Instrumental Variables for Dynamic Spatial Models with Interactive Effects*

Ayden Higgins†
University of Oxford

Preliminary and Incomplete
November 24, 2023

Latest Version Available at:
https://ahiggins.co.uk/IVIFE.pdf

Abstract

This paper studies estimation and inference in the context of a dynamic spatial model with interactive effects. An instrumental variables interactive fixed effects (IV-IFE) estimator is proposed which provides consistent and asymptotically unbiased estimates of model parameters, as long as the cross-sectional dimension of the data grows sufficiently fast relative to number of time periods and the number of instruments. This trivially includes where both the number of time periods and the number of instruments are fixed. Nonetheless, circumstances exist where the estimator can exhibit asymptotic bias, the extent to which depends, in part, on the structure of the spatial dependence. A bias corrected estimator is constructed, which, through simulation, is demonstrated to be an effective remedy to this issue. An empirical application utilises the method to study the relationship between economic growth, civil liberties and political rights in the 21st century.

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

1 Introduction

1.1 Overview

Many, if not most panel data sets exhibit some degree of dependence, whether that be across time, between cross-sectional units, or both. When researchers set out to

*This work was supported by the European Research Council through the grant ERC-2018-CoG-819086-PANEDA.

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

quantify the strength of these dependencies, linear autoregressive models are frequently used. Within this family of models, the time series autoregressive model, known simply as the AR model, is the prototype, with its roots dating back to the work of Yule, Walker, and Wold (amongst others), in the early part of the 20th century (Klein, 1997). The analogue of this model in the context of spatial data is the spatial autoregressive model (SAR), first appearing in Whittle (1954), and set out fully in Cliff and Ord (1973, 1981).[1] These models provide a flexible way to model dependent data, and are useful not just for measuring the extent to which dependence may exist, but also for modelling more complicated channels through which changes in one variable may impact another. This usefulness is only magnified in the context of panel data where linear autoregressive models provide a parsimonious way to model complicated processes that evolve both over time and across space.

Correlation in outcomes may, however, arise for a myriad of reasons, and economists may wish to distinguish between different sources. In particular, a distinction is typically made between correlation that arises due to state dependence, and that driven by unobserved factors. The former, more typically discussed in context of AR models, describes the causal impact which outcomes in the past, or in neighbouring locations, have on present outcomes (Heckman, 1981). The latter describes correlation in outcomes arising for a broader set of reasons, such as unobserved heterogeneity, or from exposure to common shocks or a common environment. When seeking to confirm or to refute the presence of state dependence, it is common practice to also control for unobserved factors, and thereby raise the threshold before which correlation in outcomes is taken as evidence in favour of true state dependence. In the context of panel models, unobserved factors have traditionally been controlled for by modelling additive effects, however, modelling interactive effects has become increasingly popular, as this provides a more general way to control for these unobserved features of the data.

This paper develops an instrumental variables (IV) estimator for dynamic spatial models with interactive effects, which possesses several benefits when compared to alternative estimation approaches which might be considered in this context. Foremost amongst these is simplicity, as estimation reduces to a univariate optimisation problem which depends only on the parameter of interest in the model, and therefore does not (directly) depend on any number of nuisance parameters which may arise as a consequence of modelling a factor structure in the error, or due to heteroskedasticity.[2] This is unlike existing estimators, and is useful not just from the point of view of practicality, but also in terms of establishing the asymptotic properties of the estimator. The

[1] Although the discussion in this paper tacitly assumes that the notion of space is tied to a physical location, it should be emphasised that the model (and estimation approach) equally apply to settings where the notion of space is interpreted more abstractly. Examples of this include models of social interaction between individuals, relations between firms, and dependencies between financial assets.

[2] Though sometimes innocuous, the possible heteroskedasticity of regression errors can pose a major obstacle to the estimation of spatial models, particularly within a likelihood framework; see, e.g., Lin and Lee (2010).

paper focuses primarily on short panels, and assumes that a large set of instruments is available, motivated by the observation that the reduced form of the model implies a rich set of internally generated instruments that can be used in estimation. Allowing for lagged outcomes (in both space and time) to appear as regressors, interactive effects, and a large number of instruments is challenging in a short panel, and to the best of the author's knowledge there is currently no alternative method which addresses all three issues. The main result of this paper establishes that, provided the cross-sectional dimension of the panel (n) grows sufficiently fast relative to the number of time periods (T), and the number of instruments (m), the estimator is consistent, unbiased and asymptotically normal. This trivially includes where the number of time periods and the number of instruments is fixed. However, when n grows more slowly, the estimator can exhibit a sizeable asymptotic bias which depends, in part, on the structure of spatial dependence in outcomes. To address this issue, a bias-corrected estimator is formulated and demonstrated through simulation to be a reliable method for conducting inference. An empirical application utilises the method to examine the relationship between economic growth, civil liberties, and political rights in the 21st century, with the findings being that increased political rights and civil liberties within a country results in higher values of GDP per capita in the long run.

2 Related Literature

Under an asymptotic where the dimension of the cross-section and the time series are both large, quasi-maximum likelihood (QML) estimators are available to estimate dynamic spatial models with interactive effects, such as those proposed by Shi and Lee (2017) and Bai and Li (2021). These estimators sometimes possess a very useful property that the implied estimates of the factors and the loadings are given by solutions to standard principal component problems. This makes their estimation, whether implicitly or explicitly, significantly more straightforward (see, e.g., Bai (2009); Moon and Weidner (2015)). In Shi and Lee (2017), the authors introduce a simple estimator which avoids directly estimating the factors and loadings by concentrating these out using principal components. Bai and Li (2021) relax the homoskedasticity assumption imposed in Shi and Lee (2017) and allow for cross-sectional (though not temporal) heteroskedasticity through modelling additional variance parameters. This, however, comes at the cost of complexity since it introduces another larger set of nuisance parameters, in addition to the factors and the loadings. Both estimators are consistent as $n, T \to \infty$, though they suffer from asymptotic bias. This is typical of QML estimators in the presence of high-dimensional nuisance parameters, in which case the incidental parameter problem usually manifests as an asymptotic bias when both dimensions of the panel are asymptotically proportional. However, when $n \to \infty$ but T is fixed, these estimators are, in general, inconsistent, in which case it becomes necessary to resort to alternative, and often more bespoke estimation strategies.

Li and Yang (2023) develop a modified QML estimator that provides consistent and asymptotically unbiased estimates of common parameters with T fixed and $n \to \infty$. Their approach is based on the idea of recentering the profile score, which has been successfully applied in other contexts as a resolution to the inconsistency of maximum likelihood estimators in the presence of high-dimensional nuisance parameters (see, e.g. McCullagh and Tibshirani (1990)). They approach the interactive effects in a similar manner to Ahn et al. (2013) which, in essence, involves shifting the incidental parameter problem from both dimensions of the panel onto only the time dimension. This, however, does not free the estimation problem entirely of nuisance parameters, and still necessitates solving a set of nonlinear adjusted score equations over the common parameters and a set of T-dimensional nuisance parameters. Moving away from a likelihood framework, Kuersteiner and Prucha (2020) study a generalised method of moments (GMM) estimator which affords a great deal of generality, not only relaxing strict exogeneity of the regressors (aside from lagged outcomes), which is assumed by all the aforementioned methods, but also allowing for a possibly endogenous weights matrix. They introduce an innovative generalised Helmert transformation which, similar to the Ahn et al. (2013) approach, shifts the incidental parameter problem onto only the time dimension. However, this transformation is engineered so as not to produce correlation in the errors and to maintain sequential exogeneity of regressors and weights matrices which allows for linear and quadratic moments to be constructed using the transformed model. Nonetheless, this also does not produce an estimation problem which is entirely free of nuisance parameters.

Recently Cui et al. (2022) have studied an IV estimator for a dynamic spatial model with interactive effects which combines the common correlated effects approach to modelling interactive effects (Pesaran, 2006) with the use of principal components. Their two-step approach is designed for large panels and yields a consistent and asymptotically unbiased estimator as both $n, T \to \infty$. However, their approach is unsuitable for short panels and they do not consider the possibility that the number of instruments may also be increasing with sample size. This paper also relates to a much wider set of literatures concerning spatial models (Kelejian and Prucha, 1998; Lee, 2002, 2003, 2004), spatial panel models with additive fixed effect (Lee and Yu, 2010a,b, 2014), panel models with interactive effects (Pesaran, 2006; Bai, 2009; Moon and Weidner, 2015), and instrumental variables with interactive effects (Ahn et al., 2001, 2013; Robertson and Sarafidis, 2015; Lee et al., 2012).

Notation: Throughout the paper, all vectors and matrices are real unless stated otherwise. 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 as $\mu_{\min}(\boldsymbol{A}) := \mu_n(\boldsymbol{A}) \leq \ldots \leq \mu_1(\boldsymbol{A}) =: \mu_{\max}(\boldsymbol{A})$. Let $\boldsymbol{P}_{\boldsymbol{A}} := \boldsymbol{A}(\boldsymbol{A}^\top \boldsymbol{A})^+ \boldsymbol{A}^\top$ and $\boldsymbol{M}_{\boldsymbol{A}} := \boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{A}}$, where $\boldsymbol{I}_n$ is the $n \times n$ identity matrix and $+$ denotes the Moore-Penrose generalised inverse. An $n \times 1$ vector of all ones is denoted $\boldsymbol{\iota}_n$, an $n \times 1$ vector of all zeros is denoted $\mathbf{0}_n$, and an $n \times m$ matrix of zeros is denoted $\mathbf{0}_{n\times m}$. For a matrix $\boldsymbol{A}$ which potentially has an increasing dimension, $\boldsymbol{\mathcal{O}}_p(1)$ is used to indicate that $\|\boldsymbol{A}\|_2 = \mathcal{O}_p(1)$ and, similarly, $\boldsymbol{o}_p(1)$ signifies that $\|\boldsymbol{A}\|_2 = o_p(1)$. Throughout, c is used to denote some arbitrary positive constant. The operation $\mathrm{vec}(\cdot)$ applied to an $n \times m$ matrix $\boldsymbol{A}$ creates an $nm \times 1$ vector $\mathrm{vec}(\boldsymbol{A})$ by stacking the columns of $\boldsymbol{A}$. The operation $\mathrm{diag}(\boldsymbol{B})$ applied to an $n \times n$ matrix $\boldsymbol{B}$ creates an $n \times n$ diagonal matrix $\mathrm{diag}(\boldsymbol{B})$ which contains the diagonal elements of $\boldsymbol{B}$ along its diagonal and $\mathrm{off}(\boldsymbol{B}) \coloneqq \boldsymbol{B} - \mathrm{diag}(\boldsymbol{B})$. A sequence of $n \times n$ matrices $\boldsymbol{C}_n$ is said to be uniformly bounded in absolute row and column sums (UB) if both the sequences $\|\boldsymbol{C}_n\|_1$ and $\|\boldsymbol{C}_n\|_\infty$ are bounded.

