UNIT ROOT INFERENCE IN GENERALLY TRENDING AND CROSS-CORRELATED FIXED-T PANELS*

Donald Robertson
University of Cambridge

Vasilis Sarafidis
Monash University

Joakim Westerlund†
Lund University
and
Centre for Financial Econometrics
Deakin University

May 14, 2016

Abstract

This paper proposes a new panel unit root test based on the generalized method of moments approach for panels with a possibly small number of time periods, T, and a large number of cross-sectional units, N. In the model that we consider the deterministic trend function is essentially unrestricted and the errors obey a multi-factor structure that allows for rich forms of unobserved heterogeneity. In spite of these allowances, the GMM estimator considered is shown to be asymptotically unbiased, $\sqrt{N}$-consistent and asymptotically normal for all values of the autoregressive (AR) coefficient, ρ, including unity, making it a natural candidate for unit root inference. Results from our Monte Carlo study suggest that the asymptotic properties are borne out well in small samples. The implementation is illustrated by using a large sample of US banking institutions to test Gibrat's Law.

JEL Classification: C12; C13; C33; C36.

Keywords: Panel data; unit root test; unobserved heterogeneity; common factors; GMM.

1 Introduction

There is a voluminous literature on panel unit root tests. The main motivation for using such procedures is that by considering not one but N time series of length T the power of panel-based tests can increase considerably relative to that achievable using univariate tests. The

*The authors would like to thank Shakeeb Khan (Joint Editor), one Associate Editor, and two anonymous referees for many valuable comments and suggestions. Thank you also to the Knut and Alice Wallenberg Foundation for financial support through a Wallenberg Academy Fellowship, and the Jan Wallander and Tom Hedelius Foundation for financial support under research grant number P2014–0112:1.

†Corresponding author: Department of Economics, Lund University, Box 7082, 220 07 Lund, Sweden. Telephone: +46 46 222 8997. Fax: +46 46 222 4613. E-mail address: joakim.westerlund@nek.lu.se.

largest branch of the literature by far is that focusing on panels where both N and T are large (see Breitung and Pesaran, 2008, for an overview). A typical study assumes that $N, T \to \infty$ such that $N/T \to 0$. The main reason for this is the presence of cross-sectional heterogeneity, such as individual-specific effects, the consistent estimation of which requires $T \to \infty$. This induces an estimation error, which can be controlled, but typically only if $N/T \to 0$, for otherwise the bias is not eliminated as $N \to \infty$ (see Westerlund and Breitung, 2013, Section 5, for a detailed discussion). This requirement may put strain on the data. Indeed, as a large body of Monte Carlo evidence shows, while the large-N requirement is usually not a problem, the large-T requirement, and in particular the requirement that T must be larger than N, poses a real restriction (see, for example, De Wachter et al., 2007; Hlouskova and Wagner, 2006), to the extent that researchers might well consider discarding data in order to have N sufficiently small relative to T.[1] Moreover, in many panels, such as those frequently encountered in applied microeconometrics, T (N) is simply too small (large) for such discarding practices to make sense, although the unit root hypothesis is still of considerable interest (see Bond et al., 2005).

Discussions like the one in the previous paragraph have motivated researchers to look for inferential procedures that are suitable in fixed-T panels. Harris and Tzavalis (1999) were among the first. They proposed a fixed-T panel unit root test based on the bias-corrected ordinary least squares (OLS) estimator of the autoregressive (AR) coefficient, ρ. Many other tests have since then been proposed (see De Blander and Dhaene, 2012, and the references provided therein). The evidence reported so far suggests that in terms of small-sample performance, not requiring T to be large can be a great advantage (see, for example, Harris and Tzavalis, 1999; Hadri and Larsson, 2005; Hlouskova and Wagner, 2006). In fact, fixed-T tests often outperform large-T tests and do so for a wide range of values of T.

However, while much progress has been made, there are (at least) two important issues that have not received much attention. First, except for De Blander and Dhaene (2012), Harris and Tzavalis (1999, 2004), and Han and Phillips (2010), who consider the case of a linear trend, the fixed-T literature has not yet ventured outside the fixed effects environment. This is noteworthy because if one admits to the possibility that time series might be trending, then the probability that the panel of multiple time series exhibits at least some trending behavior will tend to one as $N \to \infty$, in which case fixed effects-only tests are rendered invalid. Second, unlike several large-T "second-generation" unit root tests that allow for cross-sectional dependence in the form of

[1] For example, while the Penn World Tables have $N \gg T$, by considering only the OECD countries, one can make $N \ll T$.

common factors (see Moon and Perron, 2004, Phillips and Sul, 2003, Bai and Ng, 2004, Pesaran, 2007, among others), as far as we are aware, there is presently no fixed-T unit root test that is able to accommodate common factors; that is, existing fixed-T tests are "first-generation" tests (Baltagi, 2008, Chapter 12). A noteworthy exception is the recent working paper by Karavias and Tzavalis (2014), in which the authors allow for spatial correlation, representing "weak" cross-sectional dependence.[2]

The purpose of the current paper is to address both of the above mentioned issues. Specifically, a second-generation approach to unit roots in fixed-T panels is proposed that allows for both cross-sectional dependence and generally trending behavior. This is accomplished by letting the data admit to a common factor structure, in which the unobserved factors are treated as unknown parameters to be estimated along with the other parameters of the model. This parametric treatment means that the factors are virtually unrestricted, apart from some mild regularity conditions. It also provides a way to control for (unobserved) deterministic trend terms, which in our model appear naturally as additional factors. In the terminology of Bai (2009), the model that we consider constitutes an "interactive effects" model. Interestingly, since the factors are estimated, the usual problem in empirical work of deciding on which deterministic terms to include does not arise. Hence, the approach is not only general, but is in this sense also remarkably simple. This is in stark contrast to the existing large-T literature, where the factors are incidental parameters, the number of which increases with T. Not only does this complicate estimation, but it also restricts quite substantially the types of factors that can be permitted.

The estimation is carried out by modifying the generalized method of moments (GMM) approach of Robertson and Sarafidis (2015). The new estimator is shown to have a number of desirable properties. First, it circumvents the usual incidental parameter bias problem, which arises because the number of factor loading parameters that has to be estimated increases with N. This is true both in the conventional fixed effects case and in the more general interactive effects model considered here. The reason is that we do not require estimates of the loadings themselves, but only estimation of their normalized second moment, whose dimension remains fixed as N grows. Second, the estimator supports asymptotically normal inference for all values of ρ, including unity, and the well-known identification problems that arise when ρ equals unity do not emerge (see, for example, Bun and Windmeijer, 2010, for a discussion of this issue). The limiting distribution of the GMM estimator considered here is therefore continuous as ρ passes through unity (see Han and Phillips, 2010, for a similar result). Finally, the estimator and the

[2]See Chudik et al. (2011) for a detailed treatment of the concepts of weak and strong cross-sectional dependence.

associated t-statistic for a unit root appear to have satisfactory small-sample properties.

The remainder of the paper is organized as follows. Section 2 presents the model and assumptions, which are used in Section 3 to derive the GMM estimator and its asymptotic distribution. The empirical usefulness of the new estimator is evaluated using both simulated and real data in Sections 4 and 5, respectively. Section 6 concludes.

2 Model

Consider the panel data variable $y_{i,t}$, which is observed over $t = 0, 1, ..., T$ time series and $i = 1, ..., N$ cross-sectional units. The data generating process (DGP) of this variable is assumed to be given by

$$y_{i,t} = \rho y_{i,t-1} + u_{i,t}, \tag{1}$$

$$u_{i,t} = \boldsymbol{\lambda}_i' \mathbf{f}_t + \varepsilon_{i,t}, \tag{2}$$

for $t = 1, \dots T$, where $\rho \in \mathbb{R}$, $\mathbf{f}_t$ is an $r \times 1$ vector of common factors, $\boldsymbol{\lambda}_i$ is the associated vector of factor loadings, and $\varepsilon_{i,t}$ is an idiosyncratic error term. The following assumptions are assumed to hold throughout the paper.

Assumption 1. $\varepsilon_{i,t} \sim \text{iid}(0, \sigma_\varepsilon^2)$ and has finite moments up to fourth order.

Assumption 2. $\boldsymbol{\lambda}_i \sim \text{iid}(0, \boldsymbol{\Sigma}_\lambda)$ with finite moments up to fourth order and $\boldsymbol{\Sigma}_\lambda$ positive definite.

Assumption 3. $\mathbf{f}_t$ is non-stochastic such that $\|\mathbf{f}_t\| < \infty$.

Assumption 4. $y_{i,0}$ is iid and has finite moments up to fourth order.

Assumption 5. $E(\varepsilon_{i,t} | y_{i,0}, \dots, y_{i,t-1}, \boldsymbol{\lambda}_i') = 0$.

Assumptions 1–2 are mainly employed for simplicity and can be relaxed to allow for more general DGPs. For example, analogously to the existing dynamic panel GMM literature, arbitrary (unconditional) time series and cross-sectional heteroskedasticity in $\varepsilon_{i,t}$ can be allowed by using the standard sandwich formula of the covariance matrix of the GMM estimator.[3] The independence assumption across i can also be relaxed to allow for weak cross-sectional correlation. However, this would require replacing Assumption 1 with a "high-level" assumption like

[3]Note that $\varepsilon_{i,t}$ is not required to be conditionally homoskedastic.

Assumption C in Bai and Ng (2004), which we would like to avoid. Similarly, while the conditional moments of $\boldsymbol{\lambda}_i$ can be heterogeneously distributed, we avoid such generalizations in order to reduce the notational burden in the paper. The requirement that $\boldsymbol{\Sigma}_\lambda$ should be positive definite, which is standard in the common factor literature (see Bai and Ng, 2008, for a recent overview), is, on the other hand, key and cannot be dispensed with.

According to Assumption 3, $\mathbf{f}_t$ is treated as a fixed parameter vector to be estimated along with the remaining parameters of the model. Such parametric treatments are common in the related large-T literature; however, since in that literature T grows, $\mathbf{f}_t$ has to be restricted. For example, it is standard practice to assume that $\mathbf{f}_t$ has some stationary distribution with mean zero and finite moments up to fourth order (see, for example, Phillips and Sul, 2003), which of course rules out the presence of deterministic constant and trend terms. Assumption 3 does not impose any distributional assumptions of this kind and is in this sense very general.

Assumption 4 basically imposes similar conditions for the initial observation as those that apply to the error components of the model. However, as it will soon become clear, the functional form of $y_{i,0}$ remains unrestricted. For instance, one could set $y_{i,0} = \boldsymbol{\lambda}_i^{*\prime}\mathbf{f}_0 + \varepsilon_{i,0}^*$. In this case, $y_{i,0}$ could be thought of as the reduced form of $y_{i,t}$ at period $t = 0$, which is assumed to be observed. Thus, the values of $\boldsymbol{\lambda}_i^*$ and $\varepsilon_{i,0}^*$ are not necessarily identical to the values of $\boldsymbol{\lambda}_i$ and $\varepsilon_{i,0}$ that would arise had $y_{i,0}$ been assumed to follow (1).

