017**	**-0.003**	**0.004**	**-0.008**	**-0.010**	**-0.007**	**0.107**	**0.048**	**0.021**
		(0.594)	(0.534)	(0.524)	(1.142)	(1.047)	(1.027)	(1.151)	(1.063)	(1.046)
	250	**-0.020**	**-0.005**	**-0.001**	**0.009**	**0.003**	**0.002**	**0.079**	**0.036**	**0.017**
		(0.611)	(0.527)	(0.518)	(1.123)	(1.050)	(1.025)	(1.116)	(1.064)	(1.026)
	500	**-0.023**	**-0.007**	**-0.009**	**-0.025**	**0.006**	**0.007**	**0.073**	**0.026**	**0.022**
		(0.648)	(0.533)	(0.512)	(1.135)	(1.032)	(1.014)	(1.129)	(1.055)	(1.021)
TLS	100	**-0.021**	**-0.002**	**0.003**	**-0.011**	**-0.011**	**-0.007**	**0.033**	**0.026**	**0.021**
		(0.746)	(0.603)	(0.529)	(1.143)	(1.047)	(1.027)	(1.142)	(1.061)	(1.046)
	250	**-0.005**	**-0.006**	**-0.002**	**0.015**	**0.002**	**0.001**	**0.019**	**0.013**	**0.006**
		(0.770)	(0.638)	(0.589)	(1.124)	(1.050)	(1.025)	(1.110)	(1.063)	(1.026)
	500	**-0.010**	**-0.013**	**-0.012**	**-0.015**	**0.006**	**0.007**	**0.023**	**0.007**	**0.012**
		(0.772)	(0.660)	(0.608)	(1.136)	(1.032)	(1.014)	(1.125)	(1.054)	(1.021)
TLS-BC	100	**-0.015**	**0.000**	**0.004**	**-0.006**	**-0.010**	**-0.007**	**0.027**	**0.025**	**0.021**
		(0.739)	(0.595)	(0.524)	(1.143)	(1.047)	(1.027)	(1.142)	(1.061)	(1.046)
	250	**0.001**	**-0.005**	**-0.001**	**0.019**	**0.003**	**0.002**	**0.016**	**0.013**	**0.006**
		(0.767)	(0.635)	(0.585)	(1.124)	(1.050)	(1.025)	(1.110)	(1.063)	(1.026)
	500	**-0.007**	**-0.012**	**-0.011**	**-0.012**	**0.007**	**0.007**	**0.020**	**0.007**	**0.012**
		(0.770)	(0.659)	(0.606)	(1.136)	(1.032)	(1.014)	(1.125)	(1.054)	(1.021)

Overall the bias for the TLS estimator is smaller than that of the LS estimator, both with and without bias correction. This is particularly notable for β_2 where, owing to the correlation between $\boldsymbol{X}_2$, the factors, and the loadings, the LS estimator exhibits sizeable bias, particularly at smaller values of T. As in the previous design, it is also found that as sample size increases the standard error of the LS and TLS estimators generally becomes similar. Interestingly, for the autoregressive parameter α, the standard error of both estimators is observed to increase with T fixed and n increasing, and decrease for n fixed and T increasing. For the TLS estimator, this is explained by the presence of variance components associated with the lagged outcomes which are of order T/n.[31] When both n and T are large, these reduce the standard error, but when T is fixed

[31] These are $\boldsymbol{\Upsilon}^{(2)}$ and $\boldsymbol{\Upsilon}^{(4)}$; see Appendix E.

and n is large, these additional components diminish resulting in an increased standard error.[32]

Table 5 below presents empirical coverage probabilities of a 95% two-sided confidence interval. Two variants are displayed for the TLS estimator: TLS (Long) which is computed using the large-T covariance matrix estimator described in Section 5.1.2, and TLS (Short) which uses the fixed-T covariance estimator described in Section 5.1.1. In both cases bias correction is performed using the results in Section 5.1.2.

Table 5: Empirical Coverage Probability of a 95% Confidence Interval

		α			β_1			β_2		
	$n \setminus T$	10	25	50	10	25	50	10	25	50
LS	100	0.842	0.920	0.937	0.921	0.941	0.944	0.914	0.933	0.937
	250	0.744	0.913	0.934	0.924	0.940	0.946	0.926	0.935	0.945
	500	0.630	0.873	0.929	0.920	0.943	0.945	0.922	0.937	0.947
LS-BC	100	0.911	0.932	0.940	0.921	0.941	0.943	0.915	0.933	0.938
	250	0.904	0.937	0.942	0.924	0.940	0.946	0.924	0.935	0.945
	500	0.885	0.938	0.945	0.921	0.942	0.946	0.923	0.938	0.947
TLS (Long)	100	0.916	0.933	0.937	0.923	0.941	0.944	0.921	0.934	0.937
	250	0.919	0.938	0.941	0.924	0.939	0.946	0.928	0.935	0.945
	500	0.926	0.938	0.943	0.921	0.943	0.946	0.926	0.938	0.947
TLS-BC (Long)	100	0.919	0.937	0.940	0.923	0.940	0.943	0.921	0.934	0.938
	250	0.921	0.940	0.943	0.924	0.939	0.946	0.928	0.935	0.945
	500	0.927	0.939	0.944	0.921	0.943	0.946	0.926	0.938	0.947
TLS (Short)	100	0.939	0.937	0.937	0.944	0.945	0.943	0.942	0.942	0.940
	250	0.944	0.946	0.944	0.948	0.947	0.949	0.949	0.943	0.948
	500	0.950	0.946	0.946	0.946	0.952	0.950	0.949	0.946	0.951
TLS-BC (Short)	100	0.939	0.940	0.942	0.944	0.945	0.943	0.942	0.942	0.940
	250	0.943	0.947	0.947	0.948	0.947	0.949	0.949	0.943	0.948
	500	0.951	0.948	0.947	0.946	0.952	0.950	0.949	0.946	0.951