3 Model and Estimation

3.1 Dynamic Spatial Model

The model studied in this paper assumes that amongst a cross-section of n units, indexed $i = 1, \ldots, n$, over a number of time periods $t = 1, \ldots, T$, outcomes are generated according to

$$\boldsymbol{y}_t = \rho \boldsymbol{W} \boldsymbol{y}_t + \alpha \boldsymbol{y}_{t-1} + \phi \boldsymbol{W} \boldsymbol{y}_{t-1} + \boldsymbol{X}_t \boldsymbol{\beta} + \boldsymbol{\eta}_t, \tag{3.1}$$

where $\boldsymbol{y}_t$, and $\boldsymbol{\eta}_t$ are $n \times 1$ vectors of outcomes and error terms, respectively, $\boldsymbol{X}_t$ is an $n \times K$ matrix of covariates, $\boldsymbol{\beta}$ is a $K \times 1$ parameter vector, ρ, α and ϕ are scalar parameters, and $\boldsymbol{W}$ is an $n \times n$ weights matrix. It should be highlighted that throughout this paper the term 'covariates' will be applied exclusively to the variable $\boldsymbol{X}_t$. The term regressors will be used to describe $\boldsymbol{W}\boldsymbol{y}_t$, $\boldsymbol{y}_{t-1}$, $\boldsymbol{W}\boldsymbol{y}_{t-1}$ and $\boldsymbol{X}_t$ collectively. The parameters ρ and α quantify the strength of purely spatial, or purely temporal dependence in the data, while ϕ provides a measure of dependence across both time and space. The weights matrix $\boldsymbol{W}$, which is assumed to be observed, summarises the structure of spatial dependence, with larger elements indicating a closer proximity.[3] It is assumed that $\boldsymbol{W}$ has a zero diagonal, which, if coupled with the assumption that $\boldsymbol{W}$ is row-normalised, gives $\boldsymbol{W}\boldsymbol{y}_t$ and $\boldsymbol{W}\boldsymbol{y}_{t-1}$ the interpretation of being weighted averages of outcomes omitting location i. Although dependent only on a few common parameters, the dynamic spatial model belies a significant amount of heterogeneity and, in particular, it gives rise to marginal effects that are both location and time dependent. For ease of exposition, outcomes are assumed to evolve according to a first order process, that is, the process only involves first order autoregressive terms: the extension to higher order processes is immediate.

[3]Since the weights matrix $\boldsymbol{W}$ is observed, this could, in principle, be allowed to vary over time at the expense of more cumbersome notation.

3.2 Interactive Effects

In order to control for the possible presence of latent factors, it is assumed that $\boldsymbol{\eta}_t$ may be decomposed as

$$\boldsymbol{\eta}_t \coloneqq \boldsymbol{\Lambda} \boldsymbol{f}_t + \boldsymbol{\varepsilon}_t, \tag{3.2}$$

where $\boldsymbol{\Lambda}$ and $\boldsymbol{F} \coloneqq (\boldsymbol{f}_1, \ldots, \boldsymbol{f}_T)$ are, respectively, $n \times R$ and $R \times T$ matrices of loadings and factors, and $\boldsymbol{\varepsilon}_t$ is an $n \times 1$ vector of primitive error terms. This model provides a flexible means of controlling for unobserved features of the data, as common factors may vary across time and yet have a heterogeneous effect on the cross-section. It also nests as special cases several familiar models of additive effects such as individual, time or groups effects.

3.3 Instruments and Transformation

The parameter of interest in the model is $\boldsymbol{\theta} \coloneqq (\rho, \alpha, \phi, \ \boldsymbol{\beta}^\top)^\top$ which is a $P \times 1$ vector where $P \coloneqq K+3$. Estimation of this parameter is, however, complicated by two sources of endogeneity. The first stems from the factor term in the error, which, in this paper, is permitted to be arbitrarily correlated with regressors in the model. The second arises due to the presence of a spatial and temporal lags on the right-hand side of (3.1) which will, by construction, be correlated with both the error term $\boldsymbol{\varepsilon}$ and the factor term. This section addresses both of these issues by applying a transformation to the model. Before doing so, however, it is useful to first re-write the model in terms of matrices. Let $\boldsymbol{Y} \coloneqq (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_T)$, $\boldsymbol{Y}_{-1} \coloneqq (\boldsymbol{y}_0, \ldots, \boldsymbol{y}_{T-1})$, $\boldsymbol{X}_k$ be the $n \times T$ matrix containing observations on the k-th covariate with $k = 1, \ldots, K$, and $\boldsymbol{\varepsilon} \coloneqq (\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_T)$. Moreover, for brevity, define $\boldsymbol{\beta} \cdot \boldsymbol{X} \coloneqq \sum_{k=1}^{K} \beta_k \boldsymbol{X}_k$.

With this notation, the model can be rewritten more compactly in the form

$$\boldsymbol{S}(\rho)\boldsymbol{Y} = \alpha \boldsymbol{Y}_{-1} + \phi \boldsymbol{W} \boldsymbol{Y}_{-1} + \boldsymbol{\beta} \cdot \boldsymbol{X} + \boldsymbol{\Lambda} \boldsymbol{F}^\top + \boldsymbol{\varepsilon}, \tag{3.3}$$

where $\boldsymbol{S}(\rho) \coloneqq \boldsymbol{I}_n - \rho \boldsymbol{W}$. In order to allow for the possibility that components of the factor term may be correlated with regressors in the model, both the factors and loadings are treated as additional (nuisance) parameters and jointly estimated alongside the parameter of interest $\boldsymbol{\theta}$. An obvious estimation approach would be to minimise the sum of squared residuals, which gives rise to the least squares interactive fixed effects (LS-IFE) estimator, whose properties have been extensively studied; see, for example, Bai (2009) and Moon and Weidner (2015). Yet in order to consistently estimate both the n-dimensional factor loadings, and the T-dimensional factors, this approach typically requires both dimensions of the panel to diverge. As a consequence it is well established that the LS-IFE estimator of $\boldsymbol{\theta}$ is, in general, inconsistent with T fixed. Following Higgins (2022) it is, however, possible to first transform the model before applying the

LS-IFE estimator in order to resolve this fixed T inconsistency. The transformation described in that paper has three important properties: it reduces the dimension of the model, it preserves important features of the model, and it renders the error term asymptotically negligible. In the context of the dynamic spatial model (3.1), an equivalent transformation can be constructed by utilising instrumental variables.

Assuming the error $\boldsymbol{\varepsilon}$ is independent of the covariates $\boldsymbol{X}_k$, these variables may serve as their own instruments. For the spatially and temporally lagged outcomes, under some assumptions (set out in the next section) a rich set of instruments can be obtained by considering the reduced form of the model. Assume $\boldsymbol{S}(\rho)$ is invertible and can be expanded in the Neumann series

$$\boldsymbol{S}^{-1}(\rho) = \sum_{h=0}^{\infty} (\rho \boldsymbol{W})^h. \tag{3.4}$$

Since this implies

$$\boldsymbol{Y} = \sum_{h=0}^{\infty} (\rho \boldsymbol{W})^h (\alpha \boldsymbol{Y}_{-1} + \phi \boldsymbol{W} \boldsymbol{Y}_{-1} + \boldsymbol{\beta} \cdot \boldsymbol{X} + \boldsymbol{\Lambda} \boldsymbol{F}^\top + \boldsymbol{\varepsilon}), \tag{3.5}$$

then, assuming that at least one element in $\boldsymbol{\beta}$ is nonzero, it is possible to generate a multitude of instruments for $\boldsymbol{Y}$ using powers of the weights matrix interacted with the covariates $\boldsymbol{X}_k$: for example, with $\beta_1 \neq 0$, $\boldsymbol{X}_1, \boldsymbol{W}\boldsymbol{X}_1, \boldsymbol{W}^2\boldsymbol{X}_1, \boldsymbol{W}^3\boldsymbol{X}_1$ may serve as instruments.[4] Let $\boldsymbol{\mathcal{V}}$ be an $n \times m$ matrix containing a set of instruments. An $n \times m$ matrix $\boldsymbol{Q}_{\boldsymbol{\mathcal{V}}}$ can be constructed as $\boldsymbol{\mathcal{V}}(\boldsymbol{\mathcal{V}}^\top \boldsymbol{\mathcal{V}})^{-\frac{1}{2}}$.[5] Further define

$$\begin{aligned} \boldsymbol{Z}_1 &:= \boldsymbol{W}\boldsymbol{Y} & \boldsymbol{Z}_4 &:= \boldsymbol{X}_1 \\ \boldsymbol{Z}_2 &:= \boldsymbol{Y}_{-1} & &\vdots \\ \boldsymbol{Z}_3 &:= \boldsymbol{W}\boldsymbol{Y}_{-1} & \boldsymbol{Z}_P &:= \boldsymbol{X}_K. \end{aligned}$$

Then (3.3) can be premultiplied by $\boldsymbol{Q}_{\boldsymbol{\mathcal{V}}}^\top$ which gives

$$\tilde{\boldsymbol{Y}} = \boldsymbol{\theta} \cdot \tilde{\boldsymbol{Z}} + \tilde{\boldsymbol{\Lambda}} \boldsymbol{F}^\top + \tilde{\boldsymbol{\varepsilon}}, \tag{3.6}$$

where $\sim$ is used to denote transformed matrices, for example, $\tilde{\boldsymbol{Y}} := \boldsymbol{Q}_{\boldsymbol{\mathcal{V}}}^\top \boldsymbol{Y}$, and $\boldsymbol{\theta} \cdot \tilde{\boldsymbol{Z}}$ is defined analogously to $\boldsymbol{\beta} \cdot \boldsymbol{X}$. Notice that the transformation $\boldsymbol{Q}_{\boldsymbol{\mathcal{V}}}^\top$ fulfils the three roles described previously. First, with $n > m$, the system (3.6) is of dimension $m \times T$ and

[4] The model can be easily extended to include multiple weights matrices. In this event an even larger set of instruments can be generated since $\boldsymbol{S}^{-1}(\boldsymbol{\rho}) = \boldsymbol{I}_n + \sum_{s=1}^{S} \rho_{s_1} \boldsymbol{W}_{s_1} + \sum_{s_1=1}^{S} \sum_{s_2=1}^{S} \rho_{s_1}\rho_{s_2} \boldsymbol{W}_{s_1}\boldsymbol{W}_{s_2} + \sum_{s_1=1}^{S} \sum_{s_2=1}^{S} \sum_{s_3=1}^{S} \rho_{s_1}\rho_{s_2}\rho_{s_3} \boldsymbol{W}_{s_1}\boldsymbol{W}_{s_2}\boldsymbol{W}_{s_3} + \cdots$.