Assumption 5 implies that $\varepsilon_{i,t}$ is serially uncorrelated. This can also be relaxed in a straightforward way. In particular, a moving average (MA) process of order q can be accommodated by replacing Assumption 5 with $E(\varepsilon_{i,t}|y_{i,0}, \ldots, y_{i,t-1-q}, \boldsymbol{\lambda}_i') = 0$, for $t \geq 1 - q$. An autoregressive process (AR) in $\varepsilon_{i,t}$ can be accommodated by simply augmenting the right-hand side of (1) with more lags of $y_{i,t}$. This is discussed in detail in Remark 3, and then again in Section 5, where we implement our approach in the presence of MA errors. In addition, Assumption 5 implies no correlation between the idiosyncratic error and the factor loadings. This assumption is standard in dynamic panels and it could be relaxed, although the computational task becomes far more complex in this case.[4] The assumption that $\varepsilon_{i,t}$ is uncorrelated with $\boldsymbol{\lambda}_i$ is also standard in the large T-literature (see, for example, Moon and Perron, 2004; Bai and Ng, 2004; Pesaran, 2007).

The above DGP can be seen as a version of the type of components model commonly used in the large-T (panel) unit root literature, in which the deterministic and stochastic elements of

[4]For instance, in the simple one-way error components model with $u_{i,t} = \eta_i + \varepsilon_{i,t}$, the GMM estimator proposed by Arellano and Bond (1991) employs moment conditions of the form $E(y_{i,s}\Delta u_{i,t}) = 0$ for $s < t - 1$. It is easy to see that this generally requires $E(\varepsilon_{i,t}|\eta_i) = 0$, because otherwise $y_{i,s}$ will be correlated with the transformed error, since they are both functions of η_i.

the model are separated explicitly.[5] In order to illustrate this point, suppose that

$$y_{i,t} = d_{i,t} + z_{i,t}, \tag{3}$$

$$z_{i,t} = \rho z_{i,t-1} + v_{i,t}, \tag{4}$$

$$v_{i,t} = \gamma_i' \mathbf{w}_t + \varepsilon_{i,t}, \tag{5}$$

where $d_{i,t}$ represents the deterministic component of $y_{i,t}$, and $\mathbf{w}_t$ is an $m \times 1$ vector of "genuine" stochastic common factors. Multiplying the first lag of expression (3) by ρ and subtracting from the original equation yields

$$y_{i,t} = \rho y_{i,t-1} + d_{i,t} - \rho d_{i,t-1} + \gamma_i' \mathbf{w}_t + \varepsilon_{i,t}, \tag{6}$$

which is equivalent to (1) and (2) with $\boldsymbol{\lambda}_i' \mathbf{f}_t = d_{i,t} - \rho d_{i,t-1} + \gamma_i' \mathbf{w}_t$. To appreciate what implications the formulation in (6) has for the DGP in (1) and (2) it is useful to consider a couple of cases. The two most common specifications of $d_{i,t}$ by far are $d_{i,t} = \eta_i$ (fixed effects), and $d_{i,t} = \eta_i + \beta_i t$ (incidental trends). In the first specification, $d_{i,t} - \rho d_{i,t-1} = \eta_i(1-\rho)$, whereas in the second, $d_{i,t} - \rho d_{i,t-1} = \eta_i(1-\rho) + \beta_i \rho + \beta_i(1-\rho)t$. One of the implications of (6) is therefore to make the order of the trend polynomial a function of ρ. In the trend case, for example, when $\rho = 1$, (6) corresponds to (1) and (2) with $r = m+1$, $\boldsymbol{\lambda}_i = (\beta_i, \gamma_i')'$ and $\mathbf{f}_t = (1, \mathbf{w}_t')'$. On the other hand, when $|\rho| < 1$, the same model corresponds to (1) and (2) with $r = m+2$, $\boldsymbol{\lambda}_i = (\eta_i(1-\rho) + \beta_i \rho, \beta_i(1-\rho), \gamma_i')'$ and $\mathbf{f}_t = (1, t, \mathbf{w}_t')'$. Hence, when cast in terms of (1) and (2), the fact that in equation (6) the order of the trend polynomial is reduced when $\rho = 1$ implies that one of the elements of $\boldsymbol{\lambda}_i$ drops out. While this does not fit well with our assumption that $\boldsymbol{\Sigma}_\lambda$ is positive definite, in practice there is an easy way out of this, which is to make the estimation procedure conditional on r (as it is commonly done in the common factor literature; see Bai and Ng, 2008). This treatment stands in sharp contrast with the standard dynamic panel data literature (see the detailed discussion in Section 3), as well as with the large-T panel unit root literature, in which the estimation is carried out conditional on knowing the whole of $d_{i,t}$. The fact that the GMM approach developed here does not require $d_{i,t}$ to be pre-specified is a great advantage in practice, as the researcher is spared the problem of having to decide on which deterministic components to include. If r is unknown, as is usually the case in practice, then an information criterion may be used to obtain a consistent estimator. In Section 3 we elaborate on this point.

[5] See Schmidt and Phillips (1992) for a discussion in the pure time series context.

It is important to note that our approach is not limited to incidental intercept and trend terms only; on the contrary, it can accommodate virtually any trend function that is linear in the parameters, including polynomial trend functions, trigonometric functions and models of discrete and smooth structural shifts. This is again in stark contrast to the bulk of the existing large-T literature, which has not yet ventured outside the linear trend environment (see Westerlund, 2015, for an exception). The reason for this is the presence of incidental parameter bias, the analytical complexity of which increases very quickly with both the number and the non-linearity of the trend terms. The implication is that once outside the linear trend environment, bias-correction is not really an attractive option. Against this background, the asymptotic unbiasedness of the GMM estimator developed here is clearly a great advantage, as it enables testing in situations that were previously not possible.

3 The GMM-based test approach

3.1 Main results

The DGP in (1) and (2) can be written in stacked vector form as

$$\mathbf{y}_i = \rho \mathbf{y}_{i,-1} + \mathbf{F}\boldsymbol{\lambda}_i + \boldsymbol{\varepsilon}_i = \rho \mathbf{y}_{i,-1} + (\mathbf{I}_T \otimes \boldsymbol{\lambda}_i')\mathbf{f} + \boldsymbol{\varepsilon}_i, \tag{7}$$

where $\mathbf{y}_i = (y_{i,1}, \ldots, y_{i,T})'$, $\mathbf{y}_{i,-1} = (y_{i,0}, \ldots, y_{i,T-1})'$ and $\boldsymbol{\varepsilon}_i = (\varepsilon_{i,1}, \ldots, \varepsilon_{i,T})'$ are $T \times 1$ vectors, $\mathbf{F} = (\mathbf{f}_1, \ldots, \mathbf{f}_T)'$ is a $T \times r$ matrix, and $\mathbf{f} = \text{vec}(\mathbf{F}')$.

Let $\mathbf{S}_t$ be the $M_t \times T$ selector matrix of zeroes and ones that for each time period t picks out the M_t entries of $\mathbf{y}_{i,-1}$ that are uncorrelated with $\varepsilon_{i,t}$, i.e. $\mathbf{S}_t$ is such that $E[(\mathbf{S}_t \mathbf{y}_{i,-1})\varepsilon_{i,t}] = \mathbf{0}_t$. Under Assumption 5 we have $\mathbf{S}_t = (\mathbf{I}_t, \mathbf{0}_{t\times(T-t)})$, a $t \times T$ matrix; therefore, at time t the vector of valid instruments is simply given by $y_{i,0}, \ldots, y_{i,t-1}$. Define $\mathbf{S} = \text{diag}(\mathbf{S}_1, ..., \mathbf{S}_T)$, a $M \times T^2$ block diagonal matrix, where $M = \sum_{t=1}^{T} M_t = T(T+1)/2$ in the present case. The $M \times T$ matrix of instruments can now be written as

$$\mathbf{Z}_i' = \mathbf{S}(\mathbf{I}_T \otimes \mathbf{y}_{i,-1}). \tag{8}$$

It is easy to verify that $E(\mathbf{Z}_i'\boldsymbol{\varepsilon}_i) = \mathbf{0}_{M\times 1}$. As a result, multiplying (7) by $\mathbf{Z}_i'$ and taking expectations yields

$$\mathbf{m} = \rho \mathbf{m}_1 + \mathbf{S}(\mathbf{I}_T \otimes \mathbf{G})\mathbf{f}, \tag{9}$$

where $\mathbf{m} = E(\mathbf{Z}_i'\mathbf{y}_i)$, $\mathbf{m}_1 = E(\mathbf{Z}_i'\mathbf{y}_{i,-1})$ and $\mathbf{G} = E(\mathbf{y}_{i,-1}\boldsymbol{\lambda}_i')$, a $T \times r$ matrix.[6]

[6]The existence of $\mathbf{G}$ is implied by Assumptions 1–4.

The dimension of $\mathbf{G}$ can be reduced substantially by taking into account the AR structure of $y_{i,t}$. The latter implies that $\mathbf{y}_{i,-1}$ can be written as follows:

$$\mathbf{y}_{i,-1} = \mathbf{\Gamma}\mathbf{e}_1 y_{i,0} + \mathbf{\Gamma}\widetilde{\mathbf{F}}_{-1}\boldsymbol{\lambda}_i + \mathbf{\Gamma}\widetilde{\boldsymbol{\varepsilon}}_{i,-1}, \tag{10}$$

where $\mathbf{e}_t = (0, ..., 0, 1, 0, ..., 0)'$ is a $T \times 1$ vector of zeros, except for a one sitting in position t, $\widetilde{\mathbf{F}}_{-1} = (\mathbf{0}_{r\times 1}, \mathbf{f}_1, ..., \mathbf{f}_{T-1})'$ and $\widetilde{\boldsymbol{\varepsilon}}_{i,-1} = (0, \varepsilon_{i,1}, ..., \varepsilon_{i,T-1})'$. Also,

$$\mathbf{\Gamma} = (\mathbf{I}_T - \rho\mathbf{L})^{-1} = \begin{bmatrix} 1 & 0 & \dots & 0 & 0 \\ \rho & 1 & \dots & 0 & 0 \\ \rho^2 & \rho & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & 0 \\ \rho^{T-1} & \rho^{T-2} & & \rho & 1 \end{bmatrix}; \quad \mathbf{L} = \begin{bmatrix} 0 & 0 & \dots & 0 & 0 \\ 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & 0 \\ 0 & 0 & & 1 & 0 \end{bmatrix},$$

both being $T \times T$ matrices. Using the expression in (10) to substitute for $\mathbf{y}_{i,-1}$ in $E(\mathbf{y}_{i,-1}\boldsymbol{\lambda}_i')$ yields

$$E(\mathbf{y}_{i,-1}\boldsymbol{\lambda}_i') = E[(\mathbf{\Gamma}\mathbf{e}_1 y_{i,0} + \mathbf{\Gamma}\widetilde{\mathbf{F}}_{-1}\boldsymbol{\lambda}_i + \mathbf{\Gamma}\widetilde{\boldsymbol{\varepsilon}}_{i,-1})\boldsymbol{\lambda}_i'] = \mathbf{\Gamma}\mathbf{e}_1\mathbf{g}_0' + \mathbf{\Gamma}\widetilde{\mathbf{F}}_{-1}\mathbf{\Sigma}_\lambda, \tag{11}$$

where $\mathbf{g}_s = E(y_{i,s}\boldsymbol{\lambda}_i)$. This means that given ρ, $\mathbf{f}$ and $\mathbf{\Sigma}_\lambda$, the only unknown parameters in $\mathbf{G}$ are those contained in the $r \times 1$ vector $\mathbf{g}_0$. Thus, making use of (11) leads to a more parsimonious formulation of the model when compared to (9). As a result, noting that $\mathbf{f} = (\mathbf{I}_T \otimes \mathbf{F}')\mathbf{e}$, where $\mathbf{e} = \text{vec}(\mathbf{I}_T)$, the proposed vector of moment conditions is given by

$$\mathbf{m} - \rho\mathbf{m}_1 - \mathbf{S}[\mathbf{I}_T \otimes (\mathbf{\Gamma}\mathbf{e}_1\mathbf{g}_0'\mathbf{F}' + \mathbf{\Gamma}\widetilde{\mathbf{F}}_{-1}\mathbf{\Sigma}_\lambda\mathbf{F}')]\mathbf{e} = \mathbf{0}_{M\times 1}. \tag{12}$$