For the autoregressive parameter α, the coverage of the LS estimator declines with T fixed and n increasing. This is expected due to bias of order n/T from which the LS estimator is known to suffer. The bias-corrected LS estimator performs better, with coverage being much closer to its nominal value. However, this also performs poorly at

[32] In unreported additional simulation output which includes $n = 1000$ and $n = 2500$, the standard error of the TLS estimator is seen to stabilise, while it continues to increase for the LS estimator.

smaller values of T, in which case coverage is also observed to decline with T fixed and n increasing. For smaller values of T the TLS (Long) estimator performs better than the LS estimator. This is particularly noticeable for the autoregressive coefficient where the coverage probabilities of the TLS estimator improve with T fixed and n increasing. Overall the LS and LS-BC estimators require larger values of T in order to attain nominal coverage. When both n and T are large, the coverage probabilities for the LS and TLS (Long) estimators are broadly similar, with and without bias-correction. Surprisingly, the TLS (Short) estimator outperforms both the LS and TLS (Long) estimators, when n is large and T is small, and also where both n and T are large.

Finally, Table 6 below shows the percentage of times the eigenvalue ratio test correctly detects the true number of factors. As previously, the parameter ϱ_n is set equal to $T^{\frac{1}{4}}/\sqrt{n}$.

Table 6: Percentage of Estimated R equal to R_0

$n \setminus T$	10	25	50
100	84.59	98.97	99.94
300	94.50	99.96	100.00
500	97.81	99.99	100.00

Similar to the results for the static model, the test performs well with the error rate declining with either n or T increasing. The overall error rate in this second design is smaller due to the larger values of T.

7 Conclusion

This paper has introduced a method to estimate linear panel data models with interactive fixed effects which has been shown to be consistent and asymptotically unbiased when n is large and T fixed, and also when both n and T are large, provided $T/n \to 0$. This stands in contrast to the usual case where the LS estimator is generally inconsistent when n is large and T is fixed, and suffers from asymptotic bias when both n and T are large. Careful study of this estimation approach has also revealed interesting connections between the LS estimator and several method of moments-based approaches, bridging the gap between what are, at present, two quite separate literatures.

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.

Alvarez, J., Arellano, M., 2003. The time series and cross-section asymptotics of dynamic panel data estimators. Econometrica 71 (4), 1121–1159.

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.

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

Breitung, J., Hansen, P., 2021. Alternative estimation approaches for the factor augmented panel data model with small T. Empirical Economics 60, 327–351.

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

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

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

Freyberger, J., 2018. Non-parametric panel data models with interactive fixed effects. The Review of Economic Studies 85 (3), 1824–1851.

Hall, P., Heyde, C. C., 1980. Martingale Limit Theory and Its Application. Probability and Mathematical Statistics. Academic Press, New York.

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

Higgins, A., 2023. Instrumental variables for dynamic spatial models with interactive effects. Working Paper.

Higgins, A., Martellosio, F., 2023. Shrinkage estimation of network spillovers with factor structured errors. Journal of Econometrics 233 (1), 66–87.

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

Hsiao, C., 2018. Panel models with interactive effects. Journal of Econometrics 206 (2), 645–673.

Hsiao, C., Shi, Z., and, Q. Z., 2022. Transformed estimation for panel interactive effects models. Journal of Business & Economic Statistics 40 (4), 1831–1848.

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.

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

Lin, Z., 2010. Probability inequalities. Springer.

Lu, X., Su, L., 2016. Shrinkage estimation of dynamic panel data models with interactive fixed effects. Journal of Econometrics 190 (1), 148–175.

Mackey, L., Jordan, M. I., Chen, R. Y., Farrell, B., Tropp, J. A., 2014. Matrix concentration inequalities via the method of exchangeable pairs. The Annals of Probability 42 (3), 906 – 945.

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.

Vershynin, R., 2012. Introduction to the non-asymptotic analysis of random matrices. Cambridge University Press, p. 210–268.

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.

White, H., 1984. Asymptotic Theory for Econometricians, 1st Edition. Economic Theory, Econometrics, and Mathematical Economics. Academic Press, San Diego.