[5] Implicitly it has been assumed that the matrix $\boldsymbol{\mathcal{V}}$ has full column rank m. Suppose that this is not the case and that $\boldsymbol{\mathcal{V}}$ has instead rank $m_* < m$. Take a 'short' singular value decomposition $\boldsymbol{\mathcal{V}} := \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^\top$, where $\boldsymbol{U}$ is $n \times m_*$, $\boldsymbol{S}$ is $m_* \times m_*$, and $\boldsymbol{V}$ is $m_* \times m$. Define $\boldsymbol{\mathcal{V}}_* := \boldsymbol{U}\boldsymbol{S}$. One can then proceed in the same way, using $\boldsymbol{\mathcal{V}}_*$ and m_* in lieu of $\boldsymbol{\mathcal{V}}$ and m.

therefore now presents a lower dimensional estimation problem where, in particular, the transformed factor loadings $\tilde{\boldsymbol{\Lambda}}$ are $m \times R$. Second, this transformation preserves two important features of the model: the covariates, and the low-rank structure of the factor term. Intuitively $\boldsymbol{Q}_{\boldsymbol{\mathcal{V}}}^{\top}$ 'projects' onto the m-dimensional subspace spanned by the columns of $\boldsymbol{\mathcal{V}}$. If the columns of $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_K$ are included in $\boldsymbol{\mathcal{V}}$, the covariates are wholly preserved under the transformation. For the regressors $\boldsymbol{WY}$, $\boldsymbol{Y}_{-1}$ and $\boldsymbol{WY}_{-1}$, while these are not fully preserved, due to their correlation with the instruments, these will be partially preserved under $\boldsymbol{Q}_{\boldsymbol{\mathcal{V}}}^{\top}$. Finally, when the weights matrix $\boldsymbol{W}$ is nonstochastic and the covariates $\boldsymbol{X}_k$ are independent of $\boldsymbol{\varepsilon}$, as will be assumed throughout this paper, the transformation also serves to reduce the order of the error term.[6]

3.4 IV-IFE Estimator

Using a chosen set of instruments, estimation proceeds by minimising the following least squares objective function

$$\mathcal{Q}(\boldsymbol{\theta}, \tilde{\boldsymbol{\Lambda}}, \boldsymbol{F}) := \frac{1}{nT} \operatorname{tr}\left(\left(\tilde{\boldsymbol{Y}} - \boldsymbol{\theta} \cdot \tilde{\boldsymbol{Z}} - \tilde{\boldsymbol{\Lambda}} \boldsymbol{F}^{\top}\right)^{\top} \left(\tilde{\boldsymbol{Y}} - \boldsymbol{\theta} \cdot \tilde{\boldsymbol{Z}} - \tilde{\boldsymbol{\Lambda}} \boldsymbol{F}^{\top}\right)\right). \tag{3.7}$$

Both the factors and the transformed loadings can be concentrated out of (3.7), in which case one arrives at an objective function involving $\boldsymbol{\theta}$ alone,

$$\mathcal{Q}(\boldsymbol{\theta}) := \frac{1}{nT} \sum_{r=R+1}^{T} \mu_r \left(\left(\tilde{\boldsymbol{Y}} - \boldsymbol{\theta} \cdot \tilde{\boldsymbol{Z}}\right)^{\top} \left(\tilde{\boldsymbol{Y}} - \boldsymbol{\theta} \cdot \tilde{\boldsymbol{Z}}\right)\right), \tag{3.8}$$

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

$$\hat{\boldsymbol{\theta}} := \underset{\boldsymbol{\theta} \in \Theta}{\arg\min}\, \mathcal{Q}(\boldsymbol{\theta}). \tag{3.9}$$

4 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. Let $\boldsymbol{\Pi}$ denote a $T \times T$ shift matrix, which is all zeros, except those elements directly above the main diagonal which take a value of 1, $\boldsymbol{A}(\rho, \alpha, \phi) := \boldsymbol{S}^{-1}(\rho)(\alpha \boldsymbol{I}_n + \phi \boldsymbol{W})$, $\bar{\boldsymbol{\Pi}} := (\boldsymbol{\Pi}^{\top} \otimes \boldsymbol{I}_n)$, $\bar{\boldsymbol{W}} := (\boldsymbol{I}_T \otimes \boldsymbol{W})$, and $\bar{\boldsymbol{B}}(\rho, \alpha, \phi) := \boldsymbol{I}_{nT} - \rho \bar{\boldsymbol{W}} - \alpha \bar{\boldsymbol{\Pi}} - \phi \bar{\boldsymbol{W}} \bar{\boldsymbol{\Pi}}$. Moreover, let $|\boldsymbol{A}|$ denote the entrywise absolute value of a matrix, Θ denote the parameter space for $\boldsymbol{\theta}$, and $\Theta_\rho, \Theta_\alpha$ and Θ_ϕ denote the parameter spaces for ρ, α and ϕ, respectively. The following assumptions are made.

[6]See Higgins (2022) for additional details.

[7]In order to obtain the objective function (3.8), a number of factors R must be assumed. For the moment it is assumed that this number equals or exceeds the true number.

Assumption MD (Model)**.**

(i) The parameter vector $\boldsymbol{\theta}_0$ is in the interior of Θ, where Θ is a compact subset of $\mathbb{R}^P$.

(ii) The weights matrix $\boldsymbol{W}$ has a zero diagonal, is nonstochastic and UB.

(iii) For all $\rho \in \Theta_\rho$, $\alpha \in \Theta_\alpha$ and $\phi \in \Theta_\phi$, $|\det(\boldsymbol{S}(\rho))| \geq c > 0$, $|\det(\bar{\boldsymbol{B}}(\rho, \alpha, \phi))| \geq c > 0$, and $\|\boldsymbol{A}(\rho, \alpha, \phi)\|_2 < 1 - c$ holds for all (n, T), and $\boldsymbol{S}^{-1}(\rho)$, $\bar{\boldsymbol{B}}^{-1}(\rho, \alpha, \phi)$ and $\sum_{h=1}^{\infty} |\boldsymbol{A}^h(\rho, \alpha, \phi)|$ are UB.

(iv) $x_{kit}, \lambda_{0,ir}$ and $f_{0,tr}$ have uniformly bounded fourth moments.

(v) The errors ε_{it} are independent of the factors, the loadings, and the covariates, and are also independent over i and t with $\mathbb{E}[\varepsilon_{it}] = 0$, $\mathbb{E}[\varepsilon_{it}^2] =: \sigma_{it}^2 > 0$ and uniformly bounded fourth moments.

Assumption MD sets out a basic set of assumptions the model is assumed to satisfy. MD(iii) ensures that the matrix $\boldsymbol{S}(\rho)$ is invertible and can be expanded as a Neumann series for all values of ρ in the parameter space. A more primitive condition for this is that $\|\rho\boldsymbol{W}\| < 1$ for some norm $\|\cdot\|$. If, for example, the weights matrix is row-normalised, this holds when $|\rho| < 1$. Similarly, the condition $\|\rho\bar{\boldsymbol{W}} + \alpha\bar{\boldsymbol{\Pi}} + \phi\bar{\boldsymbol{W}}\bar{\boldsymbol{\Pi}}\| < 1$ for some norm is sufficient for the invertibility of $\bar{\boldsymbol{B}}(\rho, \alpha, \phi)$. Under row-nominalisation of the weights matrix, this holds if $|\rho| + |\alpha| + |\phi| < 1$. These assumptions also ensure dependence across both space and over time is sufficiently limited as to obtain by recursive substitution

$$\boldsymbol{y}_t = \sum_{h=0}^{\infty} \boldsymbol{A}^h(\rho_0, \alpha_0, \phi_0)\boldsymbol{S}^{-1}(\rho_0)(\boldsymbol{X}_{t-h}\boldsymbol{\beta}_0 + \boldsymbol{\Lambda}_0\boldsymbol{f}_{0,t-h} + \boldsymbol{\varepsilon}_{t-h}),$$

for all values of $\boldsymbol{\theta}$ in the parameter space. Assumption MD(v) assumes the errors are independent of the factors, the loadings and the covariates as in Bai (2009). This assumption also allows for (unconditional) heteroskedasticity in both dimensions of the panel.

Assumption CS (Consistency)**.**

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

(ii) $\min_{\boldsymbol{\delta}\in\mathbb{R}^P:\|\boldsymbol{\delta}\|_2=1} \sum_{r=R+R_0+1}^{T} \mu_r(\frac{1}{nT}(\boldsymbol{\delta} \cdot \tilde{\boldsymbol{Z}})^\top(\boldsymbol{\delta} \cdot \tilde{\boldsymbol{Z}})) \geq c > 0$, w.p.a.1 as $n \to \infty$.

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, is no less than R_0. Assumption CS(ii) is a multicollinearity condition and requires there to remain a sufficient amount of variation in the regressors after having been transformed by the matrix $\boldsymbol{Q}_{\boldsymbol{\mathcal{V}}}$ and then been projected orthogonal to arbitrary $R \times T$ factors and $R_0 \times m$ factor loadings. An important implication of this assumption 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 outcomes.

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

$$\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|_2 = \mathcal{O}_p\left(\sqrt{\frac{m}{n}}\right).$$

The main insight from Proposition 1 is that if $m/n \to 0$ the estimator is consistent. In the case in which the number of instruments m is fixed, this result establishes that the estimator is $\sqrt{n}$-consistent. Notice that this condition does not depend directly on the number of time periods T. However, if the matrix of instruments $\boldsymbol{\mathcal{V}}$ includes all columns for some $\boldsymbol{X}_k$, then the number of instruments will be increasing with T. For instance, suppose $K = 1$ and $\boldsymbol{\mathcal{V}} = (\boldsymbol{X}, \boldsymbol{W}\boldsymbol{X})$. In this event $m = \mathcal{O}(T)$ whereby $\hat{\boldsymbol{\theta}}$ would be consistent under Proposition 1 when $T/n \to 0$. This trivially includes where T is fixed. It is also worth highlighting that this result is obtained imposing quite weak conditions on the regression error $\boldsymbol{\eta}$. In particular heteroskedasticity is permitted in both dimensions of the panel, and the factors may be strong, weak or nonexistent.

5 Asymptotic Distribution

5.1 Main Result

Assumption AD below sets out conditions under which it is possible to derive an asymptotic expansion of the objective function around the true parameter value, and to then characterise the asymptotic distribution of the estimator. Ahead of this some additional notation is introduced. Let $\boldsymbol{\pi}_t$ be a $T \times 1$ canonical vector, that is, it consists of all zeros, except the t-th element which equals 1. The following matrices are also defined:

$$\begin{aligned}\bar{\boldsymbol{G}}(\alpha, \phi) &:= \alpha \boldsymbol{I}_{nT} + \phi \bar{\boldsymbol{W}},\\ \bar{\boldsymbol{C}}(\rho, \alpha, \phi) &:= (\boldsymbol{I}_{nT} + \bar{\boldsymbol{\Pi}}\bar{\boldsymbol{B}}^{-1}(\rho, \alpha, \phi)\bar{\boldsymbol{G}}(\alpha, \phi)),\end{aligned}$$