Our model is not identified as it stands because the vector of moment conditions in (12) contains products of unknown parameters. In particular, $\mathbf{g}_0'\mathbf{F}'$ and $\widetilde{\mathbf{F}}_{-1}\mathbf{\Sigma}_\lambda\mathbf{F}'$ are not separately identifiable without normalizing restrictions. This is the same problem encountered in the common factor literature, which is due to the fact that $\boldsymbol{\lambda}_i'\mathbf{f}_t$ is observationally equivalent to $\boldsymbol{\lambda}_i^{*\prime}\mathbf{f}_t^*$, where $\boldsymbol{\lambda}_i^* = \mathbf{H}^{-1}\boldsymbol{\lambda}_i$ and $\mathbf{f}_t^* = \mathbf{H}'\mathbf{f}_t$ for some $r \times r$ invertible matrix $\mathbf{H}$ (see Bai and Ng, 2008, for a discussion). Because of this, in what follows we set, with a slight abuse of notation, $\widetilde{\mathbf{F}}_{-1} = \widetilde{\mathbf{F}}_{-1}\mathbf{\Sigma}_\lambda^{1/2}$, $\mathbf{F} = \mathbf{F}\mathbf{\Sigma}_\lambda^{1/2\prime}$ and $\mathbf{g}_0 = \mathbf{\Sigma}_\lambda^{-1/2}\mathbf{g}_0$. Hence, using $\mathbf{f}$ to denote the vectorized version of the normalized $\mathbf{F}'$, the vector of parameters of interest is given by $\boldsymbol{\theta} = (\rho, \mathbf{f}', \mathbf{g}_0')' = (\theta_1, \boldsymbol{\theta}_2', \boldsymbol{\theta}_3')'$, which is of order $\dim(\boldsymbol{\theta}) \times 1$, where $\dim(\boldsymbol{\theta}) = 1 + (T+1)r$. The proposed GMM estimator of this parameter vector based on the above moment conditions is given by

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}\in\Theta} Q_N(\boldsymbol{\theta}),$$

where Θ is a compact subset of $\mathbb{R}^{\dim(\boldsymbol{\theta})}$ and

$$\begin{aligned} Q_N(\boldsymbol{\theta}) &= \overline{\mathbf{h}}_N(\boldsymbol{\theta})'\mathbf{W}_N\overline{\mathbf{h}}_N(\boldsymbol{\theta}), \\ \overline{\mathbf{h}}_N(\boldsymbol{\theta}) &= \frac{1}{N}\sum_{i=1}^{N}\mathbf{h}_i(\boldsymbol{\theta}), \\ \mathbf{h}_i(\boldsymbol{\theta}) &= \mathbf{Z}_i'\mathbf{y}_i - \rho\mathbf{Z}_i'\mathbf{y}_{i,-1} - \mathbf{S}[\mathbf{I}_T \otimes (\boldsymbol{\Gamma}\mathbf{e}_1\mathbf{g}_0'\mathbf{F}' + \boldsymbol{\Gamma}\widetilde{\mathbf{F}}_{-1}\mathbf{F}')]\mathbf{e}, \end{aligned}$$

where $\mathbf{W}_N$ is an $M \times M$ weighting matrix.

Remark 1. Ignoring the structure of $\mathbf{G}$ simplifies the estimation burden considerably because, letting $\mathbf{g} = \text{vec}(\mathbf{G})$, (9) becomes $\mathbf{m} = \rho\mathbf{m}_1 + \mathbf{S}(\mathbf{F} \otimes \mathbf{I}_T)\mathbf{g}$. That is, for a given value of $\mathbf{F}$, the model is linear and so it can be estimated using OLS. By contrast, the proposed GMM estimator exploits the non-linear restrictions in $\mathbf{G}$ and is therefore non-linear. This extra complication is, however, well worthwhile, in the sense that ignoring the non-linear restrictions can lead to a substantial loss of efficiency (see, for example, Wansbeek and Bekker, 1996; Robertson and Sarafidis, 2015).

Remark 2. The moment conditions in (12) can be modified to allow for MA errors, or a more general AR structure for $y_{i,t}$. Suppose first that $\varepsilon_{i,t}$ follows an MA(q) process. This case can be accommodated by setting $\mathbf{Z}_i' = \mathbf{S}(\mathbf{I}_T \otimes \mathbf{y}_{i,-(q+1)})$, where $\mathbf{y}_{i,-(q+1)} = (y_{i,0}, ..., y_{i,T-1-q})'$ and the t-th diagonal element of $\mathbf{S}$ is given by $\mathbf{S}_t = (\mathbf{I}_t, \mathbf{0}_{t\times(T-t-q)})$, which has dimension $t \times (T-q)$. In Section 5 the estimator is implemented with $q \in \{1, 2\}$.

Consider next the case when $\varepsilon_{i,t}$ is serially uncorrelated, but $y_{i,t}$ follows an AR(2) process;

$$y_{i,t} = \rho_1 y_{i,t-1} + \rho_2 y_{i,t-2} + u_{i,t},$$

where we assume for notational simplicity that $y_{i,0}$ and $y_{i,-1}$ are observed. This model can be written in vector form as

$$\mathbf{y}_i = \rho_1\mathbf{y}_{i,-1} + \rho_2\mathbf{y}_{i,-2} + (\mathbf{I}_T \otimes \boldsymbol{\lambda}_i')\mathbf{f} + \boldsymbol{\varepsilon}_i, \tag{13}$$

where $\mathbf{y}_{i,-2} = (y_{i,-1}, y_{i,0}, ..., y_{i,T-2})'$ and the remaining vectors are as before. In this case the vector of instruments is given by $\mathbf{y}_{i,-1}^{+} = (y_{i,-1}, y_{i,0}, \ldots, y_{i,T-1})'$, which has the following DGP:

$$\mathbf{y}_{i,-1}^{+} = \boldsymbol{\Gamma}\mathbf{e}_2 y_{i,0} + \boldsymbol{\Pi}\mathbf{e}_1 y_{i,-1} + \boldsymbol{\Gamma}\widetilde{\mathbf{F}}_{-1}^{+}\boldsymbol{\lambda}_i + \boldsymbol{\Gamma}\widetilde{\boldsymbol{\varepsilon}}_{i,-1}^{+}, \tag{14}$$

where $\boldsymbol{\Gamma} = (\mathbf{I}_T - \rho_1\mathbf{L} - \rho_2\mathbf{L}^2)^{-1}$, $\widetilde{\mathbf{F}}_{-1}^{+} = (\mathbf{0}_{r\times 2}, \mathbf{f}_1, ..., \mathbf{f}_{T-1})'$, $\widetilde{\boldsymbol{\varepsilon}}_{i,-1}^{+} = (\mathbf{0}_{1\times 2}, \varepsilon_{i,1}, ..., \varepsilon_{i,T-1})'$ and $\boldsymbol{\Pi}$ is a $T \times T$ matrix that is a function of ρ_1, ρ_2 and T. In this case, $\mathbf{Z}_i' = \mathbf{S}(\mathbf{I}_T \otimes \mathbf{y}_{i,-1}^{+})$ with

$\mathbf{S}_t = (\mathbf{I}_{t+1}, \mathbf{0}_{(t+1)\times(T-t)})$, a $(t+1)\times(T+1)$ matrix. Thus, pre-multiplying equation (13) by $\mathbf{Z}_i'$, taking expectations (with the normalization enforced) and rearranging yields

$$\mathbf{m} - \rho_1\mathbf{m}_1 - \rho_2\mathbf{m}_2 - \mathbf{S}[\mathbf{I}_T \otimes (\mathbf{\Gamma e}_2\mathbf{g}_0'\mathbf{F}' + \mathbf{\Pi e}_1\mathbf{g}_{-1}'\mathbf{F}' + \mathbf{\Gamma}\widetilde{\mathbf{F}}_{-1}^{+}\mathbf{F}')]\mathbf{e} = \mathbf{0}_{M\times 1}, \tag{15}$$

where $\mathbf{m}_2 = E(\mathbf{Z}_i'\mathbf{y}_{i,-2})$.

Let

$$\mathbf{\Delta} = E[\nabla_{\boldsymbol{\theta}}\mathbf{h}_i(\boldsymbol{\theta}_0)], \quad \mathbf{\Omega} = E[\mathbf{h}_i(\boldsymbol{\theta}_0)\mathbf{h}_i(\boldsymbol{\theta}_0)'],$$

where $\boldsymbol{\theta}_0$ denotes the true value of $\boldsymbol{\theta}$. The following assumptions are sufficient to establish the asymptotic properties of our estimator:

Assumption 6.

(a) $\mathbf{W}_N \xrightarrow{p} \mathbf{W}$ as $N \to \infty$, where $\mathbf{W}$ is a positive definite matrix;

(b) $\boldsymbol{\theta}_0$ belongs to the interior of Θ;

(c) $E[\mathbf{h}_i(\boldsymbol{\theta})] = \mathbf{0}_{M\times 1}$ if and only if $\boldsymbol{\theta} = \boldsymbol{\theta}_0$;

(d) $\mathbf{\Delta}$ and $\mathbf{\Omega}$ exist and both are full rank matrices.

Remark 3. One implication of Assumption 6 (d) is that M, the number of moment conditions, is at least as large as the number of parameters in $\boldsymbol{\theta}$. Also, while Assumptions 1–6 (a)–(b) could be sufficient to ensure identification in the standard fixed effects case, Assumption 6 (c) and (d) are required to ensure identification and regular asymptotics in the more general interactive effects model considered here.

Theorem 1. *Under Assumptions 1–6, as $N \to \infty$,*

$$\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} N(\mathbf{0}_{\dim(\boldsymbol{\theta})\times 1}, \mathbf{\Sigma}_{\hat{\boldsymbol{\theta}}}),$$

where $\mathbf{\Sigma}_{\hat{\boldsymbol{\theta}}} = (\mathbf{\Delta}'\mathbf{W}\mathbf{\Delta})^{-1}(\mathbf{\Delta}'\mathbf{W}'\mathbf{\Omega}\mathbf{W}\mathbf{\Delta})(\mathbf{\Delta}'\mathbf{W}\mathbf{\Delta})^{-1}$.

Proof. See Appendix.