$\boldsymbol{S} := \boldsymbol{S}(\rho_0)$, $\bar{\boldsymbol{B}} := \bar{\boldsymbol{B}}(\rho_0, \alpha_0, \phi_0)$, $\bar{\boldsymbol{G}} := \bar{\boldsymbol{G}}(\alpha_0, \phi_0)$ and $\bar{\boldsymbol{C}} := \bar{\boldsymbol{C}}(\rho_0, \alpha_0, \phi_0)$. Moreover, let

$$\begin{aligned}
\boldsymbol{h}_1 &:= \bar{\boldsymbol{W}}\bar{\boldsymbol{B}}^{-1}\mathrm{vec}(\boldsymbol{X}\cdot\boldsymbol{\beta}_0 + \boldsymbol{\Lambda}_0\boldsymbol{F}_0^\top) & \boldsymbol{r}_{1,h} &:= \bar{\boldsymbol{W}}\bar{\boldsymbol{B}}^{-1}\bar{\boldsymbol{G}}\mathrm{vec}(\boldsymbol{A}^h\boldsymbol{S}^{-1}\boldsymbol{X}_{-h}\boldsymbol{\beta}_0\boldsymbol{\pi}_T^\top)\\
\boldsymbol{h}_2 &:= \bar{\boldsymbol{\Pi}}\bar{\boldsymbol{B}}^{-1}\mathrm{vec}(\boldsymbol{X}\cdot\boldsymbol{\beta}_0 + \boldsymbol{\Lambda}_0\boldsymbol{F}_0^\top) & \boldsymbol{r}_{2,h} &:= \bar{\boldsymbol{C}}\mathrm{vec}(\boldsymbol{A}^h\boldsymbol{S}^{-1}\boldsymbol{X}_{-h}\boldsymbol{\beta}_0\boldsymbol{\pi}_T^\top)\\
\boldsymbol{h}_3 &:= \bar{\boldsymbol{W}}\bar{\boldsymbol{\Pi}}\bar{\boldsymbol{B}}^{-1}\mathrm{vec}(\boldsymbol{X}\cdot\boldsymbol{\beta}_0 + \boldsymbol{\Lambda}_0\boldsymbol{F}_0^\top) & \boldsymbol{r}_{3,h} &:= \bar{\boldsymbol{W}}\bar{\boldsymbol{C}}\bar{\boldsymbol{G}}\mathrm{vec}(\boldsymbol{A}^h\boldsymbol{S}^{-1}\boldsymbol{X}_{-h}\boldsymbol{\beta}_0\boldsymbol{\pi}_T^\top)\\
\boldsymbol{h}_4 &:= \mathrm{vec}(\boldsymbol{X}_1) & \boldsymbol{r}_{4,h} &= \boldsymbol{0}_{nT}\\
&\vdots & &\vdots\\
\boldsymbol{h}_P &:= \mathrm{vec}(\boldsymbol{X}_K) & \boldsymbol{r}_{P,h} &= \boldsymbol{0}_{nT}
\end{aligned},$$

$\boldsymbol{r}_p := \sum_{h=0}^{\infty} \boldsymbol{r}_{p,h}$, $\boldsymbol{\mathcal{H}} := (\boldsymbol{h}_1, \ldots, \boldsymbol{h}_P)$, $\boldsymbol{\mathcal{R}} := (\boldsymbol{r}_1, \ldots, \boldsymbol{r}_P)$, and $\bar{\boldsymbol{M}} := (\boldsymbol{M}_{\boldsymbol{F}_0} \otimes \boldsymbol{Q}_{\boldsymbol{\mathcal{V}}} \boldsymbol{M}_{\tilde{\boldsymbol{\Lambda}}_0} \boldsymbol{Q}_{\boldsymbol{\mathcal{V}}}^\top)$.

Assumption AD (Asymptotic Distribution)**.**

(i) $R = R_0$.

(ii) $\frac{1}{n} \tilde{\boldsymbol{\Lambda}}_0^\top \tilde{\boldsymbol{\Lambda}}_0 \xrightarrow{p} \boldsymbol{\Sigma}_{\tilde{\boldsymbol{\Lambda}}_0}$ as $n \to \infty$, with $\mu_{R_0}(\boldsymbol{\Sigma}_{\tilde{\boldsymbol{\Lambda}}_0}) > 0$ and $\mu_1(\boldsymbol{\Sigma}_{\tilde{\boldsymbol{\Lambda}}_0}) < \infty$.

(iii) $\mu_{R_0}(\frac{1}{T} \boldsymbol{F}_0^\top \boldsymbol{F}_0) > 0$ and $\mu_1(\frac{1}{T} \boldsymbol{F}_0^\top \boldsymbol{F}_0) < \infty$.

(iv) The elements of the matrices $\bar{\boldsymbol{M}} \tilde{\boldsymbol{\mathcal{H}}}$ and $\bar{\boldsymbol{M}} \tilde{\boldsymbol{\mathcal{R}}}$ have uniformly bounded fourth moments for all (n, T).

Assumption AD(i) assumes that the true number of factors is known. In the absence of lagged outcomes Moon and Weidner (2015) show that the asymptotic distribution of the LS-IFE is unaffected by overstatement of the number of factors, under certain conditions. Simulation evidence suggests that also holds true for the IV-IFE estimator, but to establish this formally lies beyond the scope of this paper. Methods to detect the correct number of factors are discussed elsewhere and, in particular, the eigenvalue ratio test described in Higgins (2022) (Section 6.3) is applicable in the present context. Assumptions AD(ii) and AD(iii) assume the factors and the transformed factor loadings are both strong, that is to say that the factor term has a nonnegligible impact on the variance of the regression error.

Theorem 1 (Asymptotic Distribution)**.** *Under Assumptions MD, CS and AD, with* $\gamma_{nm}^2 := m^2 T / n \to c \geq 0$ *as* $n, m \to \infty$,

$$\sqrt{nT}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}\left(\boldsymbol{\Delta}^{-1} \boldsymbol{\psi}, \boldsymbol{\Delta}^{-1} \boldsymbol{\Omega} \boldsymbol{\Delta}^{-1}\right),$$

where,

$$\boldsymbol{\psi}_n := \frac{1}{\sqrt{nT}} \begin{pmatrix} \operatorname{tr}\left(\boldsymbol{\Sigma} \bar{\boldsymbol{M}} \bar{\boldsymbol{W}} \bar{\boldsymbol{B}}^{-1}\right) \\ \mathbf{0}_{(K+2)} \end{pmatrix},$$

$\boldsymbol{\psi} := \operatorname{plim}_{n\to\infty} \boldsymbol{\psi}_n$, $\boldsymbol{\Delta}_n := (nT)^{-1} (\boldsymbol{\mathcal{H}} + \boldsymbol{\mathcal{R}})^\top \bar{\boldsymbol{M}} (\boldsymbol{\mathcal{H}} + \boldsymbol{\mathcal{R}})$, $\boldsymbol{\Delta} = \operatorname{plim}_{n\to\infty} \boldsymbol{\Delta}_n$, $\boldsymbol{\Omega}_n := (nT)^{-1} (\boldsymbol{\mathcal{H}} + \boldsymbol{\mathcal{R}})^\top \bar{\boldsymbol{M}} \boldsymbol{\Sigma} \bar{\boldsymbol{M}} (\boldsymbol{\mathcal{H}} + \boldsymbol{\mathcal{R}})$, $\boldsymbol{\Omega} := \operatorname{plim}_{n\to\infty} \boldsymbol{\Omega}_n$, *and* $\boldsymbol{\Sigma}$ *is an* $nT \times nT$ *matrix with diagonal elements* $\sigma_{11}^2, \ldots, \sigma_{nT}^2$ *and remaining elements equal to zero.*

Without further restrictions, the order of the term $\boldsymbol{\psi}_n$ is γ_{nm}.[8] As a consequence, if the dimension of the cross-section grows sufficiently fast relative to the number of instruments and the number of time periods such that $\gamma_{nm}^2 \to 0$ (this trivially includes where both m and T are fixed), then the IV-IFE estimator is asymptotically unbiased with a standard sandwich form for the covariance matrix. On the other hand, if $\gamma_{nm}^2 \to c > 0$, the estimator may be asymptotically biased. In any given instance, the precise

[8] Although $\boldsymbol{\psi}_n$ is stochastic, it nonetheless has a nonstochastic bound.

order of the bias depends on the structure of the weights matrix, and also how close to orthogonal the columns of the weights matrix are to the instruments. To see this, suppose, for simplicity, that $\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{I}_{nT}$ with $\sigma^2 > 0$. Using (**??**) in Appendix **??**, as $n \to \infty$ and $m/n, T/n \to 0$,[9]

$$\begin{aligned}\psi_{n,1} &= \frac{\sigma^2}{\sqrt{nT}} \mathrm{tr}(\bar{\boldsymbol{W}} \bar{\boldsymbol{S}}^{-1} \bar{\boldsymbol{P}}) + o(1) \\ &= \sigma^2 \sqrt{\frac{T}{n}} \mathrm{tr}(\boldsymbol{W} \boldsymbol{S}^{-1} \boldsymbol{P}_{\mathcal{V}}) + o(1),\end{aligned}$$

where $\bar{\boldsymbol{P}} \coloneqq (\boldsymbol{I}_T \otimes \boldsymbol{P}_{\mathcal{V}})$. Define

$$\cos(\vartheta) = \frac{\mathrm{tr}(\boldsymbol{W} \boldsymbol{S}^{-1} \boldsymbol{P}_{\mathcal{V}})}{\|\boldsymbol{W} \boldsymbol{S}^{-1}\|_F \|\boldsymbol{P}_{\mathcal{V}}\|_F},$$

which measures the angles between the columns of the matrices $\boldsymbol{W} \boldsymbol{S}^{-1}$ and $\boldsymbol{P}_{\mathcal{V}}$. Then

$$\psi_{n,1} = \sigma^2 \sqrt{\frac{mT}{n}} \cos(\vartheta) \|\boldsymbol{W} \boldsymbol{S}^{-1}\|_F + o(1),$$

using $\|\boldsymbol{P}_{\mathcal{V}}\|_F = \sqrt{m}$. The terms $\cos(\vartheta)$ and $\|\boldsymbol{W} \boldsymbol{S}^{-1}\|_F$ can be understood to represent, respectively, the contribution to the bias stemming from the closeness of the instruments and the weights matrix (in terms of the angle ϑ), and from the structure of the weights matrix itself. Since $\cos(\vartheta) \in [-1, 1]$, the sign of the bias will depend on how aligned the columns of $\boldsymbol{W}$ are with those of $\boldsymbol{P}_{\mathcal{V}}$. In the extreme, if $\boldsymbol{P}_{\mathcal{V}} \boldsymbol{W} = \boldsymbol{0}_{n \times n}$, then $\psi_{n,1}$ will be exactly zero.[10] More generally, with $c_u \geq \mu_i(\boldsymbol{W}) \geq c_l > 0$, $\cos(\vartheta) \to 0$ as $n \to \infty$, provided m does not grow too fast. One the other hand, using

$$\mu_i(\boldsymbol{W} \boldsymbol{S}^{-1}) = \frac{\mu_i(\boldsymbol{W})}{1 - \rho \mu_i(\boldsymbol{W})},$$

and

$$\sqrt{\sum_{i=1}^{n} \mu_i^2(\boldsymbol{W} \boldsymbol{S}^{-1})} \leq \|\boldsymbol{W} \boldsymbol{S}^{-1}\|_F,$$ [11]

then with $c_u \geq \mu_i(\boldsymbol{W}) \geq c_l > 0$, $\|\boldsymbol{W} \boldsymbol{S}^{-1}\|_F \to \infty$ as $n \to \infty$. It is, therefore, evident that there is a tension between these terms when it comes to determining the exact order of $\boldsymbol{\psi}_n$. The following example illustrates the role that the structure of the weights matrix, summarised by its eigenvalues, has on the magnitude of the asymptotic bias.