An analytical expression for $\boldsymbol{\Delta}$ is given in Appendix. If $\mathbf{W} = \boldsymbol{\Omega}^{-1}$, the covariance matrix of $\hat{\boldsymbol{\theta}}$ reduces to $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}} = (\boldsymbol{\Delta}'\boldsymbol{\Omega}^{-1}\boldsymbol{\Delta})^{-1}$, which is optimal. The same holds if $\boldsymbol{\Omega}^{-1}$ is replaced with a consistent estimate. This can be achieved by setting

$$\hat{\boldsymbol{\Omega}} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{h}_i(\hat{\boldsymbol{\theta}}^1)\mathbf{h}_i(\hat{\boldsymbol{\theta}}^1)',$$

where $\hat{\boldsymbol{\theta}}^1$ denotes any first-stage consistent GMM estimate of $\boldsymbol{\theta}$, such as the one obtained by setting $\mathbf{W} = \mathbf{I}_M$. In agreement with the jargon in the previous literature, the estimator based on $\mathbf{W} = \hat{\boldsymbol{\Omega}}^{-1}$ ($\mathbf{W} = \mathbf{I}_M$) is henceforth referred to as the "two-step" ("one-step") GMM estimator. Notice that if interest lies only in testing the unit root null hypothesis, one can compute $\hat{\boldsymbol{\Omega}}$ by setting the first entry in $\hat{\boldsymbol{\theta}}^1$ within $\mathbf{h}_i(\hat{\boldsymbol{\theta}}^1)$ equal to unity. Alternatively, one can leave the first entry in $\hat{\boldsymbol{\theta}}^1$ unrestricted in the estimation and impose the unit restriction only when making inference. This is the practice we have followed in our simulations.

Remark 4. According to Theorem 1 there is no asymptotic bias despite the generality of the DGP considered, that is, $(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ is centered at zero even when scaled by $\sqrt{N}$. The reason is that the GMM approach developed here does not require estimating $\boldsymbol{\lambda}_1, ..., \boldsymbol{\lambda}_N$, which implies that the number of parameters that needs to be estimated remains fixed as N grows large (see Bai, 2013, for a similar approach based on maximum likelihood).

Remark 5. Theorem 1 holds for all values of ρ_0, and in this sense it presents a unified asymptotic result for the GMM estimator (see also the discussion provided in Section 3.2). This is in contrast with the existing literature, in which the asymptotic distribution of dynamic panel estimators depends critically on whether $|\rho_0| < 1$, $\rho_0 = 1$ or $\rho_0 > 1$. In fact, the only exceptions known to us are the GMM estimators of Han and Phillips (2010), and Kruiniger (2007, 2009, 2013), which have limit distributions that are continuous for $\rho_0 \in (-1, 1]$, but not for $\rho_0 > 1$.

Let $\hat{\boldsymbol{\theta}} = (\hat{\rho}, \hat{\mathbf{f}}', \hat{\mathbf{g}}_0')' = (\hat{\theta}_1, \hat{\boldsymbol{\theta}}_2', \hat{\boldsymbol{\theta}}_3')'$, and denote by $\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\theta}}}$ the associated GMM estimator of $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}}$. The GMM-based t-statistic for testing $H_0 : \rho_0 = c$ is given by

$$t_{\hat{\rho}}(c) = \frac{\sqrt{N}(\hat{\rho} - c)}{\hat{\sigma}_{\hat{\rho}}}.$$

where $\hat{\sigma}^2_{\hat{\rho}}$ is the first diagonal element of $\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\theta}}}$. Since $\hat{\sigma}^2_{\hat{\rho}}$ is consistent for $\sigma^2_{\hat{\rho}}$, under the conditions of Theorem 1, we have

$$t_{\hat{\rho}}(\rho_0) = \frac{\sqrt{N}(\hat{\rho} - \rho_0)}{\hat{\sigma}_{\hat{\rho}}} \xrightarrow{d} N(0, 1) \tag{16}$$

as $N \to \infty$. The test is also consistent. This can be appreciated by noting that

$$t_{\hat{\rho}}(c) = \frac{\sqrt{N}(\hat{\rho} - \rho_0)}{\hat{\sigma}_{\hat{\rho}}} + \frac{\sqrt{N}(\rho_0 - c)}{\hat{\sigma}_{\hat{\rho}}}. \tag{17}$$

While the first term on the right-hand side converges to $N(0,1)$ by (16), the second is $O_p(\sqrt{N})$ whenever $\rho_0 \neq c$. The power of the test will therefore tend to one as $N \to \infty$.

Remark 6. The asymptotic distribution of most (if not all) unit root test statistics depends on the deterministic specification of the fitted test regression, which need not be equal to the true one. In time series, this implies that different deterministic specifications have their own critical values, whereas in panels, it implies that different specifications have their own mean and variance correction factors (see Westerlund and Breitung, 2013, Section 3). According to (16), the GMM-based t-statistic has a unique and practically very useful property in that, under the null hypothesis, it is asymptotically invariant to $\mathbf{f}_t$ and hence to any trend function that it may contain. Therefore, mean and variance correction factors that depend on a particular deterministic specification are not required with our approach.

So far the number of factors has been treated as known, which need not be the case in practice. A natural approach towards this end is to treat the estimation of r as a model selection problem, which can be handled using an information criterion. Let us denote by r_0 the true value of r, and let $\hat{\boldsymbol{\theta}}(r)$ be the estimator of $\hat{\boldsymbol{\theta}}$ based on r factors. The estimator of r_0 considered in the present paper is similar to those of Ahn et al. (2013), and Robertson and Sarafidis (2015), and is given by

$$\hat{r} = \arg \min_{r=0,\ldots,r_{max}} BIC(r) \tag{18}$$

with $r_0 \leq r_{max}$ and $BIC(r) = N \cdot Q_N(\hat{\boldsymbol{\theta}}(r)) - \ln(N) \cdot b(r)$, where $Q_N(\hat{\boldsymbol{\theta}}(r))$ is the value of the objective function evaluated at $\hat{\boldsymbol{\theta}}(r)$, which when multiplied by N is identically the value of the overidentifying restrictions J-statistic for the two-step estimator, and $b(r)$ is a penalty function that is strictly decreasing in r. As usual, $b(r)$ is not given but has to be set by the researcher. In our simulations (see Section 4), we set $b(r) = \mathrm{df}(r)/T^{0.3}$, where $\mathrm{df}(r) = M - \dim(\boldsymbol{\theta})$ is the degrees of freedom of the model based on r factors.

Theorem 2. *Under the conditions of Theorem 1, as* $N \to \infty$,

$$\hat{r} \xrightarrow{p} r_0.$$

Proof. The proof of this theorem follows directly from the results in Ahn et al. (2013), and Robertson and Sarafidis (2015). It is therefore omitted.

Theorem 2 says that asymptotically knowing $\hat{r}$ is as good as knowing r_0. This means that the result reported in Theorem 1 holds even if r_0 is replaced by $\hat{r}$. To see that this must be the case, consider

$$\begin{aligned} P(\sqrt{N}[\hat{\boldsymbol{\theta}}(\hat{r}) - \boldsymbol{\theta}_0] \leq \delta) &= P(\sqrt{N}[\hat{\boldsymbol{\theta}}(\hat{r}) - \boldsymbol{\theta}_0] \leq \delta | \hat{r} = r_0) P(\hat{r} = r_0) \\ &+ P(\sqrt{N}[\hat{\boldsymbol{\theta}}(\hat{r}) - \boldsymbol{\theta}_0] \leq \delta | \hat{r} \neq r_0) P(\hat{r} \neq r_0). \end{aligned}$$

Because $P(\hat{r} = r_0) \to 1$ and $P(\hat{r} \neq r_0) \to 0$ by Theorem 2, while the first term on the right-hand side converges to $\lim_{N\to\infty} P(\sqrt{N}[\hat{\boldsymbol{\theta}}(\hat{r}) - \boldsymbol{\theta}_0] \leq \delta | \hat{r} = r_0) = \lim_{N\to\infty} P(\sqrt{N}[\hat{\boldsymbol{\theta}}(r_0) - \boldsymbol{\theta}_0] \leq \delta)$, the second term converges to zero. It follows that

$$|P(\sqrt{N}[\hat{\boldsymbol{\theta}}(\hat{r}) - \boldsymbol{\theta}_0] \leq \delta) - P(\sqrt{N}[\hat{\boldsymbol{\theta}}(r_0) - \boldsymbol{\theta}_0] \leq \delta)| \to 0. \tag{19}$$

The fact that the asymptotic distribution of $\sqrt{N}[\hat{\boldsymbol{\theta}}(\hat{r}) - \boldsymbol{\theta}_0]$ is equal to that of $\sqrt{N}[\hat{\boldsymbol{\theta}}(r_0) - \boldsymbol{\theta}_0]$ is worthy of some discussion. As alluded to in Section 2, and as we discuss in detail in Section 3.2, the consistency of $\hat{r}$ is actually a key ingredient in our estimation strategy, especially when the DGP is governed by (6). This is so because when $\rho_0 = 1$ some of the factors in $\mathbf{f}_t$ become redundant, thus violating the requirement that $\boldsymbol{\Sigma}_\lambda$ should be positive definite. Thus, in agreement with the existing common factor literature (see, for example, Bai and Ng, 2008), the number of common factors cannot be overspecified. It is therefore useful to consider briefly the idea behind the proof of Theorem 2 when $r > r_0$. We need to show that the probability of selecting an overspecified model is negligible, that is, we need to show that

$$P[BIC(r_0) - BIC(r) > 0] \to 0 \tag{20}$$

as $N \to \infty$. Direct substitution using the definition of $BIC(r)$ yields

$$\begin{aligned} &P[BIC(r_0) - BIC(r) > 0] \\ &= P[N \cdot Q_N(\hat{\boldsymbol{\theta}}(r_0)) - N \cdot Q_N(\hat{\boldsymbol{\theta}}(r)) + \ln(N)(b(r) - b(r_0)) > 0] \\ &\leq P[N \cdot Q_N(\hat{\boldsymbol{\theta}}(r_0)) + \ln(N)(b(r) - b(r_0)) > 0], \end{aligned}$$

where the inequality is due to the fact that $Q_N(\hat{\boldsymbol{\theta}}(r)) \geq 0$. Further use of Theorem 1 reveals that $N \cdot Q_N(\hat{\boldsymbol{\theta}}(r_0)) = O_p(1)$. Hence, since $b(r) < b(r_0)$ for $r > r_0$, we have $\ln(N)(b(r) - b(r_0)) \to -\infty$ as $N \to \infty$, and therefore (20) is satisfied. It is important to note that this result only makes

use of Theorem 1 and the fact that $Q_N(\hat{\boldsymbol{\theta}}(r)) \geq 0$. There is therefore no need to evaluate the asymptotic distribution of $\sqrt{N}[\hat{\boldsymbol{\theta}}(r) - \boldsymbol{\theta}_0]$, which would be difficult, if not impossible, since $\hat{\boldsymbol{\theta}}(r)$ comes from an overspecified model. Hence, only by making the estimation of $\boldsymbol{\theta}_0$ conditional on $\hat{r}$ can we asymptotically eliminate the risk of having redundant factors.

Remark 7. Our testing procedure is valid even if the DGP is different across individuals. Consider as an example the case when there are two clusters of cross-sectional units, C_1 and C_2. Let us denote by N_ℓ the cardinality of C_ℓ, such that $N_1 + N_2 = N$, and suppose that $y_{i,t} = \rho y_{i,t-1} + \eta_i(1-\rho) + \varepsilon_{i,t}$ for $i \in C_1$, and $y_{i,t} = \rho y_{i,t-1} + \eta_i(1-\rho) + \beta_i\rho + \beta_i(1-\rho)t + \varepsilon_{i,t}$ for $i \in C_2$. Hence, if $|\rho| < 1$, then the number of cross-sectional units following a stationary process with a non-zero mean (trend) is given by N_1 (N_2). Our procedure is valid because the vector of moment conditions remains exactly as in (12), except that the value of $\boldsymbol{\Sigma}_\lambda$ will be different compared to the case where $N_1 = 0$. In the example considered here the last diagonal entry of $\boldsymbol{\Sigma}_\lambda$ will be equal to $(1-\pi)E(\beta_i^2)(1-\rho_0)^2$, where $\pi = N_1/N$. In Section 4 we use Monte Carlo simulations to investigate the effects of this type of clustering in small samples.

3.2 Discussion

Under our assumptions, the model in (1) is identified for all values of $\rho_0 \in \mathbb{R}$, including $\rho_0 = 1$. This is an important result that merits some discussion, as it is in contrast to known results for many existing GMM estimators for dynamic panel data models, such as the ones of Anderson and Hsiao (1981), and Arellano and Bond (1991), which can fail to identify ρ_0 when $\rho_0 = 1$ (see, for example, Bond et al., 2005, for a discussion). The reason for this is that here unobserved heterogeneity is absorbed by the factors, the dimension of which can be consistently estimated. That is, our GMM estimator is obtained conditional on $\hat{r}$, which is again as good as knowing r_0 itself. To fix ideas, suppose that (6) holds with $T = 2$, $d_{i,t} = \eta_i$ and $\boldsymbol{\gamma}_i = \mathbf{0}_{m\times 1}$, which is (1) and (2) with $\mathbf{f}_t = 1$ and $\boldsymbol{\lambda}_i = \eta_i(1-\rho_0)$. As a result, there is no "genuine" factor structure. It follows that

$$y_{i,t} = \eta_i(1-\rho_0) + \rho_0 y_{i,t-1} + \varepsilon_{i,t}. \tag{21}$$

The Arellano and Bond (1991) "first-differenced" GMM estimator, hereafter DIF, of this model involves taking first-differences in order to eliminate $\eta_i(1-\rho_0)$;

$$\Delta y_{i,2} = \rho_0 \Delta y_{i,1} + \Delta\varepsilon_{i,2}. \tag{22}$$

When $\rho_0 = 1$, however, $\eta_i(1-\rho_0) = 0$ and therefore first-differencing results in a lack of identification. Indeed, in the $T = 2$ case considered here, there is a single valid moment condition in (22), which is given by $E(y_{i,0}\Delta\varepsilon_{i,2}) = 0$. Identification of ρ_0 requires $E(y_{i,0}\Delta y_{i,1}) \neq 0$. But if $\rho_0 = 1$, then

$$E(y_{i,0}\Delta y_{i,1}) = E(y_{i,0}[(\rho_0 - 1)y_{i,0} + \eta_i(1-\rho_0) + \varepsilon_{i,1}]) = E(y_{i,0}\varepsilon_{i,1}) = 0,$$

and so ρ_0 is unidentified. However, notice that under $\rho_0 = 1$, the conventional pooled OLS estimator of (21) (without common factors) is still consistent, as it is based on a valid restriction ($\eta_i(1-\rho_0) = 0$), although it becomes inconsistent for any other value of $|\rho_0| < 1$.

The discussion above suggests that it is desirable to have an estimator that is consistent regardless of the value taken by ρ_0. To see how such an estimator may be devised, let us consider the proposed GMM estimator, but without exploiting the restrictions for $\mathbf{G}$ described in (11), which are irrelevant for this argument. The appropriate moment conditions have the following form:

$$\begin{bmatrix} m_{01} \\ m_{02} \\ m_{12} \end{bmatrix} - \rho \begin{bmatrix} m_{00} \\ m_{01} \\ m_{11} \end{bmatrix} - g_0(1-\rho_0) \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} - g_1(1-\rho_0) \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} = \mathbf{0}_{3\times 1}. \tag{23}$$

where $m_{st} = E(y_{i,s}y_{i,t})$ and $g_s = E(y_{i,s}\eta_i)$. These moment conditions actually correspond to those employed by DIF. The idea here is that the term involving $g_0(1-\rho_0)$ can be removed by simply subtracting the first row of (23) from the second, yielding[7]

$$(m_{02} - m_{01}) - \rho_0(m_{01} - m_{00}) = E[y_{i,0}(\Delta y_{i,2} - \rho\Delta y_{i,1})] = 0. \tag{24}$$

Identification of ρ_0 based on (23) therefore requires $E(y_{i,0}\Delta y_{i,1}) \neq 0$, which is again violated if $\rho_0 = 1$. In terms of the general DGP considered in this paper, $\rho_0 = 1$ causes a violation of the full rank condition for $\mathbf{\Delta}$, specified in Assumption 6 (d). In order to see this, note that $m_{01} = E(y_{i,0}y_{i,1}) = E[y_{i,0}(\rho_0 y_{i,0} + \eta_i(1-\rho_0) + \varepsilon_{i,1})] = E(y_{i,0}^2)$, implying

$$\mathbf{\Delta} = -\begin{bmatrix} m_{00} & 1 & 0 \\ m_{01} & 1 & 0 \\ m_{11} & 0 & 1 \end{bmatrix} = -\begin{bmatrix} E(y_{i,0}^2) & 1 & 0 \\ E(y_{i,0}^2) & 1 & 0 \\ E(y_{i,1}^2) & 0 & 1 \end{bmatrix},$$

which has rank two only. But while GMM based on (23) fails to identify $\rho_0 = 1$, it is easy to see that the corresponding restricted GMM estimator that imposes $\eta_i(1-\rho_0) = 0$ does in fact

[7]The last row in (23) can be ignored because it is a moment condition that is consumed to estimate an extra parameter that does not appear elsewhere, namely, $g_1(1-\rho)$.

identify $\rho_0 = 1$, as $\mathbf{\Delta}$ has full rank in this case. But the restriction $\eta_i(1-\rho_0) = 0$ is equivalent to setting $r_0 = 0$. Hence, provided that r_0 is consistently estimated, the proposed GMM estimator is consistent regardless of the value of ρ_0.

It is important to note that the above mentioned failure of (23) to identify $\rho_0 = 1$ is due to the fact that under (6) $\boldsymbol{\lambda}_i$ is a function of ρ_0. If $\boldsymbol{\lambda}_i$ is not a function of ρ_0, then full identification based on (23) is possible for all values of ρ_0, including unity.

Remark 8. It is instructive to compare the approach considered here with the one of Bond et al. (2005), in which the popular "system" GMM estimator, hereafter SYS, is employed to test for a unit root. SYS is equivalent to the GMM estimator that results from imposing $g_0 = g_1$ in (23). In the fixed effects example considered in (21) this restriction holds true because

$$\begin{aligned} g_1 &= E(y_{i,1}\eta_i) = E[(\rho_0 y_{i,0} + (1-\rho_0)\eta_i + \varepsilon_{i,1})\eta_i] = \rho_0 E(y_{i,0}\eta_i) + (1-\rho_0)E(\eta_i^2) \\ &= \rho_0 E(\eta_i^2) + (1-\rho_0)E(\eta_i^2) = E(\eta_i^2) = g_0, \end{aligned}$$

where we assume that $E(y_{i,0}\eta_i) = E(\eta_i^2)$. The effect of this restriction can be appreciated by noting that (23) reduces to

$$\begin{bmatrix} m_{01} \\ m_{02} \\ m_{12} \end{bmatrix} - \rho_0 \begin{bmatrix} m_{00} \\ m_{01} \\ m_{11} \end{bmatrix} - g_0(1-\rho_0) \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \mathbf{0}_{3\times 1}, \tag{25}$$

with

$$\mathbf{\Delta} = - \begin{bmatrix} E(y_{i,0}^2) & 1 \\ E(y_{i,0}^2) & 1 \\ E(y_{i,1}^2) & 1 \end{bmatrix}.$$

Since $\mathbf{\Delta}$ has full column rank, $\rho_0 = 1$ is identified. Equivalently, subtracting the second row in (25) from the third yields

$$(m_{12} - m_{02}) - \rho_0(m_{11} - m_{01}) = E[\Delta y_{i,1}(y_{i,2} - \rho_0 y_{i,1})] = 0,$$

which, together with (24), make up the moment conditions for SYS. However, the aforementioned estimator has at least two potential shortcomings when compared to the proposed estimator. Firstly, it requires that $y_{i,t}$ is "mean stationary", that is, $E(y_{i,t}|\eta_i)$ is constant over t, which of course need not be the case in practice (see Bun and Sarafidis, 2015, for a detailed discussion). For example, if the model contains incidental trends, mean stationarity is obviously violated, rendering SYS invalid. Secondly, SYS is not equipped to handle the presence of genuine factors.

Remark 9. The above discussion focuses on the fixed effects case. The implications of the introduction of incidental trends depend on whether we are considering DIF or SYS. It can be shown that if $\rho_0 = 1$ the moment conditions employed by DIF have the same form as in (23), except that the coefficients of $[1, 1, 0]'$ and $[0, 1, 1]'$ do not depend on ρ_0 anymore. This estimator can identify $\rho_0 = 1$. SYS is, on the other hand, inconsistent, even in absence of genuine factors, since mean-stationarity is violated. Of course, in practice the deterministic component of the DGP is never really known, which is true also when it comes to the presence of genuine factors. In view of this, the fact that the new estimator does not require any a priori knowledge regarding $\mathbf{f}_t$ is a great advantage.

4 Monte Carlo simulations

4.1 Design

The DGP is of the same form as the one in (6), where we consider both the fixed effects ($d_{i,t} = \eta_i$) and incidental trends ($d_{i,t} = \eta_i + \beta_i t$) cases. Thus, while in the former (latter) case, under the unit root null hypothesis $y_{i,t}$ is a random walk without (with) a drift, under the alternative hypothesis, $y_{i,t}$ is stationary (trend-stationary). We also consider a "hybrid DGP", in which half of the cross-sectional units are generated with fixed effects, while the other half are generated with incidental trends. Under the null, $\rho_0 = 1$, while under the alternative, $\rho_0 \in \{0.95, 0.99\}$. In both cases, $\eta_i \sim \text{iid}\, N(0, \sigma^2_{\eta,i})$ and $\beta_i \sim \text{iid}\, N(0, \sigma^2_{\beta,i})$, where $\sigma_{\eta,i} \sim \text{iid}\, U[0, 2]$ and $\sigma_{\beta,i} \sim \text{iid}\, U[0, 2]$. As for $\varepsilon_{i,t}$, we set $\varepsilon_{i,t} \sim \text{iid}\, N(0, \sigma^2_{\varepsilon,t})$, where $\sigma_{\varepsilon,t} \sim \text{iid}\, U[0, 2]$, and $\varepsilon_{i,t} = v_{i,t} + 0.7 \cdot v_{i,t-1}$, $v_{i,t} \sim N(0, \sigma^2_{v,t}/(1 + 0.7^2))$ for $t \geq 0$ and $\sigma_{v,t} \sim \text{iid}\, U[0, 2]$. Hence, not only is there substantial cross-sectional heterogeneity in the DGP, but we also allow $\varepsilon_{i,t}$ to be heteroskedastic and correlated across time. The initial observation is generated as $y_{i,0} = u_{i,0} = d_{i,0} + \boldsymbol{\gamma}_i' \mathbf{w}_0 + \varepsilon_{i,0}$, where $d_{i,0} = \eta_i$ in both the fixed effects and incidental trends cases and $\varepsilon_{i,0} \sim \text{iid}\, N(0, 1)$. We consider two values for m, the dimension of $\mathbf{w}_t$, namely $m = 0$ and $m = 1$. Hence, while under the former parametrization, $\boldsymbol{\gamma}_i' \mathbf{w}_t = 0$, under the latter, $\boldsymbol{\gamma}_i' \mathbf{w}_t = \gamma_i w_t$, where γ_i and w_t are scalars satisfying $\gamma_i \sim \text{iid}\, N(0, \sigma^2_{\gamma,i})$, $w_t \sim \text{iid}\, N(0, 1)$ and $\sigma_{\gamma,i} \sim \text{iid}\, U[0, 2]$. Hence, in this DGP, $r_0 \in \{0, 1, 2, 3\}$; $r_0 = 0$ in the fixed effects case with $m = 0$ and $\rho_0 = 1$, $r_0 = 1$ in the fixed effects (incidental trends) case with $m = 0$ and $\rho_0 < 1$ ($\rho_0 = 1$), $r_0 = 2$ in the fixed effects (incidental trends) case with $\rho_0 < 1$ and $m = 1$ ($m = 0$), and $r_0 = 3$ in the incidental trends case with $\rho_0 < 1$ and $m = 1$. We set $N \in \{100, 400, 1600\}$ and the number of effective time periods, T^*, is set to

$T^* \in \{5, 7, 9, 11\}$.[8] All experiments are based on 5000 replications.