Example 1 (Group Structure)**.** Suppose that the cross-section is partitioned into G

[9] Notice (**??**) in Appendix **??** does not require the symmetry of $\boldsymbol{W}$.
[10] Notice $\boldsymbol{W}$ need not be full rank.
[11] See e.g. Fact 9.11.3. in Bernstein (2009).

disjoint groups, where within each groups, all units are mutually connected with equal weights. This produces a weights matrix of the from

$$\boldsymbol{W} = \begin{pmatrix} \boldsymbol{W}_1 & 0 & \dots & 0 \\ 0 & \boldsymbol{W}_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \boldsymbol{W}_G \end{pmatrix} \quad \text{with} \quad \boldsymbol{W}_g = \frac{1}{n_g - 1} \begin{pmatrix} 0 & 1 & \dots & 1 \\ 1 & 0 & \dots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \dots & 0 \end{pmatrix}, \tag{5.1}$$

for $g = 1, \dots, G$, where n_g denotes the number of units in group g which satisfies $n_1 + \ldots + n_G = n$. Let $n_{\text{min}} := \min_{1 \le g \le G} n_g$ and $n_{\text{max}} := \max_{1 \le g \le G} n_g$. Assuming $|\rho| < 1$, then the bounds established in Appendix **??** give,

$$\sqrt{\frac{T}{n}} \mu_{\text{min}}(\boldsymbol{\Sigma}) \sum_{i=1}^{m} \frac{\mu_{n-i+1}(\boldsymbol{W})}{1 - \rho\mu_{n-i+1}(\boldsymbol{W})} \le \psi_1 \le \sqrt{\frac{T}{n}} \mu_{\text{max}}(\boldsymbol{\Sigma}) \sum_{i=1}^{m} \frac{\mu_i(\boldsymbol{W})}{1 - \rho\mu_i(\boldsymbol{W})}, \tag{5.2}$$

which holds w.p.a.1 as $n \to \infty$ and where $\sigma^2_{\text{min}} := \mu_{\text{min}}(\boldsymbol{\Sigma})$ and $\sigma^2_{\text{max}} := \mu_{\text{max}}(\boldsymbol{\Sigma})$. Notice that with the weights matrix structured as in (5.1), $\mu_1(\boldsymbol{W}_g) = 1$ and $\mu_i(\boldsymbol{W}_g) = -1/(n_g - 1)$ for $i = 1, \dots, n_g - 1$. As a result, $\boldsymbol{W}$ will have G eigenvalues of 1, $n_1 - 1$ being $-1/(n_1 - 1)$, $n_2 - 1$ being $-1/(n_1 - 1)$, and so on. Using this

$$\sum_{i=1}^{m} \frac{\mu_{n-i+1}(\boldsymbol{W})}{1 - \rho\mu_{n-i+1}(\boldsymbol{W})} \ge -\frac{m}{n_{\text{min}} + (\rho - 1)},$$

and

$$\begin{aligned} \sum_{i=1}^{m} \frac{\mu_i(\boldsymbol{W})}{1 - \rho\mu_i(\boldsymbol{W})} &= \frac{m}{1 - \rho} && \text{if} \qquad G \ge m, \\ \sum_{i=1}^{m} \frac{\mu_i(\boldsymbol{W})}{1 - \rho\mu_i(\boldsymbol{W})} &\le \frac{G}{1 - \rho} - \frac{(m - G)}{n_{\text{max}} + (\rho - 1)} && \text{if} \qquad G < m. \end{aligned}$$

In this way, it can be seen that the order of the bias can be related to the number of groups and the maximum and minimum group size. In one scenario, if $n, m, G \to \infty$ and $n_{\text{min}}, n_{\text{max}}$ stay fixed, then (5.2) does not provide a sharper order than γ_{nm}. On the other hand, if $n, m, n_{\text{min}} \to \infty$, $m/n_{\text{min}} \to 0$ and G is held fixed, then

$$\psi_1 = \mathcal{O}\left(\sqrt{\frac{T}{n}}\right),$$

that is, the order of the bias will not depend on the number of instruments.

6 Further Results

6.1 Inference

In order to conduct inference, it is possible to estimate the vector $\boldsymbol{\psi}$ and the matrix $\boldsymbol{\Delta}$ and then construct a bias corrected estimator. This, coupled with a consistent estimator of $\boldsymbol{\Omega}$, allows for the construction of asymptotically valid confidence intervals. The following results are built on consistent estimators for the projectors $\boldsymbol{M}_{\boldsymbol{F}_0}$ and $\boldsymbol{M}_{\tilde{\boldsymbol{\Lambda}}_0}$. Notice, however, that the minimisers of (3.7) with respect to the factors and the transformed loadings are not unique. In order to resolve this indeterminacy, estimators are defined in the following manner. Assume $m > T$. Consider a (short) singular value decomposition of $(nT)^{-\frac{1}{2}}(\tilde{\boldsymbol{Y}} - \hat{\boldsymbol{\theta}} \cdot \tilde{\boldsymbol{Z}}) =: \sum_{t=1}^{T} s_t \boldsymbol{u}_t \boldsymbol{v}_t^\top$ with singular values $s_T \leq \ldots \leq s_1$. Then define $\hat{\tilde{\boldsymbol{\Lambda}}} \coloneqq (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_{R_0})$ and $\hat{\boldsymbol{F}} \coloneqq (s_1 \boldsymbol{v}_1, \ldots, s_{R_0} \boldsymbol{v}_{R_0})$. 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 $\boldsymbol{M}_{\tilde{\boldsymbol{\Lambda}}_0}$.

Define the estimators

$$
\begin{aligned}
\hat{\psi}_1 \coloneqq& \frac{1}{\sqrt{nT}} \operatorname{tr}\left(\hat{\boldsymbol{\Sigma}} \hat{\bar{\boldsymbol{M}}} \bar{\boldsymbol{W}} \hat{\bar{\boldsymbol{B}}}^{-1}\right), \\
\hat{\boldsymbol{\Delta}} \coloneqq& \frac{1}{nT} \boldsymbol{\mathcal{Z}}^\top \hat{\bar{\boldsymbol{M}}} \boldsymbol{\mathcal{Z}}, \\
\hat{\boldsymbol{\Omega}} \coloneqq& \frac{1}{nT} \boldsymbol{\mathcal{Z}}^\top \hat{\bar{\boldsymbol{M}}} \hat{\boldsymbol{\Sigma}} \hat{\bar{\boldsymbol{M}}} \boldsymbol{\mathcal{Z}},
\end{aligned}
$$

where $\boldsymbol{\mathcal{Z}} \coloneqq (\operatorname{vec}(\boldsymbol{Z}_1), \ldots, \operatorname{vec}(\boldsymbol{Z}_P))$, $\hat{\boldsymbol{\Sigma}} \coloneqq \Gamma(\operatorname{vec}(\boldsymbol{Y} - \hat{\boldsymbol{\theta}} \cdot \boldsymbol{Z})\operatorname{vec}(\boldsymbol{Y} - \hat{\boldsymbol{\theta}} \cdot \boldsymbol{Z})^\top)$, $\hat{\bar{\boldsymbol{B}}} \coloneqq \bar{\boldsymbol{B}}(\hat{\rho}, \hat{\alpha}, \hat{\phi})$, and $\Gamma(\boldsymbol{A}) \coloneqq (\boldsymbol{A} \odot (\boldsymbol{\iota}_T \boldsymbol{\iota}_T^\top \otimes \boldsymbol{I}_n))$ for an $nT \times nT$ matrix $\boldsymbol{A}$ serves as a truncation kernel.[12]

Proposition 2 (Inference)**.** *Under Assumptions MD, CS and AD, with $\gamma_{nm}^2 \to c \geq 0$ as $n, m \to \infty$,*

$$
\begin{aligned}
\|\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}\|_2 &= \mathcal{O}_p(1), \\
\|\hat{\boldsymbol{\Delta}} - \boldsymbol{\Delta}\|_2 &= \mathcal{O}_p(1), \\
\|\hat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}\|_2 &= \mathcal{O}_p(1),
\end{aligned}
$$

where $\hat{\boldsymbol{\psi}} \coloneqq (\hat{\psi}_1, 0, \ldots, 0)^\top$.

On the basis of Proposition 2, a bias corrected estimator $\tilde{\boldsymbol{\theta}} \coloneqq \hat{\boldsymbol{\theta}} - (nT)^{-\frac{1}{2}} \hat{\boldsymbol{\Delta}}^{-1} \hat{\boldsymbol{\psi}}$ can be constructed and inference may proceed in standard fashion thereafter.

[12] $\odot$ denotes the Hadamard product.

6.2 Specification Testing

Notice that Assumption MD implies the following moment conditions

$$\mathbb{E}\left[\bar{\boldsymbol{M}}\mathrm{vec}(\boldsymbol{\eta})\right] = \mathbb{E}\left[\left(\boldsymbol{M}_{\boldsymbol{F}_0} \otimes \boldsymbol{Q}_{\boldsymbol{\mathcal{V}}}\boldsymbol{M}_{\tilde{\boldsymbol{\Lambda}}_0}\boldsymbol{Q}_{\boldsymbol{\mathcal{V}}}^\top\right)\mathrm{vec}(\boldsymbol{\eta})\right] = \boldsymbol{0}_{nT}. \tag{6.1}$$

Indeed, the objective function (3.7) can be equivalently reformulated as

$$\mathcal{Q}(\boldsymbol{\theta}) = \min_{\tilde{\boldsymbol{\Lambda}},\boldsymbol{F}} \frac{1}{nT}\left((\boldsymbol{M}_{\boldsymbol{F}} \otimes \boldsymbol{M}_{\tilde{\boldsymbol{\Lambda}}}\boldsymbol{Q}_{\boldsymbol{\mathcal{V}}}^\top)\mathrm{vec}(\boldsymbol{Y} - \boldsymbol{\theta}\cdot\boldsymbol{Z})\right)^\top\left((\boldsymbol{M}_{\boldsymbol{F}} \otimes \boldsymbol{M}_{\tilde{\boldsymbol{\Lambda}}}\boldsymbol{Q}_{\boldsymbol{\mathcal{V}}}^\top)\mathrm{vec}(\boldsymbol{Y} - \boldsymbol{\theta}\cdot\boldsymbol{Z})\right).$$

In this way it becomes apparent that the estimator can be interpreted as an (unweighted) GMM estimator based on the moment conditions (6.1).[13] A model specification test can be constructed on this basis and takes the form of the following J-statistic

$$\mathcal{J} \coloneqq \mathrm{vec}(\hat{\boldsymbol{\eta}})^\top\left(\boldsymbol{M}_{\hat{\boldsymbol{F}}} \otimes \boldsymbol{Q}_{\boldsymbol{\mathcal{V}}}\boldsymbol{M}_{\hat{\tilde{\boldsymbol{\Lambda}}}}\boldsymbol{Q}_{\boldsymbol{\mathcal{V}}}^\top\right)\mathrm{vec}(\hat{\boldsymbol{\eta}}),$$

with $\hat{\boldsymbol{\eta}} \coloneqq \mathrm{vec}(\boldsymbol{Y} - \hat{\boldsymbol{\theta}}\cdot\boldsymbol{Z})$. Let

$$\boldsymbol{M}_{\mathcal{J}} \coloneqq \left(\bar{\boldsymbol{M}} - \bar{\boldsymbol{M}}(\boldsymbol{\mathcal{H}}+\boldsymbol{\mathcal{R}})\left((\boldsymbol{\mathcal{H}}+\boldsymbol{\mathcal{R}})^\top\bar{\boldsymbol{M}}(\boldsymbol{\mathcal{H}}+\boldsymbol{\mathcal{R}})\right)^{-1}(\boldsymbol{\mathcal{H}}+\boldsymbol{\mathcal{R}})^\top\bar{\boldsymbol{M}}\right),$$

$\ell \coloneqq (T-R_0)(m-R_0) - P$, and

$$\sigma_{\mathcal{J}}^2 \coloneqq \mathrm{tr}((\boldsymbol{\mathcal{M}}^{(4)} - \boldsymbol{\Sigma}^2)(\boldsymbol{M}_{\mathcal{J}} \odot \boldsymbol{M}_{\mathcal{J}})) + \ 2\boldsymbol{\iota}_{nT}^\top\mathrm{off}(\boldsymbol{\Sigma}(\boldsymbol{M}_{\mathcal{J}} \odot \boldsymbol{M}_{\mathcal{J}})\boldsymbol{\Sigma})\boldsymbol{\iota}_{nT},$$

where $\boldsymbol{\mathcal{M}}^{(4)}$ is an $nT \times nT$ matrix with diagonal elements $\mathbb{E}[\varepsilon_{11}^4], \ldots, \mathbb{E}[\varepsilon_{nT}^4]$ and all remaining elements equal to zero. The following are assumed.

Assumption JT (J-Test)**.**

(i) The errors ε_{it} have uniformly bound eighth moments.

(ii) $\ell^{-1}\sigma_{\mathcal{J}}^2 \geq c > 0$ w.p.a.1.

Assumption JS(i) strengthens the number of finite moments required of the primitive error ε. Assumption JS(ii) ensures the standardised test statistic is well defined.

Theorem 2 (J-Test)**.** *Under Assumptions MD, CS, AD and JS, with $\gamma_{nm}^2 \to c \geq 0$ as $n, \ell \to \infty$,*

$$\frac{\mathcal{J} - \varphi_{\mathcal{J}}}{\sigma_{\mathcal{J}}} \xrightarrow{d} \mathcal{N}(0,1),$$

where

$$\varphi_J \coloneqq \mathrm{tr}(\boldsymbol{\Sigma}\boldsymbol{M}_{\mathcal{J}}) - \boldsymbol{\psi}_n^\top\boldsymbol{\Delta}_n^{-1}(\boldsymbol{\mathcal{H}}+\boldsymbol{\mathcal{R}})^\top\bar{\boldsymbol{M}}(\boldsymbol{\mathcal{H}}+\boldsymbol{\mathcal{R}})\boldsymbol{\Delta}_n^{-1}\boldsymbol{\psi}_n.$$

[13] See Section 6 in Higgins (2022) for further discussion.

The first term in of φ_J is a standard degrees of freedom adjustment. If, for example, the errors ε_{it} were independently distributed with unit variances then $\mathrm{tr}(\boldsymbol{\Sigma M}_{\mathcal{J}}) = \ell$. Indeed, in such a case $\sigma^2_{\mathcal{J}}$ would also collapse to 2ℓ which, absent of the second term in φ_J, would produce the result

$$\frac{\mathcal{J}-\ell}{\sqrt{2\ell}} \xrightarrow{d} \mathcal{N}(0,1).$$

Such a result is familiar in the context of model specification testing with many instruments; see, for example, Donald et al. (2003) and Anatolyev and Gospodinov (2011). However, in this instance the asymptotic bias of the estimator $\hat{\boldsymbol{\theta}}$ also effects the asymptotic distribution of the test statistic and produces an additional term in φ_J.

6.3 Additive Effects

The bias $\boldsymbol{\psi}_n$ that appears in Theorem 1 originates from the implicit transformation of the model through the projectors $\boldsymbol{M}_{\boldsymbol{F}_0}$ and $\boldsymbol{M}_{\tilde{\boldsymbol{\Lambda}}_0}$ used to purge the factor term from the error. This does not, as one might expect, arise as a consequence of having estimated these projector matrices. Thus, such bias would occur even were the factors and loadings observed. Suppose, for example, that the data are generated according to

$$\boldsymbol{Y} = \rho \boldsymbol{WY} + \alpha \boldsymbol{Y}_{-1} + \phi \boldsymbol{WY}_{-1} + \boldsymbol{\beta} \cdot \boldsymbol{X} + \boldsymbol{\Lambda \Gamma} + \boldsymbol{LF}^\top + \boldsymbol{\varepsilon},$$

where $\boldsymbol{\Lambda}$ and $\boldsymbol{F}$ are observed matrices of dimension $n \times R_\lambda$ and $T \times R_f$, respectively, and $\boldsymbol{L}$ and $\boldsymbol{\Gamma}$ are unknown $n \times R_\lambda$ and $T \times R_f$ matrices. This nests as special cases familiar models of additive effects such as individual and time effects which corresponds to $\boldsymbol{\Lambda} = \boldsymbol{\iota}_n$ and $\boldsymbol{F} = \boldsymbol{\iota}_T$. One may construct the estimator

$$\hat{\boldsymbol{\theta}}_{AE} \coloneqq \left(\boldsymbol{\mathcal{Z}}^\top \bar{\boldsymbol{M}} \boldsymbol{\mathcal{Z}}\right)^{-1} \boldsymbol{\mathcal{Z}}^\top \bar{\boldsymbol{M}} \mathrm{vec}(\boldsymbol{Y}),$$

where $\bar{\boldsymbol{M}}$ is now known. As $n \to \infty$ and $\gamma^2_{mn} \to c \geq 0$, under conditions analogous to Assumptions MD and CS, this estimator will be asymptotically equivalent to $\hat{\boldsymbol{\theta}}$. Nonetheless, although the leading order bias of both estimators is the same, there remain terms of a lower order that arise in the asymptotic expansion of $\hat{\boldsymbol{\theta}}$, as a consequence of the factor and the loadings being unknown that would not arise for $\hat{\boldsymbol{\theta}}_{AE}$.

7 Numerical Examples

This section describes results from two Monte Carlo experiments designed to give a sense of how the properties of the IV-IFE estimator are affected by the structure of the weights matrix.

7.1 Circular World

In this first experiment outcomes are generated according to

$$\boldsymbol{Y} = \rho_0 \boldsymbol{W}\boldsymbol{Y} + \alpha_0 \boldsymbol{Y}_{-1} + \beta_{0,1}\boldsymbol{X}_1 + \beta_{0,2}\boldsymbol{X}_2 + \boldsymbol{\Lambda}_0 \boldsymbol{F}_0^\top + \boldsymbol{\varepsilon},$$

with $\rho_0 = 0.5$, $\alpha_0 = -0.25$, $\beta_{0,1} = 1$, and $\beta_{0,2} = -1$. The covariate $\boldsymbol{X}_1$ is generated as

$$\boldsymbol{X}_1 = \boldsymbol{\Lambda}_0 \boldsymbol{F}_0^\top + \boldsymbol{\epsilon},$$

where $R_0 = 2$, and $\lambda_{0,ir}$, $f_{0,tr}$ and ϵ_{it} are drawn from independent standard normal distributions. The covariate $\boldsymbol{X}_2$ is uncorrelated with the factors and the loadings with elements drawn from independent standard normal distributions. The error term is generated as $\operatorname{vec}(\boldsymbol{\varepsilon}) := \boldsymbol{\Sigma}^{\frac{1}{2}} \operatorname{vec}(\boldsymbol{u})$, where $\boldsymbol{\Sigma}$ is a diagonal matrix with nonzero elements drawn uniformly from the interval $(0, 2)$ and $u_{it} \sim \mathcal{N}(0, 1)$. The matrix of instruments is specified as $\boldsymbol{\mathcal{V}} := (\boldsymbol{X}_1, \boldsymbol{W}\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{W}\boldsymbol{X}_2)$, with the weights matrix initially being generated in the following way:

$$\boldsymbol{W} := \begin{pmatrix} 0 & 1 & 1 & 0 & \dots & 1 \\ 1 & 0 & 1 & 1 & \dots & 0 \\ 1 & 1 & 0 & 1 & \dots & 0 \\ 0 & 1 & 1 & 0 & \dots & 1 \\ \vdots & \vdots & \vdots & \vdots & \ddots & 1 \\ 1 & 0 & 0 & 1 & 1 & 0 \end{pmatrix}, \tag{7.1}$$

before then being row-normalised. This matrix represents a 'circular world' in which all the cross-sectional units have a unique location on a circle, and are connected only to their nearest two 'neighbours' on either side (see, for example, Das et al. (2003)). Tables 1a and 1b below present, respectively, bias and coverage of a 95% two-sided confidence interval for the IV-IFE estimator, across various n and T combinations, averaged over 1000 Monte Carlo draws.

Table 1a: Bias $\hat{\boldsymbol{\theta}}$ - Circular World

	$\hat{\rho}$			$\hat{\alpha}$			$\hat{\beta}_1$			$\hat{\beta}_2$		
$n \setminus T$	6	9	12	6	9	12	6	9	12	6	9	12
100	0.057	0.095	0.097	0.022	0.042	0.030	-0.012	-0.015	-0.016	0.021	0.029	0.019
300	0.028	0.040	0.052	0.015	0.017	0.022	-0.008	-0.007	-0.010	0.006	0.005	0.006
500	0.017	0.027	0.031	0.009	0.014	0.014	-0.006	-0.005	-0.005	0.002	0.005	0.005

Table 1b: Coverage 95% Confidence Interval $\hat{\boldsymbol{\theta}}$ - Circular World

	$\hat{\rho}$			$\hat{\alpha}$			$\hat{\beta}_1$			$\hat{\beta}_2$		
$n \setminus T$	6	9	12	6	9	12	6	9	12	6	9	12
100	0.654	0.119	0.013	0.789	0.345	0.454	0.916	0.913	0.912	0.912	0.879	0.900
300	0.797	0.418	0.008	0.827	0.675	0.395	0.937	0.924	0.919	0.939	0.941	0.928
500	0.845	0.545	0.257	0.862	0.695	0.567	0.942	0.932	0.931	0.946	0.938	0.942

The bias of the estimator is relatively small and declines quickly as n increases. However, in terms of coverage the performance of the estimator is much worse, especially for the spatial coefficient ρ. Indeed, while in all cases coverage of ρ improves as n increases, it is very sensitive to the value of T and rapidly declines as this becomes larger. Tables 2a and 2b below present, respectively, bias and coverage of a 95% two-sided confidence interval, based on the bias corrected estimator $\tilde{\boldsymbol{\theta}}$ described in Section 6.1.