4.2 Results

The following results are reported: (i) the mean of $\hat{\rho}$; (ii) the standard deviation (STD) of $\hat{\rho}$; (iii) the empirical rejection frequency of $t_{\hat{\rho}}(1)$ at the nominal 5% level (for $\rho_0 \neq 1$ we report size-adjusted rejection frequency); (iv) the empirical rejection frequency of the overidentifying restrictions J-statistic (again at the nominal 5% level). To speed up the calculations, all results are based on setting $r_{max} = r_0 + 1$ with the estimation of r_0 carried out as explained in the previous section, using the $BIC(r)$ criterion.

Table 1 contains the results for the case when $m = 0$ and $T^* = 5$. The full set of moment conditions equals $M = 15$ in this case, noting that when $\varepsilon_{i,t}$ follows an MA(1) process only instruments up to $y_{i,t-2}$ are valid. As a rule of thumb, when $N/M \geq 10$, the results reported are based on the two-step estimator, whereas when $N/M < 10$, the results are based on the one-step estimator. This is in order to account for the well-known result that for small N and large M, the size of two-step GMM-based test statistics tends to be distorted.

The performance of the proposed estimator and the corresponding $t_{\hat{\rho}}(1)$-statistic is more than satisfactory. In particular, although for N small there appears to be some small bias, the bias goes away as N increases. Moreover, the size of the $t_{\hat{\rho}}(1)$-statistic is close to the nominal level in all experiments considered. The power of the $t_{\hat{\rho}}(1)$-statistic is satisfactory as well, and for $\rho_0 = .95$ it approaches unity rather quickly as N increases, which is reflection of the consistency of the test.

In terms of STD, the results indicate that the estimator performs best in the fixed effects case, followed by the incidental trends case, and then the hybrid case. This is to be expected since the fixed effects-only model implies more degrees of freedom when compared to the other models. The incidental trends and hybrid models involve the same number of parameters. However, as is made clear in Remark 7, the variance of the trend slope, β_i, in the hybrid model is half the value of the variance in the incidental trends model. The fact that the STD is relatively high in the hybrid model is therefore expected, as is the fact that there is little variation in the results depending on whether $\varepsilon_{i,t}$ is serially uncorrelated or not.

The results reported in Table 2 for the case when $m = 1$ are very similar to those reported in Table 1 for $m = 0$. A noticeable difference is the dispersion of the estimator, which is much

[8] In absence of serial correlation in $\varepsilon_{i,t}$, $T^* = T$, whereas when $\varepsilon_{i,t}$ follows an MA process, $T^* = T + 1$. For the incidental trend and hybrid DGPs, we set the minimum value of T^* equal to 6, in order to have enough moment conditions to be able to make inference based on the overidentifying restrictions J-statistic.

larger in Table 2 than in Table 1. This is due to the fact that there are more parameters to estimate with essentially the same amount of information.

Table 3 reports the empirical rejection frequency for several values of T^*. We see that, as long as N is not too small, size remains close to the nominal level in all cases. On the other hand, power appears to increase quite substantially with higher values of T^*. For instance, in the fixed effects-only model when $m = 0$, $N = 1600$ and $\rho_0 = .99$, power increases from .589 when $T^* = 5$ to .993 when $T^* = 11$. The corresponding increase in power in the model with incidental trends is from .208 to .777.

Table 4 reports the power of the overidentifying restrictions J-statistic for the case when $\varepsilon_{i,t}$ follows an MA(1) process. This is an important empirical scenario, as serial correlation in $\varepsilon_{i,t}$ invalidates a subset of the proposed moment conditions. As we can see, in all cases considered the power of the test is high.

5 Application

5.1 Gibrat's Law

In this section we make use of our methodology in order to examine the empirical validity of the well known "Law of Proportionate Effect", or simply Gibrat's Law (Gibrat, 1931) using data from the US banking industry. Gibrat's Law postulates that the growth rate of firms is independent of their initial size. Analytically, we have

$$\Delta \ln s_{i,t} = \delta + (\rho - 1) \ln s_{i,t-1} + u_{i,t}, \tag{26}$$

where $s_{i,t}$ denotes the size of firm i at time t and $u_{i,t}$ is an error term. For $\rho < 1$ larger firms tend to grow at a lower rate compared to smaller firms, while for $\rho > 1$ the process is explosive and growth rate is proportional to firm size. For $\rho = 1$ Gibrat's law holds true because firms' growth rate is independent of their initial size. This means that Gibrat's Law can be examined by testing for a unit root in $\ln s_{i,t}$.

Gibrat's law has become very popular because it provides an explanation for what has been identified as an empirical regularity where the distribution of firms' size is often highly skewed across several industries. In particular, many sectors are characterised by a log-normal distribution with a larger number of small to medium scale firms and relatively few large firms; see Steindl (1965). Simon and Bonini (1958) argue that under (approximate) constant returns to scale it is natural to expect that the probability for a given firm to increase/decrease in size in

proportion to its existing size is the same, on average, for all firms in the industry that lie above a critical minimum size value. On the other hand, some of the more recent empirical evidence (see, for example, Sutton, 1997; Caves, 1998) appears to suggest that while Gibrat's law tends to be confirmed in small subsamples of well-established, mature, large firms, this is not always the case for larger samples that include small and young firms, since the latter often have higher growth rate than their larger counterparts.

As the discussion in the above paragraph suggests, the relation between firm size and growth rate remains an open issue. In this section, we therefore provide new evidence based on our newly developed test, which has a number of advantages when compared to existing tests. First, the test is valid for all values of ρ, including $\rho > 1$. Second, unlike many other tests, the test developed here can be implemented without knowing the elements in $\mathbf{f}_t$. This advantage is particularly relevant in the present context, because the size of the firm is likely to depend on its age, a variable that is not in our sample. As an illustration, let us denote by $a_{i,t}$ the age of firm i, and let β_i denote the impact of age on the size of the firm. Since age increases at the same rate for all firms, we may write $\beta_i a_{i,t} = \beta_i(a_{i,0} + t) = \gamma_i + \beta_i t$, where $\gamma_i = \beta_i + a_{i,0}$ with $a_{i,0}$ being the age of the firm at the beginning of the sample. Hence, under said conditions, the effect of age can be captured using two factors, one is a constant, while the other is a linear time trend. Of course, a priori we cannot say for sure that age has an effect, and so it is more convenient to treat age as an unobserved factor to be estimated from the data. Another possibility is that the factors represent in part common shocks due to for example the global financial crisis.

5.2 Data

The data set consists of a panel of $N = 4,022$ depository financial institutions, for which we have annual observations covering the period 2003–2011. These data have been collected from the electronic database maintained by the Federal Deposit Insurance Corporation (FDIC) (see http://www.fdic.gov). Two measures of bank size are considered; (i) fixed assets (FA), and (ii) number of employees (EMP). Both variables are transformed by taking logs and FA is deflated using the GDP deflator.

5.3 Results

As starting values for the factors we consider ($\sqrt{T}$ times) the eigenvectors corresponding to the largest r eigenvalues of the $T \times T$ matrix $\sum_{i=1}^{N} (\mathbf{y}_i - \alpha \mathbf{y}_{i,-1}) (\mathbf{y}_i - \alpha \mathbf{y}_{i,-1})' / N$, where $\alpha = \{0, .1, .2, ..., 1\}$, as well as fixed effects, linear trends, and a large number of random initializa-

tions from the normal and uniform distributions. We fit a maximum of $r_{max} = 3$ factors and use the $BIC(r)$ criterion to pick the most appropriate number, given that it passes the J test at the 5% level. To gauge against possible serial correlation in the errors, the GMM approach is implemented assuming MA(q) errors, where $q \in \{1, 2\}$. If the model is misspecified, this is likely to show up in the J test.

Table 5 reports results obtained based on the two-step GMM estimator. The results are very similar for EMP and FA. In particular, the point estimate of ρ_0 is below unity and the unit root null hypothesis is rejected even at the 1% level, suggesting that Gibrat's Law is not supported by the data. The null hypothesis of instrument validity/correct model specification is also not rejected. For EMP (FA) the best fitting model according to $BIC(r)$ has one (two) factors. Ocular inspection reveals that in case of FA while the first factor resembles a trend, the second factor has a less clear cut shape. This demonstrates the importance of allowing for nonlinear effects, casting doubt on existing results based on fixed effects-only unit root tests. The estimated factor for EMP is very similar to the first factor for FA. In fact, the correlation between the two factors is 0.915.

Our results imply that during the sampling period investigated, Gibrat's Law is violated implying that the growth rate of financial institutions is negatively correlated to their initial size, that is, smaller institutions appear to grow faster than their larger counterparts.

6 Conclusion

This paper develops a GMM-based approach that enables unit root testing in panels where N is large and T is finite. The assumption that T finite makes our test suitable for both micro and small-T macro panels. The DGP considered is very general and accommodates an unrestricted trend function and cross-sectional dependence in the form of common factors. These allowances make the new approach one of the most general around. Indeed, as far as we are aware, this is the only fixed-T unit root test approach that can be applied in the presence of cross-sectional dependence and/or a potentially non-linear trend function. The approach is also relatively simple to implement. In particular, since deterministic terms are treated as additional common factors, which are estimated, there is no need to model the deterministic part. Our results show that the new GMM-based unit root test statistic is asymptotically invariant to both the true and fitted deterministic trend function. Hence, unlike existing tests, with the new test there is no need for any mean and/or variance correction factors that reflect the fitted deterministic specification.

The limiting distribution of the GMM t-statistic is normal and this holds true regardless of the value of the AR coefficient, ρ_0. Hence, again unlike most existing tests, with this test there is no discontinuity in the asymptotic distribution at unity. The asymptotic properties are verified in small samples using both simulated and raw data.

References

Abadir, K. M., and J. R. Magnus (2005). *Matrix Algebra, Econometric Exercises 1*. Cambridge University Press, New York.

Ahn, S. C., Y. H. Lee, and P. Schmidt (2013). Panel Data Models with Multiple Time-Varying Individual Effects. *Journal of Econometrics* **174**, 1–14.

Anderson, T. W., and C. Hsiao (1981). Estimation of Dynamic Models with Error Components. *Journal of American Statistical Association* **76**, 598–606.

Arellano, M., and S. Bond (1991). Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations. *Review of Economic Studies* **58**, 277–297.