Table 2a: Bias $\tilde{\boldsymbol{\theta}}$ - Circular World

	$\tilde{\rho}$			$\tilde{\alpha}$			$\tilde{\beta}_1$			$\tilde{\beta}_2$		
$n \setminus T$	6	9	12	6	9	12	6	9	12	6	9	12
100	-0.011	-0.025	-0.024	0.000	-0.002	-0.004	0.007	0.009	0.007	-0.002	-0.007	-0.004
300	-0.003	-0.005	-0.007	0.002	0.000	-0.001	0.001	0.002	0.002	-0.002	-0.001	-0.002
500	-0.001	-0.002	-0.003	0.000	-0.001	-0.001	0.001	0.001	0.000	-0.002	-0.001	-0.001

Table 2b: Coverage 95% Confidence Interval $\tilde{\boldsymbol{\theta}}$ - Circular World

	$\tilde{\rho}$			$\tilde{\alpha}$			$\tilde{\beta}_1$			$\tilde{\beta}_2$		
$n \setminus T$	6	9	12	6	9	12	6	9	12	6	9	12
100	0.953	0.927	0.877	0.914	0.889	0.912	0.917	0.931	0.937	0.938	0.935	0.946
300	0.963	0.959	0.948	0.944	0.929	0.942	0.960	0.939	0.950	0.945	0.950	0.951
500	0.950	0.966	0.964	0.939	0.931	0.936	0.945	0.958	0.956	0.950	0.955	0.954

When compared to tables 1a and 1b, bias correction produces moderate improvements in the bias of the estimator, and substantial improvements in coverage. This is particularly apparent when considering the coverage probabilities for the spatial coefficient ρ when $T = 12$.

7.2 Multiple Stars

In the second Monte Carlo experiment, outcomes are generated in the same way as before, aside from the structure of the weights matrix. In particular the weights matrix is specified to be similar to (5.1) in that it is assumed to have a block structure, except that now within each group all cross-sectional units are connected to only a single central unit, which reciprocates the link, and to no others. More specifically, to generate $\boldsymbol{W}$, the first G cross-sectional units are assigned to group $g = 1, \ldots, G$, after which the remaining $n - G$ units are assigned to one of these G groups with probability g/G.

$$\boldsymbol{W} = \begin{pmatrix} \boldsymbol{W}_1 & 0 & \dots & 0 \\ 0 & \boldsymbol{W}_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \boldsymbol{W}_G \end{pmatrix} \quad \text{with} \quad \boldsymbol{W}_g = \begin{pmatrix} 0 & (n_g-1)^{-1} & \dots & (n_g-1)^{-1} \\ 1 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & \dots & 0 \end{pmatrix},$$

Two different values of G are considered: 5 and 25. The noteworthy feature of this design is that the rank of the weights matrix will be equal to $2G$, in which case

$$\boldsymbol{\psi}_n = \boldsymbol{\mathcal{O}}\left((T \wedge 2G) \times \sqrt{\frac{T}{n}}\right).$$

Thus, with a larger number of groups the bias of the IV-IFE estimator is expected to be relatively large, and, conversely, the bias should be relatively small with a smaller number of groups. Tables 3a and 3b below display coverage of a 95% two-sided confidence interval for both the uncorrected and corrected estimates when $G = 5$. Tables 4a and 4b display corresponding results for $G = 25$. Tables which present the corresponding bias can be found in Appendix ??.

Table 3a: Coverage 95% Confidence Interval $\hat{\boldsymbol{\theta}}$ - Multiple Stars $G = 5$

	$\hat{\rho}$			$\hat{\alpha}$			$\hat{\beta}_1$			$\hat{\beta}_2$		
$n \setminus T$	6	9	12	6	9	12	6	9	12	6	9	12
100	0.902	0.843	0.853	0.908	0.884	0.916	0.950	0.938	0.964	0.949	0.954	0.947
300	0.928	0.921	0.916	0.929	0.926	0.933	0.955	0.951	0.950	0.951	0.961	0.945
500	0.939	0.939	0.929	0.929	0.949	0.939	0.941	0.937	0.962	0.937	0.956	0.952

Table 3b: Coverage 95% Confidence Interval $\tilde{\boldsymbol{\theta}}$ - Multiple Stars $G = 5$

	$\hat{\rho}$			$\hat{\alpha}$			$\hat{\beta}_1$			$\hat{\beta}_2$		
$n \setminus T$	6	9	12	6	9	12	6	9	12	6	9	12
100	0.932	0.951	0.964	0.938	0.927	0.949	0.950	0.938	0.965	0.948	0.954	0.947
300	0.953	0.944	0.960	0.946	0.940	0.938	0.955	0.951	0.950	0.950	0.961	0.946
500	0.954	0.949	0.947	0.933	0.954	0.942	0.939	0.937	0.962	0.937	0.956	0.953

Table 4a: Coverage 95% Confidence Interval $\hat{\boldsymbol{\theta}}$ - Multiple Stars $G = 25$

	$\hat{\rho}$			$\hat{\alpha}$			$\hat{\beta}_1$			$\hat{\beta}_2$		
$n \setminus T$	6	9	12	6	9	12	6	9	12	6	9	12
100	0.815	0.798	0.353	0.850	0.860	0.757	0.951	0.962	0.939	0.941	0.949	0.938
300	0.918	0.892	0.698	0.889	0.920	0.812	0.947	0.952	0.943	0.954	0.959	0.956
500	0.935	0.913	0.861	0.924	0.949	0.931	0.944	0.958	0.947	0.954	0.950	0.954

Table 4b: Coverage 95% Confidence Interval $\tilde{\boldsymbol{\theta}}$ - Multiple Stars $G = 25$

	$\hat{\rho}$			$\hat{\alpha}$			$\hat{\beta}_1$			$\hat{\beta}_2$		
$n \setminus T$	6	9	12	6	9	12	6	9	12	6	9	12
100	0.956	0.951	0.962	0.915	0.924	0.953	0.950	0.950	0.940	0.944	0.954	0.952
300	0.956	0.961	0.961	0.925	0.959	0.932	0.952	0.949	0.947	0.955	0.961	0.953
500	0.948	0.965	0.960	0.937	0.952	0.945	0.946	0.956	0.953	0.953	0.952	0.954

As expected, when $G = 5$ the bias of the estimator is small, though nonetheless coverage is still improved by bias correction. When $G = 25$ the bias is more pronounced, with bias correction leading to significant improvements in coverage of ρ, particularly for larger values of T.

8 Empirical Application

8.1 Economic Growth, Civil Liberties and Political Rights in the 21st Century

In this section the IV-IFE estimator is applied to study the impact of civil liberties and political rights on economic growth. This study is in the spirit of Acemoglu et al. (2019) who study the long term impact of democracy on GDP per capita using, amongst other methods, a dynamic panel model with fixed effects. The data consists of a balanced panel containing data on 180 countries observed between the years 2001 and 2020. The dependent variable is the log of GDP per capita measured in 2015 US dollars.[14] The main explanatory variable is a dichotomous indicator of civil liberties and political rights d_{it} derived from the Freedom in the World Index compiled by Freedom House. This variable encodes

$$d_{it} = \begin{cases} 0 \text{ if classified as } \textit{not free}, \\ 1 \text{ if classified as } \textit{free} \text{ or } \textit{partially free}. \end{cases}$$

In order to quantify spatial dependence in outcomes, a weights matrix is generated based on the distance between countries at their nearest point. In order to generate this, a (equirectangular) projection of the globe is taken from the World Bank which consists of very high resolution coordinates of international boundaries. These coordinates are rounded to generate a lower resolution projection which describes the shape of countries using fewer data points, based on which the great-circle distance is calculated between every pair of coordinates using the haversine formula. For each country pair ij, let δ_{ij} denote the shortest distance between i and j, and let e denote half the distance of the equator. Then the $n \times n$ weights matrix $\boldsymbol{W}$ is generated by setting element w_{ij} equal to $\delta_{ij} = 1 - \delta_{ij}/e$ if $\delta_{ij}/e < \tau$, and $\delta_{ij} = 0$ if $\delta_{ij}/e \geq \tau$, where τ is a prespecified threshold.[15]

[14] The data are obtained from the World Bank.
[15] In estimation this threshold is set to 0.1.

Comparing the study conducted here to that of Acemoglu et al. (2019) there are a few important differences to highlight. The first is that while Acemoglu et al. (2019) use an unbalanced panel spanning the period $1960-2010$, the panel constructed here is balanced and covers the period $2001-2020$. Using a more recent dataset is useful since observations are available for most countries in the world over this period. This may help to mitigate any possible selection issues, in the sense that wealthier, democratic countries are disproportionately represented in the earlier years of the panel constructed by Acemoglu et al. (2019). Also, since the primary focus of this study is to consider spatial dependence, the omission of data for particularly influential countries may be consequential. Second, while Acemoglu et al. (2019) do consider that there may be spatial correlation in the data, and report additional estimation results after controlling for a spatial lag of GDP, they do not fully explore the possibility of this as an additional mechanism through which a transition to democracy may accrue more substantial long term effects on GDP. Thirdly, the authors control only for individual and time effects and not interactive effects.

8.2 Dynamic Model

To begin with a first order dynamic model is estimated, that is,

$$\boldsymbol{y}_t = \alpha \boldsymbol{y}_{t-1} + \beta \boldsymbol{d}_t + \boldsymbol{\eta}_t. \tag{8.1}$$

This specification allows for dynamics but is absent of a spatial component and allows for a more direct comparison to be made with the results obtained in Acemoglu et al. (2019). The long term impact is calculated as $\gamma := \beta/(1-\alpha)$ and is the main statistic of interest. Table 1 below provides four different sets of results. The columns headed 'FE' provides output for the model (8.1) estimated by least squares and controlling for individual and time effects, the column headed 'IV-IFE $\boldsymbol{\mathcal{V}}^{(1)}$' provides estimation results from applying the IV-IFE estimator using instrument set $\boldsymbol{\mathcal{V}}^{(1)}$, the column headed 'IV-IFE $\boldsymbol{\mathcal{V}}^{(2)}$' provides estimation results from applying the IV-IFE estimator using instrument set $\boldsymbol{\mathcal{V}}^{(2)}$, and the column headed 'ANRR' provides a range of results for the equivalent specification to (8.1) found in Acemoglu et al. (2019). The first set of of instruments is $\boldsymbol{\mathcal{V}}^{(1)} := (\boldsymbol{d}_1, \ldots, \boldsymbol{d}_T)$, that is, consists of all past, present and future values of $\boldsymbol{d}_t$. The second set expands this to also include spatial lags of $\boldsymbol{d}_t$, i.e. $\boldsymbol{\mathcal{V}}^{(2)} := (\boldsymbol{d}_1, \ldots, \boldsymbol{d}_T, \boldsymbol{W}\boldsymbol{d}_1, \ldots, \boldsymbol{W}\boldsymbol{d}_T)$. The number of factors in estimation is selected in the following way. First, by Proposition 1, $\boldsymbol{\theta}_0$ can be consistently estimated with knowledge only of an upper bound on the number of factors. Thus, inputting a large value for R, consistent estimates of the coefficients can be obtained.[16] Using these coefficient estimates, a pure factor model can be constructed and the true number of factors detected using the eigenvalue ratio test described in Higgins (2022). Finally, the model

[16] Here $R = 5$ is used.

is re-estimated inputting the detected number of factors (in this case 1) to obtain the final estimates.