Bai, J. (2009). Panel Data Models with Interactive Fixed Effects. *Econometrica* **77**, 1229–1279.

Bai, J. (2013). Fixed-Effects Dynamic Panel Models, a Factor Analytical Method. *Econometrica* **81**, 285–314.

Bai, J., and S. Ng (2004). A Panic Attack on Unit Roots and Cointegration. *Econometrica* **72**, 1127–1177.

Bai, J. and S. Ng (2008). Large Dimensional Factor Analysis. *Foundations and Trends in Econometrics* **3**, 89–163.

Bai, J., and S. Ng (2010). Panel Unit Root Tests With Cross-Section Dependence: A Further Investigation. *Econometric Theory* **26**, 1088–1114.

Baltagi, B. (2008). *Econometric Analysis of Panel Data, Fourth Edition*. John Wiley & Sons, New York.

Blundell, R., and S. Bond (1998). Initial Conditions and Moment Restrictions in Dynamic Panel Data Models. *Journal of Econometrics* **87**, 115–143.

Bond, S., C. Nauges, and F. Windmeijer (2005). Unit Roots: Identification and Testing in Micro Panels. Unpublished manuscript.

Breitung, J., and H. Pesaran (2008). Unit Roots and Cointegration in Panels. In L. Matyas and P. Sevestre (ed.), *The Econometrics of Panel Data*.

Bun, M., and F. Kleibergen (2013). Identification and Inference in Moments Based Analysis of Linear Dynamic Panel Data Models. UvA-Econometrics Discussion Paper 2013/07.

Bun, M., and F. Windmeijer (2010). The Weak Instrument Problem of the System GMM Estimator in Dynamic Panel Data Models. *Econometrics Journal* **13**, 95–126.

Bun, M., and V. Sarafidis (2015). Dynamic Panel Data Models. In B. H. Baltagi (ed.), *The Oxford Handbook of Panel Data*, Chapter 4, Oxford University Press, Oxford.

Caves, R. E. (1998). Industrial Organization and New Findings on the Turnover and Mobility of Firms. *Journal of Economic Literature* **36**, 1947-1982.

Chesher, A. (1979). Testing the Law of Proportionate Effect. *Journal of Industrial Economics* **27**, 403–411.

Chudik, A. M., H. Pesaran and E. Tosetti (2011). Weak and Strong Cross Section Dependence and Estimation of Large Panels. *Econometrics Journal* **14**, C45–C90.

De Blander, R., and G. Dhaene (2012). Unit Root Tests for Panel Data with AR(1) Errors and Small T. *Econometrics Journal* **15**, 101–124.

De Silva, S., K. Hadri and A. R. Tremayne (2009). Panel Unit Root Tests in the Presence of Cross-Sectional Dependence: Finite Sample Performance and an Application. *Econometrics Journal* **12**, 340–366.

De Wachter, S., R. Harris, and E. Tzavalis (2007). Panel Data Unit Roots Tests: The Role of Serial Correlation and the Time Dimension. *Journal of Statistical Planning and Inference 137*, 230–244.

Hadri, K., and R. Larsson (2005). Testing for Stationarity in Heterogeneous Panel Data where the Time Dimension is Finite. *Econometrics Journal* **8**, 55–69.

Han, C., and P. C. B. Phillips (2010). GMM Estimation for Dynamic Panels with Fixed Effects and Strong Instruments at Unity. *Econometric Theory* **26**, 119–151.

Harris, R., and E. Tzavalis (1999). Inference for Unit Roots in Dynamic Panels where the Time Dimension is Fixed. *Journal of Econometrics* **91**, 201–226.

Harris, R., and E. Tzavalis (2004). Testing for Unit Roots in Dynamic Panels in the Presence of a Deterministic Trend: Re-examining the Unit Root Hypothesis for Real Stock Prices and Dividends. *Econometric Reviews* **23**, 149–166.

Hlouskova, J., and M. Wagner (2006). The Performance of Panel Unit Root and Stationarity Tests: Results from a Large Scale Simulation. *Econometric Reviews* **25**, 85–116.

Im, K. S., M. H. Pesaran and Y. Shin (2003). Testing for Unit Roots in Heterogeneous Panels. *Journal of Econometrics* **115**, 53–74.

Karavias, Y., and E. Tzavalis (2014). Testing for Unit Roots in Panels with Structural Changes, Spatial and Temporal Dependence when the Time Dimension is Finite. Unpublished manuscript.

Kruiniger, H. (2007). An Efficient Linear GMM Estimator for the Covariance Stationary AR(1)/Unit Root Model for Panel Data. *Econometric Theory* **23**, 519–535.

Kruiniger, H. (2009). GMM Estimation and Inference in Dynamic Panel Data Models with Persistent Data. *Econometric Theory* **25**, 1348–1391.

Kruiniger, H. (2013). Quasi ML Estimation of the Panel AR(1) Model with Arbitrary Initial Conditions. *Journal of Econometrics* **173**, 175–188.

Moon, H. R., and B. Perron (2004). Testing for Unit Root in Panels with Dynamic Factors. *Journal of Econometrics* **122**, 81–126.

Levin, A., C. Lin, and C.-J. Chu (2002). Unit Root Tests in Panel Data: Asymptotic and Finite-sample Properties. *Journal of Econometrics* **108**, 1–24.

Pesaran, M. H. (2007). A Simple Panel Unit Root Test in the Presence of Cross Section Dependence. *Journal of Applied Econometrics* **22**, 265–312.

Phillips, P. C. B., and D. Sul (2003). Dynamic Panel Estimation and Homogeneity Testing Under Cross Section Dependence. *Econometrics Journal* **6**, 217–259.

Robertson, D., and V. Sarafidis (2015). IV Estimation of Panels with Factor Residuals. *Journal of Econometrics* **185**, 526–541.

Sarafidis, V., and T. Wansbeek (2012). Cross-Sectional Dependence in Panel Data Analysis. *Econometric Reviews* **31**, 483–531.

Schmidt, P., and P. C. B. Phillips (1992). LM Tests for a Unit Root in the Presence of Deterministic Trends. *Oxford Bulletin of Economics and Statistics* **54**, 257–287.

Simon, H. A., and C. P. Bonini (1958). The Size Distribution of Business Firms. *American Economic Review* **58**, 607–617.

Steindl, J. (1965). *Random Processes and the Growth of Firms: a Study of the Pareto Law*. Hafner Publishing, New York.

Sutton, J. (1997). Gibrat's Legacy. *Journal of Economic Literature* **35**, 40–59.

Wansbeek, T., and P. Bekker (1996). On IV, GMM and ML in a Dynamic Panel Data Model. *Economics Letters* **51**, 145–152.

Westerlund, J., and J. Breitung (2013). Lessons from a Decade of IPS and LLC. *Econometric Reviews* **32**, 547–591.

Westerlund, J. (2015). The Effect of Recursive Detrending on Panel Unit Root Tests. *Journal of Econometrics* **185**, 453–467.

Appendix

Proof of Theorem 1

The proof of Theorem 1 follows directly from Theorems 2.6 and 3.4 in Newey and McFadden (1994). In particular, it is straightforward to establish that the conditions listed therein are fulfilled under our Assumptions 1–6. For example, continuity of the moment functions holds, because $\mathbf{h}_i(\boldsymbol{\theta})$ is essentially a vector of elementary functions of $\boldsymbol{\theta}$, which belongs to the interior of Θ. Moreover, condition (iv) of Theorem 2.6 in Newey and McFadden (1994) holds, because of the compactness of Θ and the fact that $y_{i,t}$ has finite moments up to second order. ■

Derivatives

In this section we make heavy use of the results of Abadir and Magnus (2005). Readers are referred to this book for a more detailed treatment of the arguments used here.

Letting

$$\mathbf{U}(\theta_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3) = \mathbf{S}(\mathbf{I}_T \otimes (\boldsymbol{\Gamma}\mathbf{e}_1\mathbf{g}_0'\mathbf{F}' + \boldsymbol{\Gamma}\widetilde{\mathbf{F}}_{-1}\mathbf{F}'))\mathbf{e},$$

we have

$$\begin{aligned}
\mathbf{h}_i(\boldsymbol{\theta}) &= \mathbf{Z}_i'\mathbf{y}_i - \rho\mathbf{Z}_i'\mathbf{y}_{i,-1} - \mathbf{U}(\theta_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3), \\
\overline{\mathbf{h}}_N(\boldsymbol{\theta}) &= \hat{\mathbf{m}} - \rho\hat{\mathbf{m}}_1 - \mathbf{U}(\theta_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3).
\end{aligned}$$

Clearly,

$$\begin{aligned}
\mathbf{U}(\theta_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3) &= \operatorname{vec}(\mathbf{U}(\boldsymbol{\theta})) = (\mathbf{e}' \otimes \mathbf{S})\operatorname{vec}(\mathbf{I}_T \otimes (\boldsymbol{\Gamma}\mathbf{e}_1\mathbf{g}_0'\mathbf{F}' + \boldsymbol{\Gamma}\widetilde{\mathbf{F}}_{-1}\mathbf{F}')) \\
&= (\mathbf{e}' \otimes \mathbf{S})(\mathbf{I}_T \otimes \mathbf{K}_{T,T} \otimes \mathbf{I}_T)[\operatorname{vec}(\mathbf{I}_T) \otimes \operatorname{vec}(\boldsymbol{\Gamma}\mathbf{e}_1\mathbf{g}_0'\mathbf{F}' + \boldsymbol{\Gamma}\widetilde{\mathbf{F}}_{-1}\mathbf{F}')] \\
&= (\mathbf{e}' \otimes \mathbf{S})(\mathbf{I}_T \otimes \mathbf{K}_{T,T} \otimes \mathbf{I}_T)[\operatorname{vec}(\mathbf{I}_T) \otimes \mathbf{I}_{T^2}\operatorname{vec}(\boldsymbol{\Gamma}\mathbf{e}_1\mathbf{g}_0'\mathbf{F}' + \boldsymbol{\Gamma}\widetilde{\mathbf{F}}_{-1}\mathbf{F}')] \\
&= (\mathbf{e}' \otimes \mathbf{S})(\mathbf{I}_T \otimes \mathbf{K}_{T,T} \otimes \mathbf{I}_T)[\operatorname{vec}(\mathbf{I}_T) \otimes \mathbf{I}_{T^2}][1 \otimes \operatorname{vec}(\boldsymbol{\Gamma}\mathbf{e}_1\mathbf{g}_0'\mathbf{F}' + \boldsymbol{\Gamma}\widetilde{\mathbf{F}}_{-1}\mathbf{F}')] \\
&= (\mathbf{e}' \otimes \mathbf{S})(\mathbf{I}_T \otimes \mathbf{K}_{T,T} \otimes \mathbf{I}_T)(\mathbf{e} \otimes \mathbf{I}_{T^2})\operatorname{vec}(\boldsymbol{\Gamma}\mathbf{e}_1\mathbf{g}_0'\mathbf{F}' + \boldsymbol{\Gamma}\widetilde{\mathbf{F}}_{-1}\mathbf{F}') \\
&= \mathbf{A}\operatorname{vec}(\boldsymbol{\Gamma}\mathbf{e}_1\mathbf{g}_0'\mathbf{F}' + \boldsymbol{\Gamma}\widetilde{\mathbf{F}}_{-1}\mathbf{F}'),
\end{aligned}$$

where $\mathbf{A} = (\mathbf{e}' \otimes \mathbf{S})(\mathbf{I}_T \otimes \mathbf{K}_{T,T} \otimes \mathbf{I}_T)(\mathbf{e} \otimes \mathbf{I}_{T^2})$, a $M \times T^2$ matrix. Here, $\mathbf{K}_{k,n}$ is the $kn \times kn$ commutation matrix of zeroes and ones such $\mathbf{K}_{k,n}\operatorname{vec}\mathbf{H} = \operatorname{vec}\mathbf{H}'$ for any $k \times n$ matrix $\mathbf{H}$. As a

result,

$$\begin{aligned}
\mathrm{D}_{\theta_1}\mathbf{U}(\theta_1,\boldsymbol{\theta}_2,\boldsymbol{\theta}_3) &= \mathbf{A}[(\mathbf{F}\mathbf{g}_0\mathbf{e}_1' + \mathbf{F}\widetilde{\mathbf{F}}_{-1}')\otimes\mathbf{I}_T]\mathrm{D}_\rho\,\boldsymbol{\Gamma}(\rho);\\
\mathrm{D}_{\theta_2}\mathbf{U}(\theta_1,\boldsymbol{\theta}_2,\boldsymbol{\theta}_3) &= \mathbf{A}[\mathbf{I}_T\otimes(\boldsymbol{\Gamma}\mathbf{e}_1\mathbf{g}_0' + \boldsymbol{\Gamma}\widetilde{\mathbf{F}}_{-1}) + (\mathbf{F}\otimes\boldsymbol{\Gamma})\mathbf{K}_{r,T}(\mathbf{B}\otimes\mathbf{I}_r)];\\
\mathrm{D}_{\theta_3}\mathbf{U}(\theta_1,\boldsymbol{\theta}_2,\boldsymbol{\theta}_3) &= \mathbf{A}(\mathbf{F}\otimes\boldsymbol{\Gamma}\mathbf{e}_1),
\end{aligned}$$

where

$$\mathrm{D}_\rho\,\boldsymbol{\Gamma}(\rho) = \mathrm{vec}\begin{bmatrix} 0 & 0 & \dots & 0 & 0\\ 1 & 0 & \dots & 0 & 0\\ 2\rho & 1 & \dots & 0 & 0\\ \vdots & \vdots & \ddots & \vdots & 0\\ (T-1)\rho^{T-2} & (T-2)\rho^{T-3} & & 1 & 0 \end{bmatrix},$$

and $\mathbf{B}$ is a $T\times T$ matrix with zeros everywhere except in the first lower-diagonal that takes the value of one. We can therefore show that

$$\nabla_{\boldsymbol{\theta}}\overline{\mathbf{h}}_N(\boldsymbol{\theta}) = -[\hat{\mathbf{m}}_1 + \mathrm{D}_{\theta_1}\mathbf{U}(\theta_1,\boldsymbol{\theta}_2,\boldsymbol{\theta}_3),\ \mathrm{D}_{\theta_2}\mathbf{U}(\theta_1,\boldsymbol{\theta}_2,\boldsymbol{\theta}_3),\ \mathrm{D}_{\theta_3}\mathbf{U}(\theta_1,\boldsymbol{\theta}_2,\boldsymbol{\theta}_3)].$$

Table 1: Simulation results for the case when $m = 0$ and $T^* = 5$.

		Fixed effects				Incidental trends				Hybrid			
ρ	N	Mean	STD	$t_{\hat{\rho}}(1)$	J	Mean	STD	$t_{\hat{\rho}}(1)$	J	Mean	STD	$t_{\hat{\rho}}(1)$	J
$\varepsilon_{i,t}$ is serially uncorrelated													
1	100	.997	.021	.065	.059	.989	.054	.065	.079	.989	.076	.061	.060
1	400	1.00	.008	.069	.045	1.00	.025	.059	.043	1.00	.028	.054	.062
1	1600	1.00	.003	.051	.047	1.00	.009	.052	.050	1.00	.012	.052	.051
.99	100	.982	.046	.097	.043	.978	.073	.062	.069	.988	.076	.091	.058
.99	400	.987	.026	.203	.044	.987	.019	.091	.056	.986	.029	.125	.059
.99	1600	.989	.009	.589	.049	.989	.011	.208	.052	.990	.010	.196	.050
.95	100	.958	.026	.354	.071	.938	.072	.234	.032	.939	.071	.266	.032
.95	400	.956	.016	.967	.067	.948	.024	.363	.048	.948	.027	.398	.046
.95	1600	.953	.012	1.00	.061	.950	.010	.533	.051	.950	.012	.598	.050
$\varepsilon_{i,t}$ follows an MA(1) process													
1	100	.996	.025	.062	.043	.991	.057	.061	.081	.988	.081	.053	.061
1	400	1.00	.011	.067	.045	1.00	.029	.062	.055	1.00	.031	.055	.058
1	1600	1.00	.003	.049	.048	1.00	.011	.054	.054	1.00	.014	.052	.055
.99	100	.980	.051	.103	.042	.977	.077	.072	.094	.985	.081	.101	.060
.99	400	.986	.022	.194	.049	.988	.022	.074	.062	.985	.033	.117	.062
.99	1600	.987	.010	.569	.046	.988	.012	.183	.053	.988	.012	.176	.057
.95	100	.960	.030	.337	.044	.941	.076	.240	.079	.936	.076	.248	.054
.95	400	.956	.019	.946	.053	.947	.023	.347	.064	.946	.030	.340	.051
.95	1600	.954	.013	1.00	.051	.949	.011	.468	.054	.949	.013	.571	.053

Notes: "Mean" and "STD" refer to mean and standard deviation of $\hat{\rho}$, respectively. The results reported for the J-statistic are the empirical sizes. In the "hybrid" model, half of the units are generated with fixed effects, while the other half is generated with incidental trends.

Table 2: Simulation results for the case when $m = 1$ and $T^* = 5$.

		Fixed effects				Incidental trends				Hybrid			
ρ	N	Mean	STD	$t_{\hat{\rho}}(1)$	J	Mean	STD	$t_{\hat{\rho}}(1)$	J	Mean	STD	$t_{\hat{\rho}}(1)$	J
$\varepsilon_{i,t}$ is serially uncorrelated													
1	100	.995	.049	.043	.041	.983	.096	.045	.081	.985	.087	.042	.077
1	400	1.00	.020	.065	.044	.994	.047	.070	.063	.998	.051	.064	.072
1	1600	1.00	.006	.060	.046	.999	.011	.055	.044	1.00	.019	.054	.061
.99	100	.988	.051	.056	.041	.968	.095	.048	.079	.974	.102	.083	.081
.99	400	.993	.020	.092	.051	.983	.047	.110	.072	.987	.053	.123	.079
.99	1600	.993	.008	.307	.064	.989	.011	.182	.035	.992	.017	.189	.063
.95	100	.961	.062	.227	.073	.936	.086	.134	.042	.933	.089	.126	.042
.95	400	.964	.037	.798	.079	.942	.046	.535	.038	.946	.054	.547	.048
.95	1600	.952	.013	.937	.063	.950	.010	.967	.047	.950	.013	.935	.054
$\varepsilon_{i,t}$ follows an MA(1) process													
1	100	.992	.053	.042	.063	.981	.107	.064	.085	.983	.091	.035	.072
1	400	.996	.027	.057	.056	.995	.051	.061	.067	.999	.053	.043	.066
1	1600	1.00	.008	.055	.057	1.00	.017	.056	.053	1.00	.022	.049	.059
.99	100	.984	.054	.053	.034	.972	.098	.064	.082	.981	.105	.062	.083
.99	400	.987	.027	.089	.059	.984	.051	.102	.076	.983	.059	.079	.074
.99	1600	.991	.013	.305	.047	.986	.015	.173	.063	.988	.021	.116	.059
.95	100	.964	.069	.212	.084	.939	.092	.125	.065	.937	.096	.133	.047
.95	400	.959	.043	.763	.083	.945	.054	.529	.042	.945	.058	.446	.046
.95	1600	.948	.017	.894	.061	.948	.013	.941	.057	.948	.015	.915	.054

Notes: See Table 1.

Table 3: Rejection frequencies for different values of T^* when $\varepsilon_{i,t}$ is serially uncorrelated.

		$T^*=7$			$T^*=9$			$T^*=11$		
ρ_0	N	FE	TR	HY	FE	TR	HY	FE	TR	HY
					$m=0$					
1	100	.051	.050	.053	.052	.045	.047	.053	.039	.040
	400	.047	.084	.079	.063	.047	.048	.047	.045	.056
	1600	.050	.053	.057	.054	.046	.049	.051	.051	.046
.99	100	.081	.051	.068	.072	.068	.067	.108	.078	.083
	400	.198	.102	.112	.155	.101	.112	.231	.159	.126
	1600	.857	.193	.186	.950	.542	.569	.993	.777	.769
.95	100	.441	.217	.224	.585	.194	.219	.738	.870	.888
	400	.836	.754	.768	.986	.793	.831	.999	1.00	1.00
	1600	1.00	.713	.736	1.00	1.00	1.00	1.00	1.00	1.00
					$m=1$					
1	100	.038	.068	.057	.039	.097	.088	.040	.095	.082
	400	.050	.078	.066	.050	.078	.069	.047	.077	.061
	1600	.054	.075	.061	.055	.059	.056	.048	.061	.055
.99	100	.067	.056	.068	.073	.065	.067	.081	.075	.081
	400	.110	.094	.102	.131	.103	.121	.182	.141	.152
	1600	.819	.173	.193	.903	.373	.419	.969	.608	.690
.95	100	.701	.343	.368	.477	.176	.195	.565	.213	.328
	400	.983	.374	.381	.930	.653	.701	1.00	.833	.981
	1600	1.00	.978	.998	1.00	1.00	1.00	1.00	1.00	1.00

Notes: "FE", "TR" and "HY" refer to the fixed effects, the incidental trends and hybrid models, respectively.

Table 4: Power of the J-statistic when $\varepsilon_{i,t}$ follows an MA(1) process.

	$\rho_0=1$			$\rho_0=.99$			$\rho_0=.95$		
N	FE	TR	HY	FE	TR	HY	FE	TR	HY
				$m=0$					
100	.983	.841	.843	.984	.836	.840	.997	.838	.834
400	1.00	1.00	.994	1.00	1.00	.998	1.00	.997	.996
1600	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
				$m=1$					
100	.857	.794	.782	.947	.757	.816	.954	.769	.776
400	.957	.938	.895	.998	.839	.967	.972	.968	.920
1600	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Notes: The results for $m=0$ ($m=1$) are based on $T^*=5$ ($T^*=7$).

Table 5: Empirical results.

Measure	$\hat{\rho}$	$\hat{\sigma}_\rho$	$t_{\hat{\rho}}(1)$	p-value	J	p-value	BIC1	$\hat{r}$	q
EMP	.866	.001	-126.1	.000	22.08	.228	-57.9	1	1
FA	.822	.001	-128.0	.000	3.89	.918	-36.1	2	1

Notes: "$\hat{\rho}$" and "$\hat{\sigma}_\rho$" refer to the two-step GMM estimator of ρ_0 and its estimated standard error, "$t_{\hat{\rho}}(1)$" refers to the unit root t-statistic, "J" refers to the Hansen–Sargan statistic, "BIC1" refers to the minimizing value of the BIC1, "$\hat{r}$" refers to the estimated number of factors using the BIC1, and "q" refers to the order of the assumed MA errors.