Table 1: Dynamic Model

		$\boldsymbol{\mathcal{V}}^{(1)}$	$\boldsymbol{\mathcal{V}}^{(2)}$	
	FE	**IV-IFE**	**IV-IFE**	**ANRR**
β	**0.0161**	**0.0141**	**0.0124**	**0.0078 - 0.0097**
t-stat	3.9000	2.5343	2.286	
α	**0.9228**	**0.8435**	**0.9335**	**0.938 - 0.973**
t-stat	146.3841	24.7520	76.8929	
γ	**0.2086**	**0.0901**	**0.1865**	**0.1264 - 0.3558**
t-stat	3.7379	2.7802	2.4492	
J-stat	-	**0.2096**	**0.4032**	

The coefficient β is found to be similar using both the FE and IV-IFE estimators. Though these are slightly larger than the values found in ANRR, they share the same sign and are both significant. On the other hand the autoregressive coefficient α exhibits more substantial differences. While the FE coefficient is found to be similar to the value found by ANRR, as is the IV-IFE estimator which uses instrument $\boldsymbol{\mathcal{V}}^{(2)}$, the value found using the IV-IFE estimator and instruments $\boldsymbol{\mathcal{V}}^{(1)}$ is much smaller. As a consequence, the long run effect γ found using $\boldsymbol{\mathcal{V}}^{(1)}$ is also much smaller, and indeed falls outside of the range of values found in ANRR. This difference in value can be attributed to the fact that, ultimately, $\boldsymbol{d}_t$ is a poor instrument. Firstly, $\boldsymbol{d}_t$ does not have a great deal of variation. Indeed over the course of the period $2001 - 2020$, across all 180 countries in the panel (constituting some 3600 observations), only 65 changes in status are observed. Second, as is evident from estimates of β in Table 1, $\boldsymbol{d}_t$ is only quite weakly correlated with GDP. This then leads the IV-IFE estimator using instruments $\boldsymbol{\mathcal{V}}^{(1)}$ to underestimate the degree of persistence in the data. Motivated by the suspicion that the data truly exhibit spatial dependence in outcomes, $\boldsymbol{\mathcal{V}}^{(2)}$ uses spatial lags of $\boldsymbol{d}_t$ as additional instruments, which leads to estimation results much closer to those found using the FE estimator and in ANRR. The fact that augmenting the instrument set with spatial lags of $\boldsymbol{d}_t$ has a significant impact on the estimates already suggests that the data exhibit some degree of spatial correlation motivating the results to follow for the full dynamic spatial model.

8.3 Dynamic Spatial Model

This section provides estimation results for a first-order dynamic spatial model

$$\boldsymbol{y}_t = \rho \boldsymbol{W} \boldsymbol{y}_t + \alpha \boldsymbol{y}_{t-1} + \phi \boldsymbol{W} \boldsymbol{y}_{t-1} + \beta \boldsymbol{d}_t + \boldsymbol{\eta}_t.$$

For the IV-IFE estimator the set of instruments $\boldsymbol{\mathcal{V}}^{(2)}$ is used (described in the previous subsection), which includes all values of $\boldsymbol{d}_t$ and first order spatial lags $\boldsymbol{W}\boldsymbol{d}_t$. Unlike in the dynamic model, there is now no longer a common long run effect over the cross-section, but rather an $n \times n$ matrix of long run effects, where element ij of this matrix captures the impact that an increase in political rights and civil liberties in country i has on the GDP of a country j, in the long run. Two summary statistics of this matrix are provided. The first is the direct effect defined to be

$$\gamma_D := \frac{\beta}{n} \text{tr} \left(((1-\alpha)\boldsymbol{I}_n - (\rho + \phi)\boldsymbol{W})^{-1} \right).$$

This is the generalisation of the long run effect described in the previous subsection and is the average long term effect that an increase in political rights and civil liberties in country i has on its own GDP. Evidently, if there were no spatial spillovers and $\rho = \phi = 0$, then γ_D would collapse to equal γ. The second summary statistic is the indirect effect defined to be

$$\gamma_I := \frac{\beta}{n^2} \boldsymbol{\iota}_n^\top \left((1-\alpha)\boldsymbol{I}_n - (\rho + \phi)\boldsymbol{W})^{-1} \right) \boldsymbol{\iota}_n - \gamma_D.$$

This describes the average long term effect that an increase in political rights and civil liberties in country i has on the GDP of another country j. If there were no spatial spillovers and $\rho = \phi = 0$, then $\gamma_I = 0$.

Table 5: Dynamic Spatial Model

	FE	**IV-IFE**	**IV-IFE-BC**		**FE**	**IV-IFE**	**IV-IFE-BC**
β	**0.0158**	**0.0111**	**0.0112**	ϕ	**-0.0118**	**-0.0156**	**-0.0084**
t-stat	3.8585	2.0618	2.0781	tstat	-7.3287	-8.0175	-4.3520
α	**0.9224**	**0.9347**	**0.9374**	γ_D	**0.2042**	**0.1716**	**0.1793**
t-stat	145.5072	78.0726	78.2993	t-stat	3.7029	2.1954	2.2006
ρ	**0.0125**	**0.0164**	**0.0082**	γ_I	**0.0645**	**0.1142**	**-0.0184**
t-stat	6.9085	8.0257	4.0043	t-stat	0.9342	0.8839	-0.5809
				J-stat	-	**0.1592**	**0.0938**

Table 5 displays the estimation results for the dynamic spatial model. The column headed 'FE' display estimates controlling for individual and time effects using the 2SLS estimator described in Lee and Yu (2014). The columns headed 'IV-IFE' display IV-IFE estimates without bias correction, and the columns headed 'IV-IFE-BC' display bias corrected estimates. The estimates for β and α are similar to those obtained for the dynamic model. The spatial coefficient ρ is found to be significant and positive, indicating that higher values of GDP per capita in neighbouring countries is associated with a higher value of GDP per capita in that country itself. The coefficient ϕ is also found to be signifiant and though of similar magnitude, it is of opposite sign to ρ. Interestingly, the direct effect γ_D is found to be quite similar to the effect found for the dynamic model in spite of the addition of the spatial regressors which have been found to be relevant. This occurs because, although immediately the spatial dependence positively amplifies the impact of an increase in political rights and civil liberties, this effect is offset by a negative impact in the subsequent time period. In the long run the positive effect through ρ and the negative effect through ϕ offset each other.

9 Conclusion

To conclude, this paper introduces an IV estimator for dynamic spatial models with interactive effects that provides consistent and asymptotically unbiased estimates when cross-sectional dimension n is large relative to the number of time period T and the number of instruments m. However, circumstances exist where the estimator can exhibit asymptotic bias, which largely depends on the structure of the weights matrix, and on the set of instruments. Monte Carlo experiments suggest that bias correction is an effective remedy to this problem, and an empirical application of the methods supports the findings of Acemoglu et al. (2019) that increased political rights and civil liberties within a country results in higher values of GDP per capita in the long run. The estimation approach described in this paper applies more widely than to the dynamic spatial model alone, therefore future work will explore the application of this method to problems with endogenous regressors more generally.

References

Acemoglu, D., Naidu, S., Restrepo, P., Robinson, J. A., 2019. Democracy does cause growth. Journal of Political Economy 127 (1), 47–100.

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.

Anatolyev, S., Gospodinov, N., 2011. Specification testing in models with many instruments. Econometric Theory 27 (2), 427–441.

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

Bai, J., Li, K., 2021. Dynamic spatial panel data models with common shocks. Journal of Econometrics 224 (1), 134–160.

Bernstein, D. S., 2009. Matrix mathematics: theory, facts, and formulas, 2nd Edition. Princeton University Press, Princeton, N.J, oCLC: ocn243960539.

Cliff, A. D., Ord, J. K., 1973. Spatial autocorrelation. London: Pion.

Cliff, A. D., Ord, J. K., 1981. Spatial processes: models & applications. London: Pion.

Cui, G., Sarafidis, V., Yamagata, T., 2022. IV estimation of spatial dynamic panels with interactive effects: large sample theory and an application on bank attitude towards risk. The Econometrics Journal 26 (2), 124–146.

Das, D., Kelejian, H. H., Prucha, I. R., 2003. Finite sample properties of estimators of spatial autoregressive models with autoregressive disturbances. Papers in Regional Science 82, 1–26.

Donald, S. G., Imbens, G. W., Newey, W. K., 2003. Empirical likelihood estimation and consistent tests with conditional moment restrictions. Journal of Econometrics 117 (1), 55–93.

Heckman, J. J., 1981. Heterogeneity and state dependence. In: Studies in labor markets. University of Chicago Press, pp. 91–140.

Higgins, A., 2022. Panel Data Models with Interactive Fixed Effects and Relatively Small T. Working Paper.

Kelejian, H. H., Prucha, I. R., 1998. A generalized spatial two-stage least squares procedure for estimating a spatial autoregressive model with autoregressive disturbances. The Journal of Real Estate Finance and Eonomics 17, 99–121.

Klein, J. L., 1997. Statistical visions in time: a history of time series analysis, 1662-1938. Cambridge University Press.

Kuersteiner, G. M., Prucha, I. R., 2020. Dynamic spatial panel models: Networks, common shocks, and sequential exogeneity. Econometrica 88 (5), 2109–2146.

Lee, L.-F., 2002. Consistency and efficiency of least squares estimation for mixed regressive, spatial autoregressive models. Econometric Theory 18 (2), 252–277.

Lee, L.-F., 2003. Best spatial two-stage least squares estimators for a spatial autoregressive model with autoregressive disturbances. Econometric Reviews 22 (4), 307–335.

Lee, L.-F., 2004. Asymptotic distributions of quasi-maximum likelihood estimators for spatial autoregressive models. Econometrica 72 (6), 1899–1925.

Lee, L.-f., Yu, J., 2010a. Estimation of spatial autoregressive panel data models with fixed effects. Journal of Econometrics 154 (2), 165–185.

Lee, L.-F., Yu, J., 2010b. A spatial dynamic panel data model with both time and individual fixed effects. Econometric Theory 26 (2), 564–597.

Lee, L.-f., Yu, J., 2014. Efficient GMM estimation of spatial dynamic panel data models with fixed effects. Journal of Econometrics 180 (2), 174–197.

Lee, N., Moon, H. R., Weidner, M., 2012. Analysis of interactive fixed effects dynamic linear panel regression with measurement error. Economics Letters 117 (1), 239–242.

Li, L., Yang, Z., 2023. Spatial dynamic panel data models with interactive fixed effects: M-estimation and inference with small T. Working paper.

Lin, X., Lee, L.-f., 2010. GMM estimation of spatial autoregressive models with unknown heteroskedasticity. Journal of Econometrics 157 (1), 34–52.

McCullagh, P., Tibshirani, R., 1990. A simple method for the adjustment of profile likelihoods. Journal of the Royal Statistical Society. Series B (Methodological) 52 (2), 325–344.

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

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.

Shi, W., Lee, L.-F., 2017. Spatial dynamic panel data models with interactive fixed effects. Journal of Econometrics 197 (2), 323–347.

Whittle, P., 1954. On stationary processes in the plane. Biometrika, 434–449.